Biased and Unbiased Cross-Validation in Density Estimation

Author(s): David W. Scott and George R. Terrell

Source: *Journal of the American Statistical Association*, Dec., 1987, Vol. 82, No. 400 (Dec., 1987), pp. 1131-1146

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2289391

Biased and Unbiased Cross-Validation in Density Estimation

DAVID W. SCOTT and GEORGE R. TERRELL*

Nonparametric density estimation requires the specification of smoothing parameters. The demands of statistical objectivity make it highly desirable to base the choice on properties of the data set. In this article we introduce some biased cross-validation criteria for selection of smoothing parameters for kernel and histogram density estimators, closely related to one investigated in Scott and Factor (1981). These criteria are obtained by estimating L_2 norms of derivatives of the unknown density and provide slightly biased estimates of the average squared L_2 error or mean integrated squared error. These criteria are roughly the analog of Wahba's (1981) generalized cross-validation procedure for orthogonal series density estimators. We present the relationship of the biased cross-validation procedure to the least squares cross-validation procedure, which provides unbiased estimates of the mean integrated squared error. Both methods are shown to be based on U statistics. We compare the two methods by theoretical calculation of the noise in the cross-validation functions and corresponding cross-validated smoothing parameters, by Monte Carlo simulation, and by example. Surprisingly large gains in asymptotic efficiency are observed when biased cross-validation is compared with unbiased cross-validation if the underlying density is sufficiently smooth. The theoretical results explain some of the small sample behavior of cross-validation functions: we show that cross-validation algorithms can be unreliable for sample sizes that are "too small." To aid the practitioner in the use of these appealing automatic cross-validation algorithms and to help facilitate evaluation of future algorithms, we must address some oftentimes controversial issues in density estimation: squared loss, the integrated squared error and mean integrated squared error criteria, adaptive density estimates, sample size requirements, and assumptions about the underlying density's smoothness. We conclude that the two cross-validation procedures behave quite differently, so one might well use both in practice.

KEY WORDS: Smoothing parameter; Kernel density estimation; Histogram.

1. INTRODUCTION

1.1 Background

Much theoretical progress has been made recently with the important problem of data-based methods for choosing smoothing parameters in nonparametric curve estimation procedures since the early work of Kronmal and Tarter (1968), Woodroofe (1970), and Stone (1974). In density estimation particular attention has been paid to the least squares cross-validation (CV) algorithm described independently by Rudemo (1982) and Bowman (1984). The sequence of smoothing parameters produced by this procedure not only leads to consistent density estimates but is asymptotically optimal in a certain sense, as shown by Hall (1983) and Stone (1984). Recently Hall and Marron (1985) characterized the limiting distribution of this sequence. This theory indicates that these CV sequences converge at perhaps surprisingly slow rates. As was the case with the original kernel theory of Rosenblatt and Parzen, the new theory is asymptotic in nature, so considerable effort will be required to understand fully the practical aspects of these methods and their performance with real data. For samples of size under 100 with Gaussian kernel estimates, two simulation studies have been completed. First, Scott and Factor (1981) showed that the average behavior of some earlier CV algorithms was good for Gaussian data in the sense that the CV smoothing parameters were centered on the value predicted by minimizing mean integrated squared error (MISE). Second, Bowman (1985) presented a study using six sampling densities and eight CV algorithms. We are unaware of studies involving much larger samples.

The goal of cross-validation is to automatically provide nearly optimally calibrated nonparametric estimates, mimicking the choices of experts and perhaps surpassing them. Consistency of CV algorithms is important, but we are more concerned with understanding small-sample reliability, which we define as the smallest sample size for which there is a 90% chance of being within 10%–15% of the optimal smoothing parameter. This is a useful rule of thumb because, even for extremely large samples, density estimates with smoothing parameters outside this narrow range are either distorted or visually noisy. Highly reliable CV algorithms would provide scientific reproducibility of density estimates between laboratories, an important but elusive goal. Our objectives are similar to those of researchers trying to retain the reproducibility of multiple linear regression while introducing transformations and robust methods via artificial intelligence (Gale and Pregibon 1983). Carroll and Ruppert (1985) expressed caution about using robust methods "blindly." We will show that similar caution is appropriate for cross-validation of density estimators.

There is a rich literature on data-based smoothing algorithms for nonparametric methods. A survey of smoothing methods for density estimation may be found in Scott (1986). A more general survey was given by Titterington (1985). In our discussion we shall focus on density estimation, although the situation for nonparametric regression is parallel (Härdle and Marron 1985; Rice 1984). With regression, one must pay attention to the interactions among choices of the regression curve, the signal-to-noise ratio,

* David W. Scott is Professor, Department of Mathematical Sciences, Rice University, Houston, TX 77251. George R. Terrell is Assistant Professor, Department of Statistics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. This research was partially supported by Office of Naval Research (ONR) Contract N00014-85-K-0100 and Army Research Office Contract DAAG-29-85-K0212. This article was written while the first author was visiting Stanford University, and the authors wish to acknowledge the generous support of colleagues there, particularly the partial support of J. Friedman under ONR Contract N00014-83-K-0472. They also benefited from several discussions with Steve Marron and comments by the referees.

Journal of the American Statistical Association
December 1987, Vol. 82, No. 400, Theory and Methods

and the distribution of the noise, whereas we only need consider the density curve here.

A few notations are used extensively throughout the article. We shall denote the squared L_2 norm of a function ψ by

$$R(\psi) \equiv \|\psi\|_2^2 = \int_{-\infty}^{\infty} \psi(x)^2\, dx, \tag{1.1}$$

where R reminds us that (1.1) is one possible measure of the roughness of ψ. The square of the pth derivative of ψ will be denoted by $\psi^{(p)}(x)^2$. Integrals without limits are assumed to be over the entire real line.

At this point it is helpful to indicate the organization of the article. It would be natural at a first reading to proceed to Sections 5 and 6 after reading Sections 1.2 and only glancing at theoretical results in Sections 2–4. Sections 7–8 contain examples and discussion. Proofs of results are indicated in Section 9 and are similar to those in Hall and Marron (1985). Expanded proofs are available in Scott and Terrell (1986).

1.2 Example

We begin by plotting two CV functions for an equally spaced histogram with bin width h. Given a random sample of size n, let $\nu_h(k)$ be the bin count in the kth bin $[kh, (k + 1)h)$, where without loss of generality we may assume that the mesh includes 0. In Section 3.1, we show that the least squares CV function is

$$e_0(h) = \frac{2}{nh} - \frac{1}{n^2h} \sum_{k=-\infty}^{\infty} \nu_h(k)^2, \tag{1.2}$$

and in Section 3.2 we propose a biased CV function

$$e_1(h) = \frac{5}{6nh} + \frac{1}{12n^2h} \sum_{k=-\infty}^{\infty} [\nu_h(k + 1) - \nu_h(k)]^2. \tag{1.3}$$

The (automatic) CV smoothing parameter minimizes the sample CV function. In Figure 1, we plot e_0 and e_1 for a relatively large sample of 10,000 standard normal points [actually $N(5, 1)$], for which the asymptotic L_2 theory (Scott 1979) predicts that $h = .162$ minimizes the MISE. The difference between these plots is striking. To be sure, most of the "vertical" noise in these plots is due to a bin edge effect. This phenomenon was observed even with much smaller samples by Rudemo (1982). But the difference in noise levels has deeper implications. We claim that these pictures reveal a great deal about the theoretical and practical behavior of these CV techniques for reasonable sample sizes and suggest differences in the "horizontal" noise of smoothing parameters obtained by minimizing the two CV functions. Roughly speaking, in Figure 1b we are seeing the between-sample "vertical" variation because of the relatively small correlation between heights of adjacent or partially overlapping bins. An effort to understand these plots was the motivation for this article.

2. ASYMPTOTIC MEAN INTEGRATED SQUARED ERROR THEORY

Consider a kernel density estimate of an unknown univariate density f based on a random sample $x_1, \ldots, x_n$ with corresponding empirical cdf F_n:

$$\hat{f}(y) = \int K_h(x, y)\, dF_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x_i, y),$$

indexed by a smoothing parameter h. Our goodness-of-fit criterion between f and $\hat{f}$ will be the usual integrated squared error (ISE):

$$\text{ISE} = \int_{-\infty}^{\infty} [\hat{f}(y) - f(y)]^2\, dy. \tag{2.1}$$

Let MISE $\equiv$ E(ISE), the mean integrated squared error.

We shall focus our attention on the fixed bandwidth symmetric kernel estimator

$$\hat{f}(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - x_i}{h}\right). \tag{2.2}$$

Denote by μ_k the "moments" of the kernel K: $\mu_k = \int t^k K(t)\, dt$. If we suppose that $\mu_0 = 1, \mu_1 = \cdots = \mu_{p-1} = 0$,

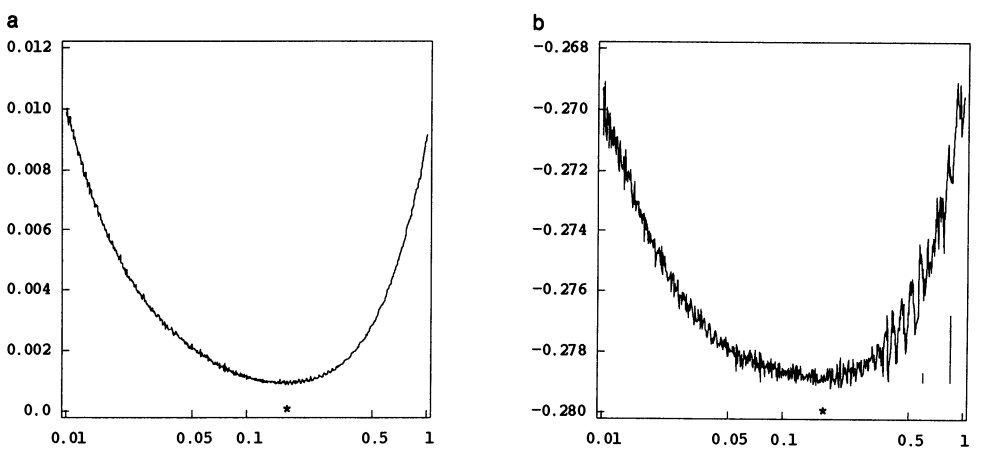

Figure 1. Biased (a) and Unbiased (b) Cross-validation Curves for a Histogram Estimator of 10,000 N(5, 1) Points. The vertical lines in the bottom right corner indicate theoretical standard deviations computed from Theorems 3.1, 3.2, and 3.3 as discussed in the text. The optimal MISE smoothing parameter is indicated by a star.

and $0 < |\mu_p| < \infty$ for some even p, then the MISE may be written as

$$\text{MISE}(h) = \text{AMISE}(h) + O(n^{-1} + h^{2p+1}), \quad (2.3)$$

where the dominant term is the asymptotic MISE (AMISE) given by

$$\text{AMISE}(h) = \frac{R(K)}{nh} + (p!)^{-2}h^{2p}\mu_p^2 R(f^{(p)}), \quad (2.4)$$

where $R(K)$ is defined in (1.1). Expression (2.3) holds if we assume that $R(K) < \infty$, $f^{(p)}$ is absolutely continuous, and $R(f^{(p+1)}) < \infty$ [generalization of Scott (1985)]. The AMISE is minimized when $h^* = O(n^{-1/(2p+1)})$, for example with $p = 2$, by

$$h^* = \{R(K)/[n\mu_2^2 R(f'')]\}^{1/5}. \quad (2.5)$$

We will be comparing several smoothing parameters, and we adopt the following easily recalled notation: for a particular sample, h_{MISE} minimizes MISE, h^* minimizes AMISE, $\tilde{h}_{\text{ISE}}$ minimizes ISE, and $\tilde{h}_{\text{UCV}}$ and $\tilde{h}_{\text{BCV}}$ minimize the unbiased and biased CV functions. Notice that the last three smoothing parameters depend upon the particular set of data.

In this article we focus on nonnegative symmetric kernel estimators and the case $p = 2$. For our CV results, conditions on the kernel and density will be slightly stronger than those given previously. Here we list several sets of conditions, first for the density f and then for the kernel:

Condition 1. f''' absolutely continuous; f^{iv} integrable; $R(f^{iv}(f)^{1/2})$ and $R(f(f^{iv})^{1/2})$ finite.

Condition 2a. $K \geq 0$ symmetric on $[-1, 1]$; K' Holder continuous; $\mu_2 > 0$.

Condition 2b. K'' absolutely continuous on $(-\infty, \infty)$; K''' continuous on $(-1, 1)$; $R(K''') < \infty$.

Throughout this article we use the Gaussian kernel and the triweight kernel, which is defined by

$$K(t) = 35/32(1 - t^2)^3 I_{[-1,1]}(t). \quad (2.6)$$

The triweight kernel is the simplest kernel satisfying Conditions 2a and 2b.

3. CROSS-VALIDATION ALGORITHMS AND THEORY

3.1 Least Squares (Unbiased) Cross-Validation

Ideally, for each sample, we would like to construct a density estimate to minimize the ISE (2.1). Least squares cross-validation attempts to address ISE rather than MISE. We shall introduce the least squares CV criterion for the generalized kernel estimator, which includes most commonly used estimators such as the histogram. Replacing $\hat{f}$ by the generalized estimator $\hat{g}$ in (2.1) and expanding yields

$$\text{ISE} = R(\hat{g}) - 2\int \hat{g}(y)f(y)\,dy + R(f). \quad (3.1)$$

Here,

$$\hat{g}(y) = \frac{1}{n}\sum_{k=1}^{n} K_{nk}(y, x_k), \quad (3.2)$$

where the kernel depends on the sample size n and may also depend on either y or x_k. The idea of Rudemo (1982) and Bowman (1984) is to find data-based expressions that, on average, agree with the first two terms in (3.1) and to omit the third term $R(f)$, which amounts to a simple fixed shift of the entire function. Consider the CV estimator

$$\text{UCV} \equiv R(\hat{g}) - \frac{2}{n}\sum_{i=1}^{n} \hat{g}_{-i}(x_i), \quad (3.3)$$

where

$$\hat{g}_{-i}(x_i) = \frac{1}{n-1}\sum_{k\neq i} K_{nk}(x_i, x_k). \quad (3.4)$$

Notice that the divisor has been changed from n to $n - 1$, but this change is not incorporated into the kernel. Now the expectations of UCV and ISE in (3.1), which is the MISE, match exactly term by term, since

$$\begin{aligned} \text{E}\hat{g}_{-i}(x_i) &= \text{E}K_{nk}(x_i, x_k) = \text{E}\int K_{nk}(y, x_k)f(y)\,dy \\ &= \text{E}\int \hat{g}(y)f(y)\,dy. \end{aligned} \quad (3.5)$$

Hence, in the fixed bandwidth case (2.2), *exactly* unbiased estimates of the shifted MISE for nonrandom h are provided by

$$\text{UCV}(h) = R(\hat{f}) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i). \quad (3.6)$$

We refer to (3.6) as an *unbiased cross-validation* (UCV) criterion because its expectation is

$$\text{E}[\text{UCV}(h)] = \text{MISE}(h) - R(f). \quad (3.7)$$

Other theoretical expressions such as (2.4) are only asymptotically correct; that is, they are biased for finite samples.

It is straightforward to see that (1.2) follows from (3.3), except for replacing $n \pm 1$ by n. Hall (1983) and Stone (1984) showed that the unbiased procedure not only provides a consistent sequence of smoothing parameters but is asymptotically optimal in a certain sense. An additional remarkable feature of this procedure is its self-adapting property. To illustrate this, consider the estimator (2.2), with a symmetric finite-support kernel. Proper analysis of its MISE in (2.3) required knowledge of the "moments" of the kernel, defined following expression (2.2). Such specification is not apparent in (3.6), which, in this case, becomes

$$\begin{aligned} \text{UCV}(h) = \frac{R(K)}{nh} &+ \sum\sum_{i\neq j}\left[\int \frac{1}{n^2h^2} K\left(\frac{x - x_i}{h}\right) K\left(\frac{x - x_j}{h}\right) dx \right. \\ &\left. - \frac{2}{n(n-1)h} K\left(\frac{x_i - x_j}{h}\right)\right]. \end{aligned} \quad (3.8)$$

An instructive exercise is to show that for p even, the expectation of (3.8) equals Equation (2.4) minus the constant $R(f)$. Thus the UCV criterion automatically "knows" the correct order of the kernel. [In fact, Stone (1984)

showed that the method even knows how many derivatives f has for the histogram and nonnegative kernel estimators.] Notice that for large sample sizes, the UCV essentially provides estimates of $R(f^{(p)})$.

In view of Figure 1, the "vertical" variability of UCV(h) is of interest. Assume now that the kernels are symmetric and have finite support on $[-1, 1]$. If we define

$$\gamma(c) = \int K(w)K(w + c)\,dw - 2K(c) \tag{3.9}$$

and let

$$c_{ij} = (x_i - x_j)/h, \tag{3.10}$$

then (3.8) becomes (replacing $n - 1$ by n)

$$\mathrm{UCV}(h) = \frac{R(K)}{nh} + \frac{2}{n^2h}\sum\sum_{i<j} \gamma(c_{ij}). \tag{3.11}$$

The following theorem provides the mean and variance of function (3.11) for fixed h.

Theorem 3.1. For the UCV kernel criterion (3.11)

$$\mathrm{E}[\mathrm{UCV}(h)] = \mathrm{AMISE}(h) - R(f) + O(n^{-1}), \tag{3.12}$$

$$\mathrm{var}[\mathrm{UCV}(h)] = 4/n[R(f^{3/2}) - R(f)^2] + O(1/n^2h + h^4/n). \tag{3.13}$$

The variance of the histogram criterion (1.2) is also given by (3.13).

We prove this in Section 9.1. The rate $O(n^{-1})$ of the leading term in (3.13) was noted by Rudemo (1982). The first term in (3.13) is nonnegative by Jensen's inequality.

Remark. For the example in Figure 1, $(\mathrm{var})^{1/2} = .00222$ for $h = h^* = .162$. This noise, indicated by the longer vertical line in the bottom right corner of Figure 1b, is much greater than the observed noise. We shall return to this point in Section 3.3.

3.2 Biased Cross-Validation

The asymptotic expansion for the MISE as given in (2.4) contains only one unknown quantity, $R(f^{(p)})$. One natural estimator is $R(\hat{f}^{(p)})$, where $\hat{f}$ is a kernel estimator. Scott, Tapia, and Thompson (1977) used this estimator in a fixed-point algorithm for choosing h in the case $p = 2$. The following lemma (proved in Sec. 9.2), however, shows that this estimator is deficient asymptotically and indicates how an improved estimator can be constructed.

Lemma 3.2. Suppose that derivatives of order $p + 2$ of the density f and kernel K exist and are continuous and that $K^{(i)}(\pm 1) = 0$ for $0 \le i \le p - 1$. Then

$$\mathrm{E}[R(\hat{f}^{(p)})] = R(f^{(p)}) + \frac{R(K^{(p)})}{nh^{2p+1}} + O(h^2). \tag{3.14}$$

Notice that for smoothing parameters of the optimal order $h_p^* = c_p n^{-1/(2p+1)}$, the kernel estimate provides a positively biased estimate of $R(f^{(p)})$, but by an asymptotically *constant* amount. Silverman (1978) based his visual "test graph" method for choosing h on another characterization of this asymptotic bias in the L_∞ norm. An improved estimate of $R(f^{(p)})$ is

$$\hat{R}(f^{(p)}) \equiv R(\hat{f}^{(p)}) - \frac{R(K^{(p)})}{nh^{2p+1}}. \tag{3.15}$$

Special Case $p = 2$. For the important case of the nonnegative kernel method when $p = 2$, let

$$\phi(c) \equiv \int K''(w)K''(w + c)\,dw. \tag{3.16}$$

Then, recalling the definition of c_{ij} in (3.10),

$$R(\hat{f}'') = \frac{R(K'')}{nh^5} + \frac{2}{n^2h^5}\sum\sum_{i<j} \phi(c_{ij}).$$

Using this together with the correction (3.15) in the AMISE expression (2.4) defines a BCV function:

$$\mathrm{BCV}(h) \equiv \frac{R(K)}{nh} + \frac{\mu_2^2}{2n^2h}\sum\sum_{i<j} \phi(c_{ij}). \tag{3.17}$$

Notice that the two $R(K'')/nh^5$ terms cancel. Observe the similarities between (3.11) and (3.17); both are U statistics but with different kernels.

In Section 9.3 we prove the following theorem.

Theorem 3.2. For a nonnegative kernel estimator satisfying Conditions 1 and 2b, the estimator BCV(h) is asymptotically normal with mean and variance

$$\mathrm{E}[\mathrm{BCV}(h)] = \mathrm{AMISE}(h) + O(n^{-1}), \tag{3.18}$$

$$\mathrm{var}[\mathrm{BCV}(h)] = \mu_2^4 R(\phi)R(f)/(8n^2h) + O(h/n^2). \tag{3.19}$$

For the histogram CV estimator given by (1.3),

$$\mathrm{var}[e_1(h)] = R(f)/(12n^2h) + O(n^{-2}). \tag{3.20}$$

For $h = O(n^{-1/5})$, var $= O(n^{-9/5})$ in (3.19). It follows from (3.18) and (2.3) that the bias in BCV(h) is $O(n^{-1})$. Thus the squared bias is $O(n^{-2})$, which is of lower order than the variance by the factor h. Hence variance dominates "vertical" mean squared error. Note that the results of Theorems 3.1 and 3.2 are not to comparable orders, since var $= O(n^{-1})$ in (3.13). This discrepancy is resolved in the next section. From (3.20) we may compute $(\mathrm{var})^{1/2} = .0000381$ at $h^* = .162$, which closely approximates the observed variation in Figure 1a, as indicated by the small vertical line in the bottom right corner.

It follows from this theorem that a consistent sequence of smoothing parameters can be found.

Corollary 3.2. Let $\tilde{h}_{\mathrm{BCV}}$ minimize (3.17) over $(0, bh^*)$ for any $b > 1$. Then

$$\operatorname{plim}_{n\to\infty}(\tilde{h}_{\mathrm{BCV}}/h^*) = 1. \tag{3.21}$$

3.3 Unbiased Cross-Validation Revisited

3.3.1. Augmented Unbiased Cross-Validation Criterion. The reason that the variation computed in Theorem 3.1 is not comparable with that of Theorem 3.2 is that Theorem 3.1 measures the vertical variation of the UCV

curve about the level MISE-$R(f)$ rather than the MISE level, which is converging to 0. The vertical variation of the entire curve has no effect on the location of the minimum in which we are interested. Bowman's (1984) method of derivation gave the following augmented UCV(h) formula, which Hall (1983) argued is the correct form for theoretical analysis:

$$\text{AUCV}(h) = R(\hat{f}) - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(x_i) + \frac{2}{n}\sum_{i=1}^{n} f(x_i) - R(f). \quad (3.22)$$

With this change, the variance is of the same order as that in Theorem 3.2.

Theorem 3.3. For nonnegative kernel estimators satisfying Conditions 1 and 2a,

$$\text{var}[\text{AUCV}(h)] = 2R(\gamma)R(f)/(n^2h) + O(h/n^2). \quad (3.23)$$

For the histogram, the variance of (1.2) augmented as in (3.22) is

$$h^2R(f'(f)^{1/2})/n + 2R(f)/(n^2h) + o(n^{-5/3}). \quad (3.24)$$

Corollary 3.3. Let h_{BCV} minimize (3.22) or, equivalently, (3.11) over (ah^*, bh^*) for arbitrarily small a and large b. Then

$$\operatorname{plim}_{n\to\infty}(\tilde{h}_{\text{UCV}}/h^*) = 1.$$

From (3.24) we compute $(\text{var})^{1/2} = .0003394$ at $h^* = .162$, indicated by the smaller vertical line in the bottom right corner of Figure 1b. This closely approximates the observed variation. This standard deviation, however, is 8.9 times larger than for the biased criterion in Figure 1a. Condition 1 is much stronger than those required by Hall and Marron (1985), because of our different approach.

3.3.2. Asymptotic Relative "Vertical" Variability. It is now a simple matter to compare Theorems 3.2 and 3.3 for kernel estimates. The relative variability may be defined by the *square root* of the ratio of the variances (3.23) and (3.19):

$$\text{ratio} = \frac{4}{\mu_2^2}\left[\frac{R(\gamma)}{R(\phi)}\right]^{1/2}, \quad (3.25)$$

which depends only on the kernel. This ratio exceeds 10 for most practical kernels satisfying Condition 2b; see the first part of Table 1. The ratio can be arbitrarily close to 0 for kernels with rough second derivatives. Such kernels should be avoided for BCV.

The extent to which the "vertical" noise is converted to "horizontal" noise is examined theoretically in Section 4 and empirically in Section 6. In terms of ratios, we will show that about half of the "vertical" noise is translated into "horizontal" noise.

4. VARIABILITY AND ASYMPTOTIC NORMALITY OF CV SMOOTHING PARAMETERS

4.1 Unbiased Cross-Validation

Hall and Marron (1985) investigated the variability of $\tilde{h}_{\text{UCV}}$ about the idealized target $\tilde{h}_{\text{ISE}}$ and showed that $\tilde{h}_{\text{UCV}} - \tilde{h}_{\text{ISE}}$ is asymptotically normal (AN). Because (as we will see in Secs. 6 and 7) $\tilde{h}_{\text{UCV}}$ and $\tilde{h}_{\text{ISE}}$ are often negatively correlated, we now compute the variation of $\tilde{h}_{\text{UCV}}$ or, equivalently, of $\tilde{h}_{\text{UCV}} - h^*$. We examine the first derivative of UCV(h) given in (3.11), since the extra terms in the augmented criterion (3.22) do not involve h. Let $\gamma_+(c)$ and $\gamma_-(c)$ define $\gamma(c)$ given in (3.9) on the intervals [0, 2] and [−2, 0], respectively:

$$\gamma_+(c) \equiv \int_{-1}^{1-c} K(w)K(w + c)\,dw - 2K(c), \quad 0 \le c \le 2, \quad (4.1)$$

and $\gamma_-(c)$ as in (4.1) with limits $-1 - c$ and 1. Consider the derivative of $\gamma_+(c_{ij})$:

$$\frac{d}{dh}\gamma_+(c_{ij}) = \frac{-c_{ij}}{h}\int_{-1}^{1-c_{ij}} K(w)K'(w + c_{ij})\,dw + \frac{2c_{ij}}{h}K'(c_{ij}), \quad 0 \le c_{ij} \le 2,$$

where the other term involving the derivative of the upper endpoint in the integral vanishes since $K'(1) = 0$. If we define

$$p(c) = c\int K(w)K'(w + c)\,dw - 2cK'(c), \quad -2 \le c \le 2, \quad (4.2)$$

and 0 elsewhere, then $\tilde{h}_{\text{UCV}}$ satisfies

$$\frac{d}{dh}\text{UCV}(h)\Big|_{h=\tilde{h}_{\text{UCV}}} = 0$$

*Table 1. Asymptotic Ratios of "Vertical" Standard Deviations of UCV and BCV Estimates of Smoothing Parameters**

$K(t) = a_m(1 - t^2)^m$	$R(\gamma)^{1/2}$	$\mu_2^2R(\phi)^{1/2}/4$	Ratio, Eq. (3.25)	$R(\rho)^{1/2}$	$\mu_2^2R(\psi)^{1/2}/4$	Ratio, Eq. (4.16)
$m = 2$ (biweight)	1.0033	.0827	12.13	1.2352	—	—
$m = 3$ (triweight)	1.0737	.0921	11.65	1.2047	.2420	4.98
$m = 4$	1.1337	.1013	11.20	1.2195	.2550	4.78
$m = 5$	1.1859	.1092	10.86	1.2458	.2685	4.64
$N(0, 1)$("$m = \infty$")	.6376	.0715	8.92	.6178	.1558	3.96

* As given in expressions (3.25) and (4.16).

or, equivalently,

$$\sum\sum_{i<j} [\gamma(c_{ij}) + \rho(c_{ij})] \Big|_{h=\tilde{h}_{UCV}} = -nR(K)/2. \quad (4.3)$$

Hall and Marron (1985) showed that the left side (which is a degenerate martingale) is AN. In Section 9.5 we compute the moments and prove the following lemma.

Lemma 4.1. Under Conditions 1 and 2a,

$$\sum\sum_{i<j} [\gamma(c_{ij}) + \rho(c_{ij})] = \text{AN}\{-n^2h^5\mu_2^2R(f'')/2, n^2hR(\rho)R(f)/2\}. \quad (4.4)$$

Now $\text{plim}_{n\to\infty} (\tilde{h}_{UCV}/h^*) = 1$, so we may replace $\tilde{h}_{UCV}$ by h^* in the variance. Hence (4.3) becomes

$$-n^2\tilde{h}_{UCV}^5\mu_2^2R(f'')/2 = \text{AN}\{-nR(K)/2, n^2h^*R(\rho)R(f)/2\}. \quad (4.5)$$

By dividing we have

$$\tilde{h}_{UCV}^5 = \text{AN}\{R(K)/[n\mu_2^2R(f'')], 2h^*R(\rho)R(f) \div [n^2\mu_2^4R(f'')^2]\}. \quad (4.6)$$

But the mean is simply $(h^*)^5$ by (2.5). Hence

$$(\tilde{h}_{UCV}/h^*)^5 = \text{AN}\{1, 2R(\rho)R(f)/[n^2(h^*)^9\mu_2^4R(f'')^2]\}. \quad (4.7)$$

Since the variance $\to 0$ as $n \to \infty$, we may apply the delta method (Serfling 1980, p. 118) with $g(x) = x^{1/5}$, which reduces the variance by the factor 25. Multiplying through by h^*, we have the following theorem.

Theorem 4.1. For a nonnegative kernel estimator satisfying conditions 1 and 2a,

$$\tilde{h}_{UCV} = \text{AN}\{h^*, 2R(\rho)R(f)/[25n^2h^{*7}\mu_2^4R(f'')^2]\}. \quad (4.8)$$

Now set $h^* = c_2n^{-1/5}$; then the standard deviation is given by

$$\sigma(\tilde{h}_{UCV} - h^*) = \frac{2^{1/2}c_2^{-7/2}}{5\mu_2^2R(f'')}[R(\rho)R(f)]^{1/2}n^{-3/10}. \quad (4.9)$$

The relative error of $\tilde{h}_{UCV}$ is $O(n^{-1/10})$.

4.2 Biased Cross-Validation

In a similar fashion we may investigate the limiting distribution of $\tilde{h}_{BCV} - h^*$. Define

$$\psi(c) = c\int K''(w)K'''(w + c)\,dw, \quad -2 \le c \le 2, \quad (4.10)$$

and 0 elsewhere. Then taking the derivative of (3.17), we find

$$\sum\sum_{i<j} [\phi(c_{ij}) + \psi(c_{ij})] \Big|_{h=\tilde{h}_{BCV}} = -2nR(K)/\mu_2^2. \quad (4.11)$$

In Section 9.6 we compute the first two asymptotic moments (AM) of (4.11) and obtain the proof of the following lemma.

Lemma 4.2. Under Conditions 1, 2a, and 2b,

$$\sum\sum_{i<j} [\phi(c_{ij}) + \psi(c_{ij})] = \text{AM}\{-2n^2h^5R(f''), n^2hR(\psi)R(f)/2\}. \quad (4.12)$$

Again $\text{plim}_{n\to\infty} (\tilde{h}_{BCV}/h^*) = 1$, so that we may use h^* in the variance. In a direct fashion we find

$$\tilde{h}_{BCV}^5 = \text{AM}\{(h^*)^5, h^*R(\psi)R(f)/[8n^2R(f'')^2]\}. \quad (4.13)$$

Theorem 4.2. Under the conditions of Lemma 4.2,

$$\tilde{h}_{BCV} = \text{AM}\{h^*, R(\psi)R(f) \div [200n^2(h^*)^7R(f'')^2]\}, \quad (4.14)$$

$$\sigma(\tilde{h}_{BCV} - h^*) = \frac{2^{1/2}c_2^{-7/2}}{20R(f'')} \times [R(\psi)R(f)]^{1/2}n^{-3/10}. \quad (4.15)$$

We remark that we believe it can be shown that $\tilde{h}_{BCV}$ is AN in (4.14).

4.3 Asymptotic Relative "Horizontal" Variability

By comparing the standard deviations in (4.9) and (4.15), we see that the asymptotic relative "horizontal" efficiency defined as the ratio of these standard deviations (not variances) is

$$\text{ratio} = \frac{4}{\mu_2^2}\left[\frac{R(\rho)}{R(\psi)}\right]^{1/2}. \quad (4.16)$$

Compare this to the ratio for the "vertical" noise given in expression (3.25). For the triweight kernel this ratio is 4.98; see Table 1. The usefulness of these results in practice is discussed in the remainder of the article.

5. IMPLEMENTATION WITH GAUSSIAN KERNEL AND AVERAGED SHIFTED HISTOGRAM ESTIMATORS

5.1 Two Introductory Examples

The development thus far requires kernels with finite support. It extends, however, to kernels with exponentially decreasing tails, as is the case with the Gaussian kernel. In this case, which was considered by Rudemo (1982) and Bowman (1984), Equations (3.11) and (3.17) may be expressed in closed form with (again replacing $n \pm 1$ by n) $R(K) = .5/\sqrt{\pi}$, $\sqrt{\pi}\gamma(c) = .5\exp(-c^2/4) - \sqrt{2}\exp(-c^2/2)$, and $32\sqrt{\pi}\phi(c) = (c^4 - 12c^2 + 12)\exp(-c^2/4)$. We plot (3.11) and (3.17) in Figures 2a and 2b for samples of size 25 and 400 from $N(0, 1)$, using data generated by IMSL routine GGNPM with seeds 1821291829 and 1943248741, respectively. For plotting purposes, we have augmented UCV(h) as in Equation (3.22). The dotted line is the (exact) MISE (Fryer 1976).

For fixed n the BCV function converges to 0 as $h \to \infty$. The BCV function barely exhibits a local minimum with

$n = 25$ (sometimes it has none; see Sec. 6), but exhibits a clear local minimum when $n = 400$. Heuristically, the BCV indicates the quality of $\tilde{h}_{\mathrm{BCV}}$ by the amount of rise to the right of the minimum. As n increases, BCV provides reasonable estimates of MISE for relatively larger values of $h > h_{\mathrm{MISE}}$, where the MISE is increasingly dominated by bias.

The UCV function does relatively well in the high bias region and less well in the high variance region, $h < h_{\mathrm{MISE}}$, as predicted by Theorem 3.3. There is no high frequency component evident in individual plots as was the case in Figure 1a for the histogram, since we are using a continuous kernel. With $n = 400$ we have selected a case where the UCV function has a minimum well to the left of h_{MISE}; see Section 6.1. (The minima in Figure 2b are .142, .330, and .389.) Rudemo, in a draft of his 1982 paper, observed this (occasional) behavior for smaller samples and speculated that it was consistent with features in the data. In Figure 2c we plot the two CV estimates along with the true density (the h_{MISE} estimate is quite similar to the BCV estimate). The density estimate reveals the illusory multimodal feature that attracted the UCV function. The UCV function also eventually converges to 0 [the augmented version to approximately $R(f)$]; the curve in Figure 2b increases monotonically to $-.579$.

For $n = 400$, evaluating (3.11) and (3.17) for each h took more than 1.1 CPU minutes on a VAX 11/750. Figure 2b required several hours of CPU time. Clearly an alternative implementation is required for even moderate sample sizes.

5.2 Averaged Shifted Histogram Implementation

To carry out an extensive Monte Carlo study, it is necessary to find a more computationally feasible method than the very slow Gaussian kernel implementation given previously. Much faster evaluations of (3.11) and (3.17) are possible with finite support kernels. Furthermore, a kernel procedure using binned data accelerates CV algorithms even more, for example, Silverman's (1982) fast Fourier transform algorithm. Another procedure that takes advantage of binned data is the averaged shifted histogram (ASH) (Scott 1985). An ASH is the (weighted) average of m histograms, each with bin width h but with bin mesh origins at integer multiples of $\delta \equiv h/m$, and is given by

$$\hat{f}_m(y) = \frac{1}{nh} \sum_{i=1-m}^{m-1} w_m(i) v_\delta(k + i) \qquad \text{for } y \text{ in } I_k, \quad (5.1)$$

where $w_m(i)$ are the weights and $v_\delta(k)$ is the bin count for the kth bin $I_k \equiv [k\delta, (k + 1)\delta]$. The weights corresponding to the triweight kernel (2.6) are

$$w_m(i) = c_m[1 - (i/m)^2]^3 \qquad \text{for } |i| < m, \quad (5.2)$$

where c_m is a normalizing constant, so $\sum w_m(i) = m$ given by

$$c_m = 35/[32(1 - 1/4m^2)(1 + 1/4m^2 + 5/24m^4)].$$

The UCV formula (3.3) for $\hat{f}_m$ is easily evaluated. The term $R(\hat{f})$ is computed directly. The term $\sum_i (\hat{f}_m)_{-i}(x_i)$ in (3.3) and (3.4) is simply equal to

$$\sum_{k=-\infty}^{\infty} v_\delta(k)\hat{s}_k - \frac{w_m(0)}{h}, \quad (5.3)$$

where $\hat{s}_k \equiv \hat{f}_m(k\delta)$ is the value of $\hat{f}_m$ in I_k. In practice the sum in (5.3) involves perhaps a few hundred terms. For $m > 10$ (i.e., δ sufficiently small) the behavior of the kernel and ASH estimators is virtually identical; in particular, similar values of the smoothing parameter h give nearly identical results.

For BCV, the asymptotic theory for the ASH involves both $R(f')$ and $R(f'')$, which is unfortunate. The frequency polygon (linear interpolator) of the ASH (FPASH), however, requires only $R(f'')$. We cannot use binned data with UCV on FPASH, since we would need to know $\hat{g}_{-i}(x_i)$; that is, we would need to know all the x_i exactly and not just $\hat{s}_k$ (or equivalently, the bin counts) as in the ASH case. Again we emphasize that for $m > 10$, the ordinary kernel, ASH, and FPASH are essentially the same for the same h.

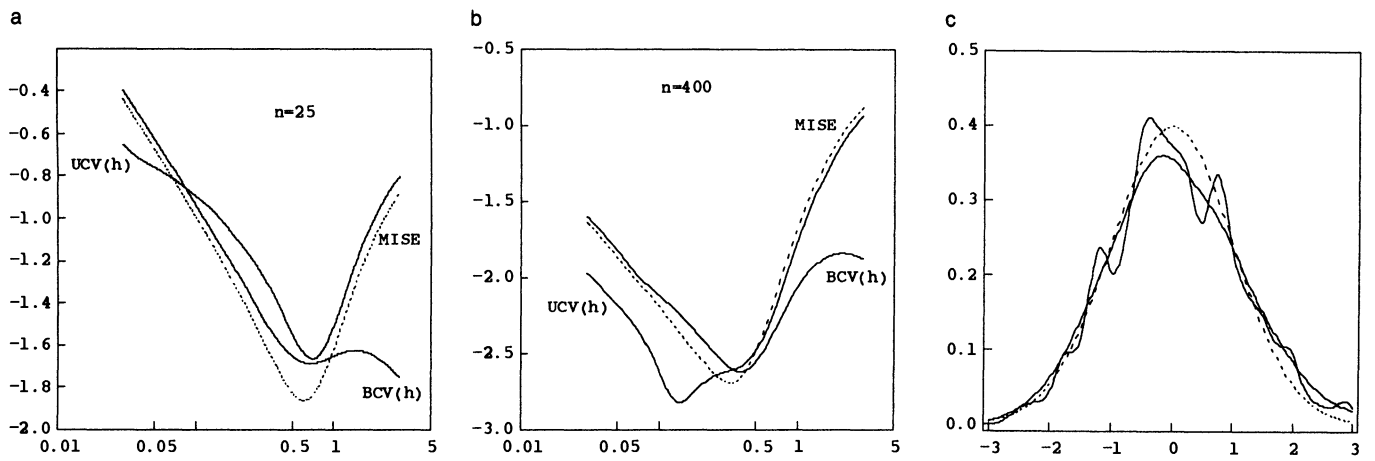

Figure 2. Examples of Biased and Unbiased Cross-validation Curves ($\log_{10}$ scale) for a Gaussian Kernel Estimator of 25 N(0, 1) Points in (a) and 400 points in (b). The exact mean integrated squared error is shown by the dotted line. The corresponding cross-validation density estimates are shown in (c), along with the true density (dotted line).

The asymptotic MISE expression for the FPASH is (Scott 1985)

$$\text{AMISE} = \frac{2R_w + \gamma_w}{3nh} + \frac{1}{4}h^4 \times \left[\sigma_w^4 + \frac{\sigma_w^2}{2m^2} + \frac{49}{720m^4}\right] R(f''),$$

where $mR_w = \Sigma\, w_m(i)^2$, $m\gamma_w = \Sigma\, w_m(i)w_m(i-1)$, and $m^3\sigma_w^2 = \Sigma\, i^2 w_m(i)$. Our estimate of $R(f'')$ turns out to be

$$\frac{1}{\delta^3}\sum_k (\hat{s}_{k+1} - 2\hat{s}_k + \hat{s}_{k-1})^2 - \frac{m^3}{nh^5}\left[2w_{m-1}^2 + 4(w_0 - w_1)^2 + 2\sum_{i=1}^{m-1}(w_{i+1} - 2w_i + w_{i-1})^2\right], \quad (5.4)$$

where we denote $w_m(i)$ by w_i and $\hat{s}_k$ is defined following Equation (5.3). These may be computed in closed form using MACSYMA (the triweight kernel formulas are available from us).

6. MONTE CARLO STUDY

6.1 Small-to-Large Sample Behavior With Gaussian Data

In this section we study the results of simulations based on samples from a standard normal distribution for sample sizes n = 25, 100, 400, 1,600, 6,400, and 25,600 with repetitions of 250, 200, 150, 100, 100, and 100, respectively. ASH and FPASH estimators with a triweight kernel were used as described previously. The δ's chosen were .15, .10, .05, .025, .02, and .01, respectively. For each sample, ISE's corresponding to four different bandwidths h (or, equivalently, m since $h = m\delta$) were computed numerically: h_{MISE}, $\tilde{h}_{\text{ISE}}$, $\tilde{h}_{\text{BCV}}$, and $\tilde{h}_{\text{UCV}}$. The value $\tilde{h}_{\text{ISE}}$, which minimizes the ISE for a particular sample, was found by searching over integer values m.

In Figure 3 we plot frequency polygons of the cross-validated smoothing parameters. The vertical lines indicate the set of h's examined (multiples of δ). In Table 2 we present some summary statistics. We note immediately that in 103/250 samples with n = 25, the BCV function had no local minima (compare Fig. 2a). The average of $\tilde{h}_{\text{UCV}}$ when n = 25 is reasonable, but only a relatively few

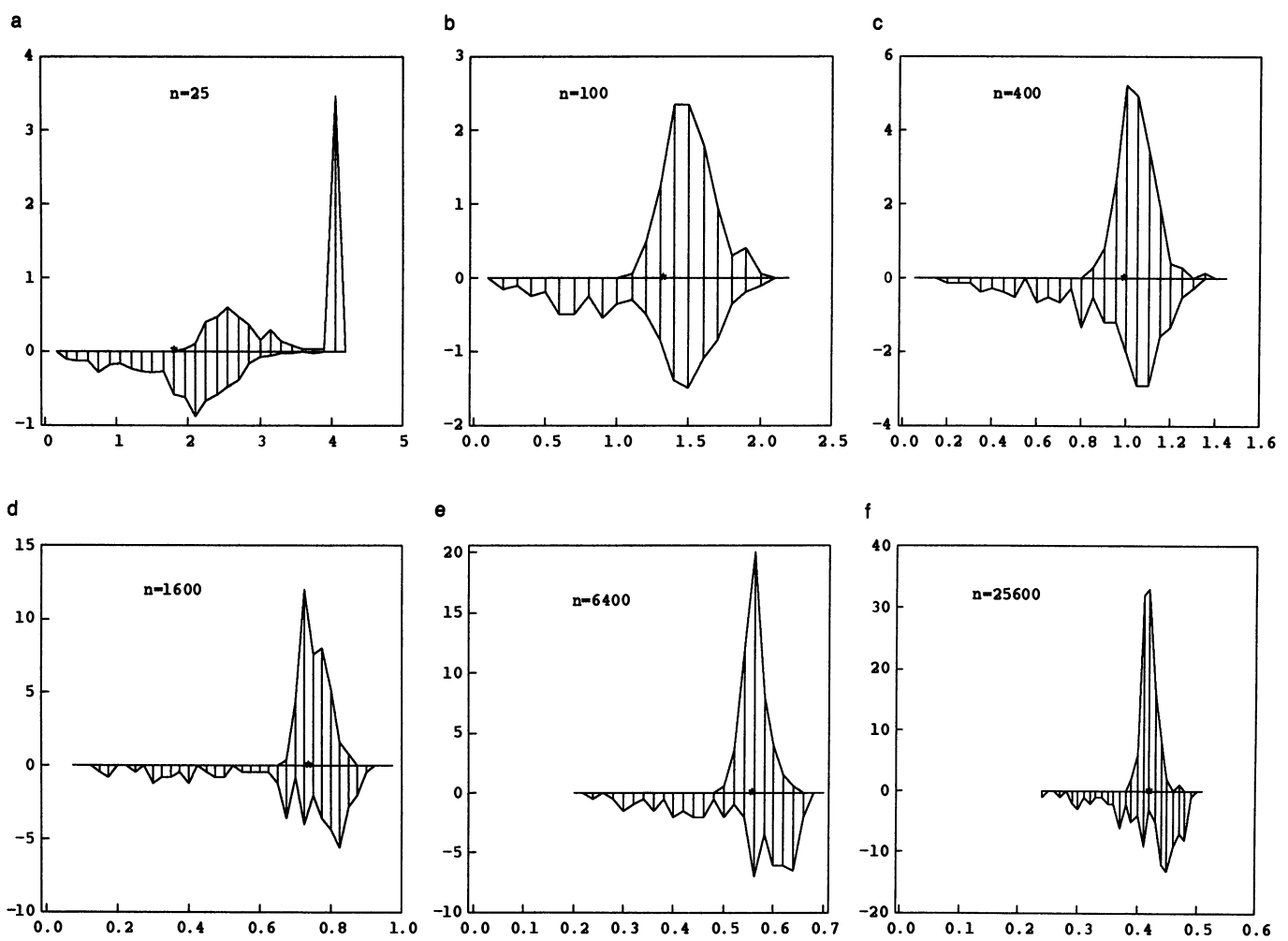

Figure 3. Histograms of Biased and Unbiased Cross-validation Smoothing Parameters for N(0, 1) Samples of Several Sizes Using an ASH Triweight Kernel Estimator. The BCV parameter histogram is in the positive direction, and the UCV histogram is in the negative direction. The location of h_{MISE} is indicated by a star on the horizontal axis.

Table 2. Monte Carlo Results of Triweight Kernel ASH Estimates of N(0, 1) Data

n	h_{MISE}	$\bar{h}_{BCV}$	$\bar{h}_{UCV}$	$\hat{\sigma}_{BCV}$	$\hat{\sigma}_{UCV}$	Ratio	σ_{BCV}	σ_{UCV}
25	1.775	—	1.907	—	.6700	—	.0951	.4732
100	1.309	1.499	1.262	.1691	.4170	2.47	.0627	.3122
400	.976	1.041	.935	.0792	.2422	3.06	.0414	.2060
1,600	.732	.753	.683	.0372	.1862	5.00	.0273	.1359
6,400	.552	.561	.535	.0246	.1054	4.27	.0180	.0896
25,600	.416	.419	.416	.0128	.0549	4.28	.0119	.0591

NOTE: The sample means and variances of the CV smoothing parameters are given, together with the theoretical standard deviations given in Theorems 4.1 and 4.2. The theoretical predictions of the standard deviations of $\tilde{h}_{BCV}$ and $\tilde{h}_{UCV}$ are denoted by σ_{BCV} and σ_{UCV}, and sample versions are indicated by a circumflex.

individual samples are close to h_{MISE}. (Of course, perhaps $\tilde{h}_{ISE}$ is not close. We check this later.) We have not found any samples where the BCV failed to have a local minimum for $n > 40$. (For other densities this threshold is higher.) On the other hand, the variance of the biased CV estimates drops dramatically beyond this threshold, so the "worst" case for $n \geq 1{,}600$ is quite close to h_{MISE} (a reasonable target, as we discuss in Sec. 7). The unbiased CV procedure continues to be attracted to spurious (rough) estimates even with $n = 25{,}600$. Its convergence to normality is also apparently slower. The asymptotic theory predicts a ratio of "vertical" standard errors of the CV curves of 11.65 (which was observed in the simulations) and a ratio of "horizontal" standard errors of CV smoothing parameters of 4.98; see Table 1. In Table 2 we see that the finite sample ratio is reasonably close to 4.98 for moderate sample sizes and that expressions (4.15) and (4.9), which yield $\sigma(\tilde{h}_{BCV}) = .250n^{-3/10}$ and $\sigma(\tilde{h}_{UCV}) = 1.243n^{-3/10}$, are remarkably accurate.

A more detailed study of the individual results for $n = 400$ and $n = 25{,}600$ is worthwhile. In Figure 4 we plot the various smoothing parameters for $n = 25{,}600$ ($n = 400$ is similar). Surprisingly, there is a negative correlation of $-.38$ between $\tilde{h}_{ISE}$ and $\tilde{h}_{UCV}$; see Section 7. The $\tilde{h}_{BCV}$ cluster more tightly around h_{MISE} (the correlation with $\tilde{h}_{ISE}$ is $-.16$). For the 150 repetitions with $n = 400$, 41 had $\tilde{h}_{UCV} \leq .85$ (.85 was the smallest observed $\tilde{h}_{BCV}$ value). In 23 of 150 samples the UCV curve had two minima, always one less and one greater than .85. Seven of these had a more reasonable local minimum near h^*. Sixteen (all $\leq .85$) were local minima compared with a reasonable $\tilde{h}_{UCV}$ near h^*. When $n = 25{,}600$, only 2 of 100 UCV curves had a second (local) minimum, but in both cases the global minimizer was more reasonable. None of the BCV curves had any other local minima over the range searched.

In Figure 5 the numerically computed ISE's of the samples with $n = 400$ are displayed. Figure 5a indicates that h^* is only occasionally grossly inefficient relative to $\tilde{h}_{ISE}$. In Figure 5b we see that the BCV almost dominates the UCV estimates with respect to ISE! Figure 5c is presented for completeness.

Using the Hall and Marron (1985) formulas for the triweight kernel and Gaussian data, we obtain $\sigma(\tilde{h}_{ISE}) = 1.304n^{-3/10}$ and $\sigma(\tilde{h}_{UCV} - \tilde{h}_{ISE}) = 2.081n^{-3/10}$. Since $\sigma(\tilde{h}_{UCV}) = 1.243n^{-3/10}$, it follows that there is indeed a negative correlation between $\tilde{h}_{UCV}$ and $\tilde{h}_{ISE}$. With the data above, we computed the sample version of $\sigma(\tilde{h}_{UCV} - \tilde{h}_{ISE})$ as .3464 and .0918 for $n = 400$ and $n = 25{,}600$, respectively, which agree closely with the theoretical predictions of .3448 and .0990. Thus, although the variability of $\tilde{h}_{UCV}$ and $\tilde{h}_{ISE}$ is similar, the negative correlation suggests that they are often on opposite sides of h_{MISE}. We have seen how $\tilde{h}_{BCV}$, which is very close to h_{MISE} for large samples, generally corresponds to estimates with ISE's smaller than using $\tilde{h}_{UCV}$.

6.2 Other Densities

Similar simulations were performed for three other densities: Cauchy, lognormal (exponential of standard

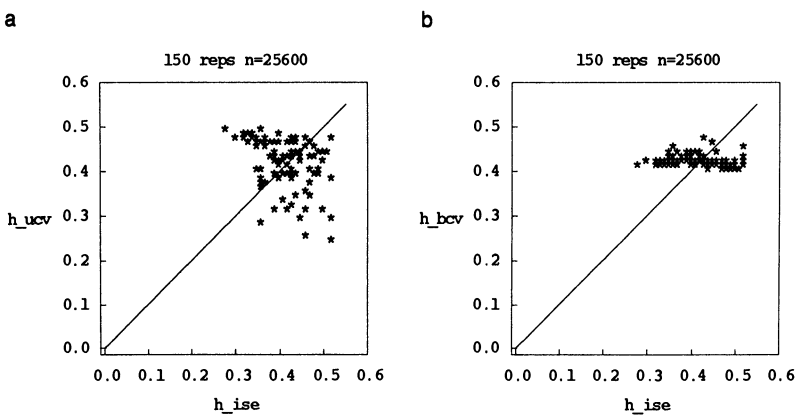

Figure 4. Scatterplots of the Various Smoothing Parameters are Shown for the Same Monte Carlo Data as in Figure 3 With Sample Size $n = 25{,}600$.

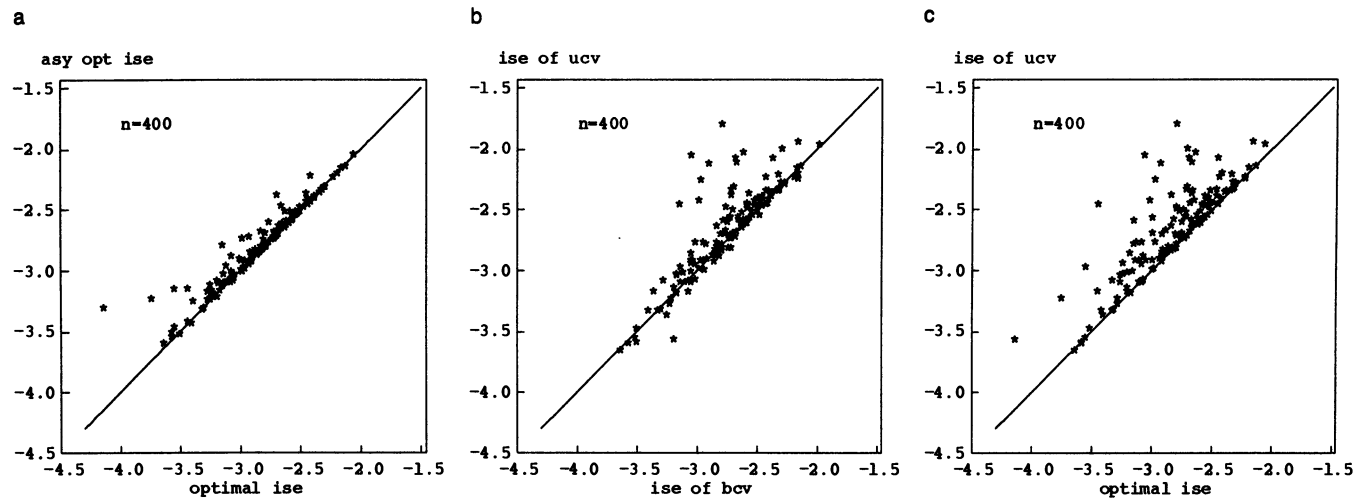

Figure 5. Using the Same Samples With n = 400 as in Figure 3, Scatterplots of the $\log_{10}$(ISE) Corresponding to the Various Smoothing Parameters Are Shown. The diagonal line is y = x.

Gaussian random variable), and a mixture given by

$$f(x) = .75\phi(x, 0, 1) + .25\phi(x, 2, \tfrac{1}{9}),$$

where $\phi(x, \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2. In Figure 6 we plot histograms of the CV estimates for 100 repetitions with $n = 1{,}600$. The Cauchy simulations ($25 \le n \le 25{,}600$) were similar to the Gaussian results in Section 6.1 except that 17% of the BCV estimates failed to exist for $n = 100$ and the ratios of "horizontal" standard errors increased to only 3.0 (see Table 3).

The lognormal and mixture results are interesting. Notice that the BCV estimates are shifted to the right from the UCV estimates. BCV failures were observed at $n = 400$. The UCV continued to perform as usual: average behavior close to h_{MISE} with high variability. The BCV (when it existed) was definitely biased upward for moderate sample sizes, although the bias vanishes by $n = 25{,}600$. We understand this phenomenon as follows: for small samples the estimates that are optimal with respect to ISE appear relatively rough or noisy. This is not a defect of L_2 error but of the use of a nonadaptive estimator.

We examine a particular example. Figure 7 is a plot of ASH estimates of a mixture sample with $n = 400$ with $\delta = .015$. For this sample, $h_{\text{MISE}} = .615$, $\tilde{h}_{\text{ISE}} = .510$, $\tilde{h}_{\text{UCV}} = .480$, and $\tilde{h}_{\text{BCV}} = .870$ (with ISE greater by 55%). L_1 and L_∞ errors are minimized for $h = .525$ and .540, respectively. The UCV estimator is best in the narrow peak, and the BCV is better in the larger peak. If we insist on a fixed bandwidth estimator, Figure 7a might be preferable [consistent with earlier recommendations of Fryer (1976) to slightly exceed h^*]. On the other hand, for small samples, occasionally large $\tilde{h}_{\text{BCV}}$'s obscure the bimodal feature.

7. EXAMPLES WITH REAL DATA

Good and Gaskins (1980) presented a large particle physics data set (the LRL data), which is interesting be-

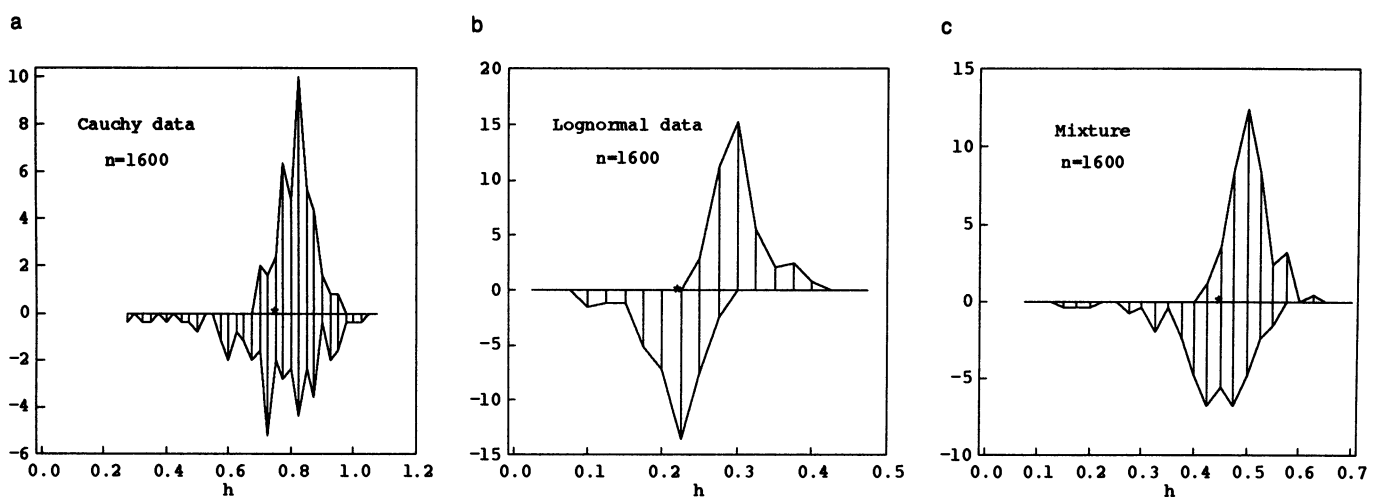

Figure 6. Similar to Figure 3, Except With n = 1,600 From Cauchy, Lognormal, and Mixture Densities.

Table 3. Partial Monte Carlo Results for Three Sampling Densities

Density	n	h_{MISE}	$\bar{h}_{BCV}$	$\bar{h}_{UCV}$	$\hat{\sigma}_{BCV}$	$\hat{\sigma}_{UCV}$	Ratio	σ_{BCV}	σ_{UCV}
Cauchy	400	1.012	1.230	1.056	.1292	.2538	1.96	.0300	.1492
	1,600	.740	.815	.751	.0547	.1448	2.65	.0198	.0984
	6,400	.549	.580	.551	.0263	.0862	3.28	.0130	.0649
	25,600	.411	.418	.415	.0144	.0371	2.57	.0086	.0428
Lognormal	400	.324	.540	.326	.1052	.0776	.74	.0050	.0248
	1,600	.218	.302	.212	.0331	.0402	1.21	.0033	.0163
	6,400	.151	.184	.150	.0137	.0209	1.52	.0022	.0108
	25,600	.107	.121	.107	.0048	.0127	2.63	.0014	.0071
Mixture	400	.612	—	.618	—	.1512	—	.0167	.0830
	1,600	.443	.504	.434	.0374	.0749	2.00	.0110	.0548
	6,400	.327	.345	.320	.0155	.0425	2.75	.0073	.0361
	25,600	.245	.252	.242	.0068	.0294	4.34	.0048	.0238

NOTE: The theoretical predictions of the standard deviations of $\bar{h}_{BCV}$ and $\bar{h}_{UCV}$ are denoted by σ_{BCV} and σ_{UCV}, and sample versions are indicated by a circumflex. This table continues Table 2 with three additional sampling densities.

cause it is prebinned with $\delta = 10\ MeV$. The authors found 13 bumps in a penalized likelihood estimate. The optimal bin width using either histogram criteria (1.2) or (1.3) gives $h = 10\ MeV$ as optimal. We also examined these data with a triweight ASH estimator. In this case $m = 2$ for UCV and $m = 4$ for BCV using the ASH and FPASH, respectively. The square roots of these estimates are shown in Figure 8. Although the 13 bumps found by Good and Gaskins are apparent in Figure 8c, it is interesting to speculate why certain small bumps are included and others excluded. It is appropriate to recall that an optimally smoothed density has a slightly noisy second derivative, as shown in Equation (3.15) when $p = 2$.

The 15,000 steel surface data points were reanalyzed in Scott (in press). The UCV curve is constant over a surprisingly wide interval, a behavior that has not been previously observed.

We have implemented a bivariate product kernel BCV algorithm. Details and an example with a data set (thought to have a bimodal density) of 320 males with heart disease are available from us. The bimodal feature was not revealed by a BCV estimate in the same manner observed in the univariate mixture example in Section 6.2.

8. DISCUSSION

8.1 Achieving Optimal ISE in Cross-Validation

One issue that recurs is whether we should prefer a smoothing parameter that minimizes MISE or whether we should minimize the ISE for the data at hand. In theory, we should address ISE. The ISE for individual samples may be decreased primarily in two ways. The first is to use a variable bandwidth estimator, although adaptive cross-validation is more delicate because of an increase in number of smoothing parameters. The second is to compensate for variation in the lower-order sample moments. Although no choice of smoothing parameter can compensate for a shift in mean, it is possible to reduce ISE due to fluctuations in the sample variance. If $\hat{\sigma} < \sigma$, then choose $h > h^*$ and vice versa, which unfortunately requires knowledge of the unknown density. Thus we do not expect much, if any improvement for CV methods attempting to minimize ISE compared with those seeking the bandwidth minimizing MISE.

It is easy to demonstrate this observation by simulation with standard Gaussian data. In Figure 9 we plot $\tilde{h}_{ISE}$ versus $\hat{\sigma}$ for the 100 repetitions with $n = 25{,}600$ used in

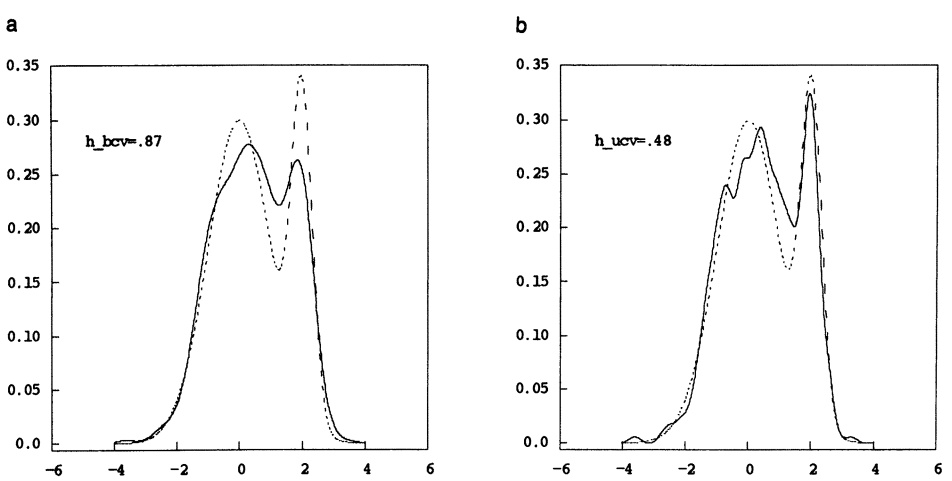

Figure 7. Biased and Unbiased Cross-Validation Density Estimates of 400 Points From a Mixture Distribution.

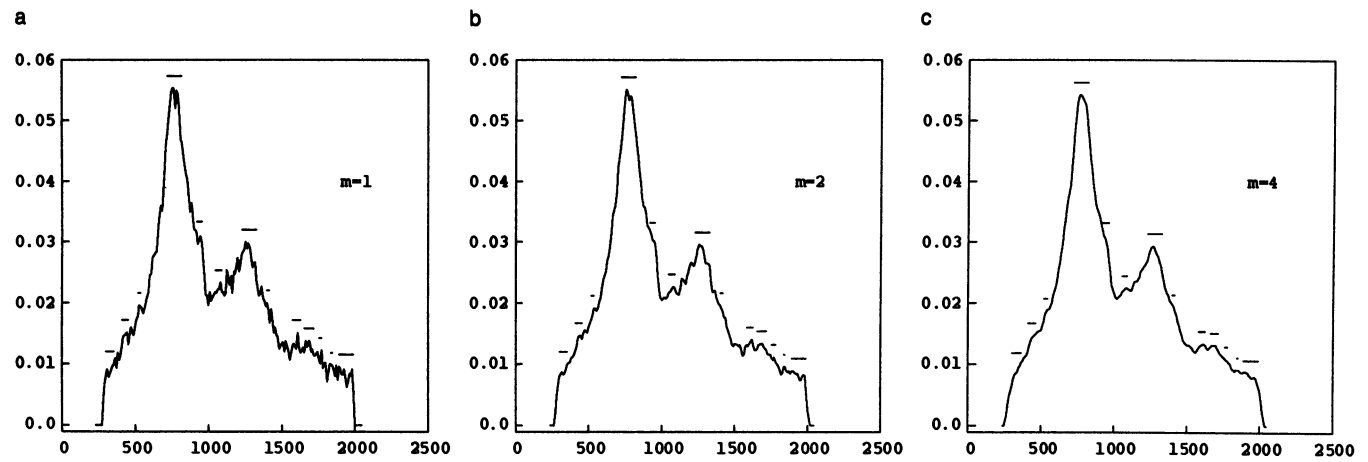

Figure 8. On a Square Root Scale, Triweight ASH Estimates of the LRL Data, With m = 1, 2, and 4. The bumps found by Good and Gaskins (1980) with a penalized-likelihood density estimator are indicated by horizontal lines above the bump.

Section 6. The correlation between $\bar{h}_{ISE}$ and $\hat{\sigma}$ is $-.688$. Shifting these samples to have 0 sample mean did not change these correlations much. Thus we see that any benefits to be gained from minimizing ISE rather than MISE are swamped by the much larger asymptotic error of the algorithms that pursue the former goal.

8.2 Partial Explanation for Improvement of BCV Over UCV

The nature of the improvement is most easily seen with the histogram, for which

$$\text{AMISE}(h) = \frac{1}{nh} + \frac{1}{12} h^2 R(f') \tag{8.1}$$

and $h^* = O(n^{-1/3})$. Following (1.2) and (1.3), consider a third estimate of (8.1):

$$e_2(h) = \frac{23}{24nh} + \frac{1}{48n^2h} \sum [\nu_h(k+1) - \nu_h(k-1)]^2. \tag{8.2}$$

Now $e_2(h)$ is based on a central difference approximation to $R(f')$, which is numerically superior to the forward difference approximation leading to $e_1(h)$. It may be shown that the "vertical" variances of e_0, e_1, and e_2 are $2R(f)/n^2h + O(h^2/n)$, $R(f)/12n^2h$, and $R(f)/192n^2h$, respectively. Again the squared bias is of lower order $O(n^{-2})$. This is a remarkable decrease in the variances. But for finite samples the use of higher-order derivative approximations will incur large bias and hence the gains are not realized except for extremely large samples. This is similar to the choice of p in Equation (2.4). Theory suggests choosing p as large as possible, whereas in practice $p = 2$ or 3 is a wiser choice. The higher-order terms cannot in general be neglected. But for moderate samples $e_1(h)$ does represent a substantial improvement over $e_0(h)$, whereas $e_2(h)$ may not.

8.3 Early UCV and BCV Algorithms

Kronmal and Tarter (1968) introduced the first UCV algorithm for a Fourier series density estimator. The algorithm provided unbiased estimates of the change in MISE as additional Fourier coefficients were introduced. Wahba (1981) had the first working BCV algorithm, which she called generalized CV. In her Fourier series estimator, the smoothing parameter is not the number of terms in the series but a shape parameter in a tapering window applied to the Fourier coefficients f_ν. By substituting unbiased sample estimators for f_ν and $|f_\nu|^2$, she derived a biased cross-validation criterion, with the (small) bias due to truncation of the series in ν. Wahba's and Kullback–Liebler methods were tested by Scott and Factor (1981). Our biased CV algorithm is essentially the analog of Wahba's procedure. We suspect, however, that Wahba's procedure is less biased than our BCV approach.

8.4 Conclusions

We have attempted to evaluate the small-sample properties and reliability of two CV algorithms. No currently

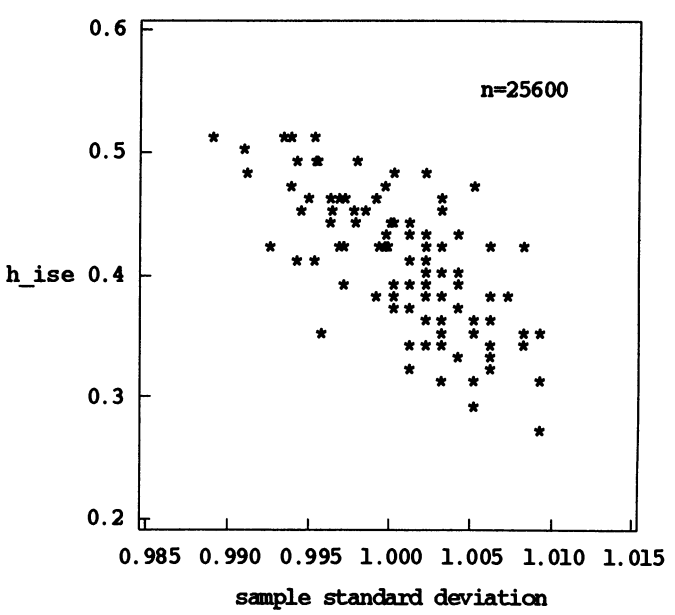

Figure 9. For the Samples in Figure 4, Scatterplots of $\bar{h}_{ISE}$ and the Sample Standard Deviation for Each Sample.

available algorithm is highly reliable for very small samples. In this situation BCV always oversmooths and UCV has a very large variance. For "large" samples, however, cross-validation is highly reliable with respect to MISE. Reliability with "medium" samples is often achieved with densities that are not too rough. From Tables 2 and 3 we see that our definition of a highly reliable CV algorithm is satisfied by the BCV estimates for sample sizes beyond 500–1,000 except for the lognormal density, which requires several thousand points. The goal of finding $\tilde{h}_{\text{ISE}}$ with CV algorithms remains largely unsolved, as pointed out in Section 6.1. It is not at all clear that using $\tilde{h}_{\text{ISE}}$ is to be preferred to h_{MISE}, given the rather peculiar manner by which the integrated squared error is further reduced.

We also find that CV performance depends strongly on sample size and the underlying density. Specifically, the conditional probability that the CV smoothing parameter is "acceptable" given n increases rather rapidly from 10% to 90%; however, the location of this transition region may begin with surprisingly large sample sizes. Further work characterizing this transition would be interesting. With finite samples we are limited in our ability to adequately estimate all densities; clearly, though, we are in a stronger position than if we made a parametric choice.

Although BCV has potentially greater reliability than UCV, it comes at the cost of additional assumptions on f. The very general optimal consistency of UCV, however, comes at a surprisingly high cost in sample size requirements if f is smooth. Asymptotically, about $(4.98)^{10/3}$ or about 211 times more points are required so that $\sigma(\tilde{h}_{\text{UCV}})$ equals $\sigma(\tilde{h}_{\text{BCV}})$ for the triweight kernel. Thus we have a tension between "customized" and "generic" CV. It would be interesting to investigate how much UCV can be improved perhaps, for example, by leaving more than one point out.

Perhaps most useful is to observe the divergence in behavior of UCV and BCV algorithms. Agreement or disagreement of the two CV parameters provides possible auxiliary information about any unusual features in the underlying density. BCV is essentially using the data to estimate the bias. This is (and should be) a difficult task because the relative contribution of the bias and variance toward the MISE is in a ratio of 1:4 near optimal smoothing. UCV provides superior bias estimates but at the cost of increased variance. Given the importance of variance at $h = h_{\text{MISE}}$, it is important to control "vertical" variance more than current UCV algorithms do.

We observe that the BCV procedure may be used to obtain approximate confidence intervals for both $\tilde{h}_{\text{UCV}}$ and $\tilde{h}_{\text{BCV}}$, assuming that the latter is asymptotically normal. BCV provides consistent estimates of $R(f'')$ as well as $R(f)$, which may be used in (4.9) and (4.15). In fact, Theorem 3.2 follows from the fact that $\hat{R}(f'') = \text{AN}\{R(f''), 2R(\phi)R(f)/(n^2h^9)\}$. Some idea of the reliability of the CV smoothing parameter can be drawn from these estimates. In addition, BCV will provide useful estimates of the MISE via expression (3.17).

For sufficiently large data sets and reasonable densities, reliability is achievable. We wish to emphasize that excellent density estimates are still possible with smaller samples but cannot be reliably calibrated by present methodology. We believe superior UCV and BCV kernel estimators can be found, since neither development attempted to optimize reliability. A referee (and others) have suggested investigating a linear combination of UCV and BCV. We prefer having these two relative different estimators to work with, partly because of different density requirements. Perhaps the more computationally intensive bootstrap methods can be used to improve reliability for small samples.

Finally, we remark that there are many other nonparametric applications where cross-validation is desirable, such as nonparametric regression, discrimination, hazard analysis, and spectral analysis. It would be interesting to see how BCV and UCV algorithms compare in these settings.

9. PROOFS OF RESULTS

9.1 Proofs of Theorems 3.1 and 3.3

We assume that Conditions 1 and 2a are satisfied. Occasionally in the proofs we tacitly assume the existence of higher-order derivatives in f when we wish to investigate explicitly error terms; however, these derivatives are not required.

9.1.1. Expectation of the UCV Function. Although this result was proved in Section 3.1, we give a different proof here to indicate the care required when computing expectations of the U statistics. Recall the definitions of γ_+ and γ_- in Equation (4.1). Since K is symmetric, it follows that γ is symmetric as well. Now

$$\begin{aligned}
\text{E}\gamma(c_{ij}) &= \int_{-\infty}^{\infty} f(x) \left[\int_{x-2h}^{x} \gamma_+ \left(\frac{x-y}{h} \right) f(y)\, dy \right. \\
&\qquad \left. + \int_{x}^{x+2h} \gamma_- \left(\frac{x-y}{h} \right) f(y)\, dy \right] dx \\
&= h \int_x f(x) \left[\int_0^2 \gamma_+(c) [f(x - hc) \right. \\
&\qquad \left. + f(x + hc)]\, dc \right] dx \\
&= 2h \int_0^2 \gamma_+(c) \left[\sum_{k=0}^{3} \frac{(-1)^k}{(2k)!} - (ch)^{2k} R(f^{(k)}) \right. \\
&\qquad \left. + o(h^6) \right] dc. \qquad (9.1)
\end{aligned}$$

For k even, it is not hard to show that

$$\begin{aligned}
&\int_0^2 c^k \gamma_+(c)\, dc \\
&\quad = \frac{1}{2} \int_{-1}^{1} K(w) \int_{-1}^{1} (s - w)^k K(s)\, ds\, dw - \mu_k, \qquad (9.2)
\end{aligned}$$

which equals $-\frac{1}{2}$, 0, $3\mu_2^2$, and $15\mu_2\mu_4$ for $k = 0, 2, 4, 6$, respectively. The result follows, since $\text{E}\gamma(c_{ij}) = -hR(f) + \frac{1}{4}\mu_2^2 h^5 R(f'') + O(h^7)$.

9.1.2. Variance of the UCV Function. Computing $E\gamma(c_{ij})^2$ parallels (9.1) with $\gamma_+(c)^2$ replacing $\gamma_+(c)$, and $E\gamma(c_{ij})^2 = hR(\gamma)R(f) + O(h^3)$. Hence $\operatorname{var}\gamma(c_{ij}) = hR(\gamma)R(f) - h^2R(f)^2 + O(h^3)$. For simplicity of notation, let $\gamma_{ij} \equiv \gamma(c_{ij})$. Now $\operatorname{cov}(\gamma_{ij}, \gamma_{kl}) = 0$; here (and from now on) we assume that distinct letters represent unequal subscripts. Let

$$I_1 \equiv \int f^{iv}(x)f(x)^2\,dx; \qquad I_2 \equiv R(f)R(f'');$$
$$I_3 \equiv R(f^{3/2}) - R(f)^2.$$

Taylor series approximations in the 3-D integral lead to $\operatorname{cov}(\gamma_{ij}, \gamma_{ik}) = \operatorname{cov}(\gamma_{ij}, \gamma_{ki}) = h^2I_3 - \mu_2^2h^6[I_1 - I_2]/2 + O(h^7)$. Using the well-known result that

$$\operatorname{var}\left[\sum\sum_{i<j}\gamma_{ij}\right] = \frac{1}{2}n(n-1)\operatorname{var}\gamma_{ij} + n(n-1)(n-2)\operatorname{cov}(\gamma_{ij}, \gamma_{ik}), \quad (9.3)$$

we complete our proof of Theorem 3.1 and explicitly give the remainder term in (3.13):

$$\operatorname{var}\operatorname{UCV}(h) = \frac{4}{n}I_3 + \frac{2R(\gamma)R(f)}{n^2h} + \frac{2\mu_2^2h^4}{n}[I_2 - I_1] + o(n^{-9/5}). \quad (9.4)$$

9.1.3. Variance of Augmented UCV Criterion. Comparing (3.11) and (3.22), we see that

$$\operatorname{var}\operatorname{AUCV}(h) = \operatorname{var}\operatorname{UCV}(h) + \frac{4}{n^2}\operatorname{var}\sum_{i=1}^{n}f(x_i) + \frac{8}{n^3h}\operatorname{cov}\left(\sum\sum_{i<j}\gamma_{ij}, \sum_i f(x_i)\right). \quad (9.5)$$

Since $Ef(x_i)^k = \int f(x)^{k+1}\,dx$, we have that the second term equals $4I_3/n$. In (9.5), $\operatorname{cov}(\gamma_{ij}, f(x_k)) = 0$. Considering the $n(n-1)$ terms for which $k = i$ or $k = j$, $E[\gamma_{ij}f(x_i)] = -hR(f^{3/2}) + h^5\mu_2^2I_1/4 + O(h^7)$; hence $\operatorname{cov}(\gamma_{ij}, f(x_i)) = -hI_3 + \mu_2^2h^5[I_1 - I_2]/4$, proving Theorem 3.3.

9.2 Proof of Lemma 3.2

$$\hat{f}^{(p)}(x) = \frac{1}{nh^{p+1}}\sum_{i=1}^{n}K^{(p)}\left(\frac{x - x_i}{h}\right).$$

$$E\,R(\hat{f}^{(p)}) = \frac{R(K^{(p)})}{nh^{2p+1}} + \frac{n-1}{nh^{2p}}\int \times\left[\int K^{(p)}(w)f(x - hw)\,dw\right]^2 dx$$

after squaring and a change of variables. The bracketed term may be approximated by

$$\sum_{i=o}^{p+2}\frac{1}{i!}(-h)^if^{(i)}(x)\int w^iK^{(p)}(w)\,dw + o(h^{p+2}).$$

Now $\int w^iK^{(p)}(w)\,dw = 0$ if $i < p$ or if $i + p$ is odd, and it equals $(-1)^pp!$ for $i = p$ and $(-1)^p(p+2)!\mu_2/2$ for $i = p + 2$. Hence the sum collapses to $h^pf^{(p)}(x) + O(h^{p+2})$. Integrating the square completes the proof. We remark that since $\mu_k = 0$ for $0 < k < p$, the error is actually $O(h^p)$ if $f^{(2p)}$ exists.

9.3 Proof of Theorem 3.2

The analysis of the moments of the BCV function is similar to that in Section 9.1, although much easier, since $\operatorname{BCV}(h)$ involves fewer terms and because more "moments" of Equation (9.6) vanish. We assume that Conditions 1, 2a, and 2b are satisfied. We remark that Condition 2b is necessary for Theorem 4.2 but stronger than necessary by one order of derivative for Theorem 3.2. From (3.16) define

$$\phi_+(c) = \int_{-1}^{1-c}K''(w)K''(w + c)\,dw, \qquad 0 \le c \le 2, \quad (9.6)$$

and $\phi_-(c)$ for $-2 \le c \le 0$. Again ϕ is symmetric. Now

$$\int_0^2 c^k\phi_+(c)\,dc = \int_{-1}^{1}K''(w)\int_0^{1-w}c^kK''(w + c)\,dc\,dw. \quad (9.7)$$

For $k = 0$, observe that $\int_0^{1-w}K''(w + c)\,dc = -K'(w)$ and $-\int_{-1}^{1}K'(w)K''(w)\,dw = 0$. For $k \ge 2$, $\int_0^{1-w}c^kK''(w + c)\,dc = k(k-1)\int_0^{1-w}c^{k-2}K(w + c)\,dc$. Noting (for m even) that

$$\int_{-1}^{1}K(s)\int_{-1}^{s}w^mK(w)\,dw = \frac{1}{2}\int_{-1}^{1}K(s)\,ds\int_{-1}^{1}w^mK(w)\,dw = \frac{\mu_m}{2},$$

and integrating by parts, we see that (9.7) equals 0, 0, 12, and $360\mu_2$ for $k = 0, 2, 4$, and 6, respectively. The analysis proceeds exactly as in Section 9.1 with ϕ_+ replacing γ_+. Let $\phi_{ij} \equiv \phi(c_{ij})$. From (9.1), it follows that $E\phi_{ij} = h^5R(f'') - h^7\mu_2R(f''') + o(h^7)$, from which (3.18) follows directly. Following Section 9.1.2, $\operatorname{var}\phi_{ij} = hR(\phi)R(f) + O(h^3)$. In this case $\operatorname{cov}(\phi_{ij}, \phi_{ik}) = O(h^{10})$. Following Equation (9.3),

$$\operatorname{var}\left[\sum\sum_{i<j}\phi(c_{ij})\right] = n^2hR(\phi)R(f)/2 + O(n^2h^3 + n^3h^{10}),$$

which, together with (3.17), proves Equation (3.19).

The asymptotic normality follows from theorem 2.1 of Hall (1984) for AN of degenerate U statistics. Let $\mu(t) = EK''((t - X)/h)$, where $h = cn^{1/5}$. Decompose $h\phi(c_{ij})$ into

$$\int\left[K''\left(\frac{t - x_i}{h}\right) - \mu(t)\right]\left[K''\left(\frac{t - x_j}{h}\right) - \mu(t)\right]dt + \int\mu(t)\left[K''\left(\frac{t - x_i}{h}\right) + K''\left(\frac{t - x_j}{h}\right)\right]dt - \int\mu(t)^2\,dt.$$

Only the last two integrals contribute to the mean, which is easily checked to be $h^6R(f'') + o(h^6)$. The variance comes from the first integral, which we denote by $H_n(x_i, x_j)$. It may be verified that the random variable $E[H_n(X,$

$Y)|X] \equiv 0$, so H_n is a degenerate Martingale. Using the notation of Hall's (1984) equation (2.1), calculations similar to the foregoing give $EH_n^2 = h^3R(\phi)R(f)$, $EH_n^4 = h^5R(\phi^2)R(f)$, and $EG_n = O(h^7)$; therefore, the conditions for Hall's theorem 2.1 hold and BCV(h) is AN.

9.4 Proof of Corollaries 3.2 and 3.3

From (2.3) and Theorem 3.2 we have that

$$\operatorname{plim}_{n\to\infty}[\mathrm{BCV}(ch^*)/\mathrm{MISE}(ch^*)] = 1,$$

$$\operatorname{plim}_{n\to\infty}[\mathrm{AMISE}(ch^*)/\mathrm{MISE}(ch^*)] = 1,$$

$$\mathrm{AMISE}(ch^*)/\mathrm{AMISE}(h^*) = \frac{c^5 + 4}{5c}, \quad (9.8)$$

so $\mathrm{MISE}(ch^*) > \mathrm{MISE}(h^*)$ for $c \neq 1$ and large n. Suppose that $c = \tilde{h}_{\mathrm{BCV}}/h^*$ does not converge to 1. Then $\Pr\{\mathrm{BCV}(\tilde{h}_{\mathrm{BCV}}) < \mathrm{BCV}(h^*)\} \to 1$ as $n \to \infty$, which contradicts the consistency results in (9.8). The proof of Corollary 3.3 was first given by Hall (1983).

9.5 Proof of Lemma 4.1

As before, define $\rho_+(c)$ from (4.2) when $0 \le c \le 2$. Then it may be shown that $\int_0^2 c^k\rho_+(c)\,dc = \frac{1}{2}, 0, -15\mu_2^2$, and $-105\mu_2\mu_4$ for $k = 0, 2, 4$, and 6, respectively. Let $\rho_{ij} \equiv \rho(c_{ij})$. Omitting details, $E\rho_{ij} = hR(f) - 5\mu_2^2h^5R(f'')/4 + o(h^6)$, from which the expectation in (4.4) may be computed. Then

$$\operatorname{var}\rho_{ij} = hR(\rho)R(f) - h^2R(f)^2 + O(h^3),$$

$$\operatorname{cov}(\rho_{ij}, \rho_{ik}) = h^2I_3 + \frac{5}{2}\mu_2^2h^6[I_2 - I_1] + o(h^6),$$

$$\operatorname{cov}(\gamma_{ij}, \rho_{ik}) = -h^2I_3 + \frac{3}{2}\mu_2^2h^6[I_1 - I_2] + o(h^6),$$

and

$$\operatorname{cov}(\gamma_{ij}, \rho_{ij}) = hR((\gamma\rho)^{1/2})R(f) + h^2R(f)^2 + O(h^3).$$

Now the variance of the left side of (4.4) may be expressed as

$$\begin{aligned} &n(n-1)[\operatorname{var}\gamma_{ij} + \operatorname{var}\rho_{ij}]/2 \\ &\quad + n(n-1)(n-2)[\operatorname{cov}(\gamma_{ij}, \gamma_{ik}) + \operatorname{cov}(\rho_{ij}, \rho_{ik})] \\ &+ n(n-1)\operatorname{cov}(\gamma_{ij}, \rho_{ij}) + 2n(n-1)(n-2)\operatorname{cov}(\gamma_{ij}, \rho_{ik}). \end{aligned} \quad (9.9)$$

Evaluating (9.9), we find

$$\operatorname{var}\sum_{i<j}\sum[\gamma_{ij} + \rho_{ij}] = \tfrac{1}{2}n^2h[R(\gamma) + R(\rho) + 2R((\gamma\rho)^{1/2})]R(f),$$

where the bracketed term may be written as $R(\gamma + \rho)$. But $(d/dc)\,\gamma_+(c) = \rho_+(c)/c$. Hence

$$\int_0^2 \gamma_+(c)\rho_+(c)\,dc = \int_0^2 c\gamma_+(c)\gamma_+'(c)\,dc = -\tfrac{1}{2}\int_0^2 \gamma_+(c)^2\,dc,$$

since $\gamma_+(2) = 0$. Since γ is symmetric, $R(\gamma + \rho) = R(\rho)$, completing the argument.

9.6 Proof of Lemma 4.2

$\int_0^2 c^k\psi_+(c)\,dc = 0, 0, -60$, and $-2{,}520\mu_2$ for $k = 0, 2, 4$, and 6, respectively.

$$E\,\psi(c_{ij}) = -5h^5R(f'') + O(h^7),$$

$$\operatorname{var}\psi(c_{ij}) = hR(\psi)R(f) + O(h^3),$$

$$\operatorname{cov}(\psi_{ij}, \psi_{ik}) = O(h^{10}); \operatorname{cov}(\phi_{ij}, \psi_{ik}) = O(h^3),$$

$$\operatorname{cov}(\phi_{ij}, \psi_{ij}) = hR((\phi\psi)^{1/2})R(f) + O(h^3),$$

and $R(\phi + \psi) = R(\psi)$ as before. The lemma follows directly.

[*Received March 1986. Revised December 1986.*]

REFERENCES

Bowman, A. W. (1984), "An Alternative Method of Cross-Validation for the Smoothing of Density Estimates," *Biometrika,* 71, 353–360.

——— (1985), "A Comparative Study of Some Kernel-Based Nonparametric Density Estimates," *Journal of Statistical Computation and Simulation,* 21, 313–327.

Carroll, R. J., and Ruppert, D. (1985), "Transformations in Regression: A Robust Analysis," *Technometrics,* 27, 1–12.

Fryer M. J. (1976), "Some Errors Associated With the Non-parametric Estimation of Density Functions," *Journal of the Institute of Mathematics and Applications,* 18, 371–380.

Gale, W. A., and Pregibon, D. (1983), "An Expert System for Regression Analysis," in *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface,* eds. K. W. Heiner, R. S. Sacher, and J. W. Wilkinson, New York: Springer-Verlag, pp. 110–117.

Good, I. J., and Gaskins, R. A. (1980), "Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data," *Journal of the American Statistical Association,* 75, 42–56.

Hall, P. (1983), "Large Sample Optimality of Least Squares Cross-Validation in Density Estimation," *The Annals of Statistics,* 11, 1156–1174.

——— (1984), "Central Limit Theorem for Integrated Squared Error of Multivariate Nonparametric Density Estimators," *Journal of Multivariate Analysis,* 14, 1–16.

Hall, P., and Marron, J. S. (1985), "Extent to Which Least-Squares Cross-Validation Minimises Integrated Square Error in Nonparametric Density Estimation," Technical Report 94, University of North Carolina, Dept. of Statistics.

Härdle, W., and Marron, J. S. (1985), "Optimal Bandwidth Selection in Nonparametric Regression Function Estimation," *The Annals of Statistics,* 13, 1465–1481.

Kronmal, R., and Tarter, M. E. (1968), "The Estimation of Probability Densities and Cumulatives by Fourier Series Methods," *Journal of the American Statistical Association,* 63, 925–952.

Rice J. (1984), "Bandwidth Choice for Nonparametric Regression," *The Annals of Statistics,* 12, 1215–1230.

Rudemo, M. (1982), "Empirical Choice of Histogram and Kernel Density Estimators," *Scandinavian Journal of Statistics,* 9, 65–78.

Scott, D. W. (1979), "On Optimal and Data-Based Histograms," *Biometrika,* 66, 605–610.

——— (1985), "Averaged Shifted Histograms: Effective Nonparametric Estimators in Several Dimensions," *The Annals of Statistics,* 13, 1024–1040.

——— (1986), "Choosing Smoothing Parameters for Density Estimators," in *Proceedings of the 17th Symposium on the Interface of Computer Science and Statistics,* Amsterdam: North-Holland, pp. 225–230.

——— (in press), Comment on "How Far Are Automatically Chosen Regression Smoothing Parameters From Their Optimum?," by W. Härdle, P. Hall, and J. S. Marron, *Journal of the American Statistical Association,* 83.

Scott, D. W., and Factor, L. E. (1981), "Monte Carlo Study of Three Data-Based Nonparametric Probability Density Estimators," *Journal of the American Statistical Association,* 76, 9–15.
Scott, D. W., Tapia, R. A., and Thompson, J. R. (1977), "Kernel Density Estimation Revisited," *Journal of Nonlinear Analysis, Theory, Methods and Applications,* 1, 339–372.
Scott, D. W., and Terrell, G. R. (1986), "Biased and Unbiased Cross-Validation in Density Estimation," Technical Report 246, Stanford University, Dept. of Statistics.
Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics,* New York: John Wiley.
Silverman, B. W. (1978), "Choosing the Window Width When Estimating a Density," *Biometrika,* 65, 1–11.
——— (1982), "Kernel Density Estimation Using the Fast Fourier Transform," Statistical Algorithm AS 176, *Applied Statistics,* 31, 93–97.
Stone, C. J. (1984), "An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates," *The Annals of Statistics,* 12, 1285–1297.
Stone, M. (1974), "Cross-Validation and Multinomial Prediction," *Biometrika,* 61, 509–515.
Titterington, D. M. (1985), "Common Structure of Smoothing Techniques in Statistics," *International Statistical Review,* 53, 141–170.
Wahba, G. (1981), "Data-Based Optimal Smoothing of Orthogonal Series Density Estimates," *The Annals of Statistics,* 9, 146–156.
Woodroofe, M. (1970), "On Choosing a Delta Sequence," *Annals of Mathematical Statistics,* 41, 1665–1671.