ON THE ASYMPTOTIC EQUIVALENCE OF HELLINGER DISTANCE AND KULLBACK-LEIBLER LOSS

Yuichiro Kanazawa*, Atsuyuki Kogure** and Sang-Gil Lee***

Hellinger distance, Kullback-Leibler loss, and variance-stabilizing integrated square error are shown to be asymptotically equivalent up to a constant. Simulation confirmed the result. Implications of the result for kernel and histogram density estimation and the connection of the result with goodness of fit tests are discussed.

Key Words and Phrase: Akaike's Information criterion, Hellinger distance, histogram, integrated absolute error, integrated square error, kernel estimator, Kullback-Leibler loss, least-squares cross-validation, likelihood cross-validation.

1. Introduction

Suppose we wish to construct the kernel density estimator or histogram $\hat{f}$ with the width of its smoothing parameter being h based on a random sample $X_1, \dots, X_n$ of size n from a density f . The kernel estimator $\hat{f}$ is defined by

$$(1.1) \quad \hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right),$$

where $K(t)$ is a kernel function and h is a window width. The histogram is determined by the counts of observations in the cells $[a + (j-1)h, a + jh)$, $j=1, \dots, k$, where a is an origin, h is a cell width, and k is the number of cells. They are closed on the left and open on the right for definiteness.

Since the parameter h primarily controls the amount of smoothing and global discrepancy between the estimator and the density, it is of interest to obtain the theoretically optimal width of h . To do so, one first selects a global measure of discrepancy between the estimator and the density. Then one minimizes, usually asymptotically, the expected value of the measure.

A few measures of discrepancy were utilized for density estimation. As a measure based on distance, Parzen (1962) used mean integrated squared error (MISE) $E\{\int (\hat{f} - f)^2\}$ to obtain the asymptotically optimal window width

$$(1.2) \quad h_{MISE}^{window} = \left[\frac{\int K(t)^2 dt}{\left\{ \int t^2 K(t) dt \right\}^2} \right]^{1/5} \left\{ \int f^{(2)}(x)^2 dx \right\}^{-1/5} n^{-1/5}$$

Received November, 1995. Revised June, 1998. Accepted July, 1998.

* Institute of Policy and Planning Sciences, The University of Tsukuba, 1-1-1 Tennohdai, Tsukuba, Ibaraki 305-8573, Japan.

** Department of Economics, Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba 263-0022, Japan.

*** Doctoral Program in Policy and Planning Sciences, University of Tsukuba, 1-1-1 Tennohdai, Tsukuba, Ibaraki 305-8573, Japan.

for the kernel estimator. Scott (1979) and Freedman and Diaconis (1981) used the MISE to obtain the asymptotically optimal cell width

$$(1.3) \quad h_{MISE}^{\text{cell}} = \left[\frac{6}{\int f^{(1)}(x)^2 dx} \right]^{1/3} n^{-1/3}$$

for the histogram. For mean integrated squared error applied to a histogram with varying cell widths, see Kogure (1987).

Devroye and Györfi (1985) employed the asymptotic upper bound of mean integrated absolute error (MIAE) $E\{ \int |\hat{f} - f| \}$ to obtain the close-to-optimal width

$$h_{MIAE}^{\text{window}} = (2\pi)^{-1/5} \left[\frac{\int K(t)^2 dt}{\left\{ \int t^2 K(t) dt \right\}^2} \right]^{1/5} \left[\frac{\left\{ \int |f^{(2)}(x)| dx \right\}^2}{\left\{ \int f(x)^{1/2} dx \right\}} \right]^{-1/5} n^{-1/5}$$

for the kernel estimator. Similarly they obtained the close-to-optimal width

$$h_{MIAE}^{\text{cell}} = \left[\frac{8 \left\{ \int f(x)^{1/2} dx \right\}^2}{\pi \left\{ \int |f^{(1)}(x)| dx \right\}^2} \right]^{1/3} n^{-1/3}$$

for the histogram.

As a measure founded on information-theoretic concepts, the expected value of Kullback-Leibler loss (EKLL) $E \int f \log(f/\hat{f})$ was carefully examined by Hall (1987) to obtain the asymptotically optimal width

$$(1.4) \quad h_{EKLL}^{\text{window}} = (4 \cdot |I|)^{1/5} \left[\frac{\int K(t)^2 dt}{\left\{ \int t^2 K(t) dt \right\}^2} \right]^{1/5} \left\{ \int_I \frac{f^{(2)}(x)^2}{f(x)} dx \right\}^{-1/5} n^{-1/5}$$

for the kernel estimator for sufficiently smooth and compactly supported densities on $|I|$. Let v_i be the number of data in the i -th cell, h the cell width, and J the total number of histogram cells. Taylor (1987) used the expected value of Akaike's information criterion (EAIC) $J - \log[\prod_{i=1}^{\# \text{ of cells}} \{v_i/(nh)\}^{v_i}]$ to obtain the asymptotically optimal cell width

$$(1.5) \quad h_{EAIC}^{\text{cell}} = \left[\frac{12|I|}{\int \{f^{(1)}(x)^2/f(x)\} dx} \right]^{1/3} n^{-1/3}$$

for the histogram where $|I|$ represents the width of the support of the data. Hall (1990) employed the Kullback-Leibler loss (KLL) to obtain the cell width h_{KLL}^{cell} that asymptotically minimizes the KLL and that is *equivalent* to the h_{EAIC}^{cell} for sufficiently smooth and compactly supported densities. Here he had to work with raw loss KLL instead of the expected loss EKLL because there is a

nonzero probability that the histogram vanishes in any given cell. The equivalence between the $h_{E_{AIC}}^{\text{cell}}$ and h_{KLL}^{cell} should not surprise us because AIC is essentially a model selection rule based on KLL as Akaike (1973) wrote: "...this [AIC] is equivalent to maximizing an information theoretic quantity which is given by the definition

$$E \log \frac{f(x|\hat{\theta})}{f(x|\theta)} = E \int f(x|\theta) \log \frac{f(x|\hat{\theta})}{f(x|\theta)} dx.$$

The integral in the right-hand side of the above equation gives the Kullback-Leibler's mean information for discrimination between $f(x|\hat{\theta})$ and $f(x|\theta)$"

As for the relationship between distance-based and information-theoretic measures of discrepancy, we know the following. For the kernel estimator, the window width h_{MHD}^{window} that minimizes mean Hellinger distance (MHD)

$$MHD(\hat{f}, f) = E \int \{\hat{f}(x)^{1/2} - f(x)^{1/2}\}^2 dx$$

was found asymptotically equivalent in Kanazawa (1993b) to the $h_{E_{KLL}}^{\text{window}}$ in (1.4). For the histogram, the cell width h_{MHD}^{cell} that minimizes the MHD was recognized to be asymptotically equivalent in Kanazawa (1993a) to the $h_{E_{AIC}}^{\text{cell}}$ in (1.5) as well. The results strongly suggest that information-theoretic Kullback-Leibler loss itself is equivalent to distance-based Hellinger distance (HD) in some sense. We will investigate the asymptotic relationship in section 2. We will present the simulation results that quantify the relationship in section 3. We will discuss its implications for kernel and histogram density estimation and point out the connection of the results to those of goodness-of-fit tests in section 4. The proof is given in section 5. For Hellinger distance applied to a histogram with varying cell widths, see Kanazawa (1988, 1992).

2. Asymptotic result

Let $\hat{f}$ be the kernel estimator or histogram. For the former, we additionally assume the following on the kernel:

- K1. $K(t)$ is a probability density on $\mathfrak{N}$.
- K2. $K(t)$ is right continuous on $\mathfrak{N}$.
- K3. $K(t)$ is of bounded variation on $\mathfrak{N}$.
- K4. $\int_{\mathfrak{N}} |t| K(t) dt < \infty$.

REMARK 1 The weak set of assumptions, corresponding to ASSUMPTION 2 in Prakasa Rao (1983, pp. 50), establishes $\sup_{x \in I} |\hat{f}(x) - f(x)| = o_p(1)$ in lemma 1 on page 35, where I is the support of the density f . The lemma is necessary to prove theorem 1. The assumptions are satisfied by such standard kernels as Epanechnikov, biweight, triangular, rectangular, and normal.

Define variance-stabilizing integrated square error (VSISE) as

$$(2.1) \quad VSISE(\hat{f}, f) = \int \frac{\{\hat{f}(x) - f(x)\}^2}{f(x)} dx.$$

We will explain why we characterize (2.1) as variance-stabilizing in section 4.

THEOREM 1 *Let the following conditions on the density f be satisfied:*

D1. f is compactly supported on I .

D2. f is bounded away from zero on I .

D3. $\sup_{x \neq y} |f(x) - f(y)|/|x - y| < \infty$.

Let h be the window width of the kernel estimator or the cell width of the histogram. Suppose $h \rightarrow 0$ and $(nh)^{-1} \log n \rightarrow 0$ as $n \rightarrow \infty$. As $n \rightarrow \infty$,

$$VSISE(\hat{f}, f) \stackrel{p}{=} 4 \cdot HD(\hat{f}, f) \stackrel{p}{=} 2 \cdot KLL(\hat{f}, f)$$

where $A \stackrel{p}{=} B$ means $A/B \rightarrow 1$ in probability.

REMARK 2 D3 is required to establish $\sup_{x \in I} |\hat{f}(x) - f(x)| = o_p(1)$ for the kernel estimator $\hat{f}(x)$ in lemma 1. It corresponds to ASSUMPTION 4 in Prakasa Rao (1983, pp. 50). D1 and D3 are required to establish the same maximal boundary for the histogram $\hat{f}(x)$.

3. Simulation Result

We run the simulation using the standard normal symmetrically truncated about the origin supported on $I = [-1.959964, 1.959964]$, $[-2.575829, 2.575829]$, and $[-3.290527, 3.290527]$ as our underlying densities $f(x)$. They satisfy conditions D1 through D3. The areas on these intervals correspond to 0.95, 0.99, and 0.999 under the untruncated standard normal respectively. We use the “minimal” random number generator of Park and Miller with the Bays-Durham shuffle in Press et al. (pp. 280) to generate a uniform random deviate between 0.0 and 1.0 and convert it to a normally distributed deviate with zero mean and unit variance by the Box-Muller transformation in Press et al. (pp. 289). The values outside the intervals are discarded to produce the standard normal deviates symmetrically truncated about the origin.

For $\hat{f}(x)$, we use both the kernel estimator and histogram. To construct the former, we used three types of widely used kernels known to satisfy conditions K1 through K4. First the Epanechnikov kernel

$$K_e(t) = \begin{cases} \frac{3}{4} \left(1 - \frac{1}{5} t^2\right) / \sqrt{5} & \text{for } |t| < \sqrt{5}, \\ 0 & \text{otherwise,} \end{cases}$$

which is known to minimize the asymptotic MISE. Second the Gaussian kernel

$$K_g(t) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right),$$

which conforms to the shape of underlying normal densities $f(x)$ except that the $f(x)$ are symmetrically truncated about the origin. Third the rectangular kernel

$$K_r(t) = \begin{cases} \frac{1}{2} & \text{for } |t| < 1, \\ 0 & \text{otherwise,} \end{cases}$$

which is known to be the crudest kernel because of its shape. We chose these kernels for their relative strengths and weaknesses. Since the efficiencies of the two symmetric kernels K_g and K_r relative to K_e are 0.9512 and 0.9295 respectively and our arguments are fundamentally asymptotic, we expect the choice of kernel to be immaterial in the sense that one of them will perform only marginally better or worse in a given situation. See Silverman (1986, pp. 42-43) for the efficiency of symmetric kernels.

With the kernel, the window width h_{EKLL}^{window} in (1.4) that asymptotically optimizes the EKLL, MHD, and MVSISE can easily be computed. For the Epanechnikov kernel the h_{EKLL}^{window} is

$$\begin{aligned} (3.1) \quad h_{EKLL}^{\text{window}} &= (4 \cdot |I|)^{1/5} \left[\frac{\int K_e(t)^2 dt}{\left\{ \int t^2 K_e(t) dt \right\}^2} \right]^{1/5} \left\{ \int_I \frac{f^{(2)}(x)^2}{f(x)} dx \right\}^{-1/5} n^{-1/5} \\ &= \left(\frac{12\sqrt{5}}{25} \cdot |I| \right)^{1/5} \left\{ \int_I \frac{f^{(2)}(x)^2}{f(x)} dx \right\}^{-1/5} n^{-1/5} \\ &= \begin{cases} 1.3827 \cdot n^{-1/5} & \text{on } [-1.960, 1.960], \\ 1.3114 \cdot n^{-1/5} & \text{on } [-2.576, 2.576], \\ 1.3057 \cdot n^{-1/5} & \text{on } [-3.291, 3.291]. \end{cases} \end{aligned}$$

We use $I = [-1.960, 1.960]$, $[-2.576, 2.576]$, and $[-3.291, 3.291]$ to abbreviate $I = [-1.959964, 1.959964]$, $[-2.575829, 2.575829]$, and $[-3.290527, 3.290527]$ respectively for the rest of this article. Similarly we calculate h_{EKLL}^{window} for the two other kernels.

Using these window widths, we first compute the kernel estimator $\hat{f}(x)$ for a sample of size n . If the support of $\hat{f}(x)$ exceeds the support I of $f(x)$, then the estimator $\hat{f}(x)$ is also truncated to be on I so that its support matches that of $f(x)$.

Then the Hellinger distance

$$HD(\hat{f}, f) = \int_I \{ \hat{f}(x)^{1/2} - f(x)^{1/2} \}^2 dx,$$

Kullback-Leibler loss

$$KLL(\hat{f}, f) = \int_I f(x) \log \frac{f(x)}{\hat{f}(x)} dx,$$

and variance-stabilizing integrated square error

$$VSISE(\hat{f}, f) = \int_I \frac{\{ \hat{f}(x) - f(x) \}^2}{f(x)} dx$$

are computed under the kernel estimator $\hat{f}(x)$. To do so, we apply the

five-point closed Newton-Cotes formula, also known as *Bode's rule*:

$$\int_{x_1}^{x_5} g(x) dx = h \left[\frac{14}{45} g_1 + \frac{64}{45} g_2 + \frac{24}{45} g_3 + \frac{64}{45} g_4 + \frac{14}{45} g_5 \right] + O(h^7 f^{(6)}),$$

where h is the fixed-width for an interval used for integration, $x_i = x_0 + ih$, $i=1, \dots, 5$ is the i -th abscissa, and $g_i = g(x_i)$, repeatedly for the support I from $x_1 = \inf I$ to $x_N = \sup I$ using $g(x) = \{\hat{f}(x)^{1/2} - f(x)^{1/2}\}^2$, $g(x) = f(x) \log\{f(x)/\hat{f}(x)\}$, and $g(x) = \{\hat{f}(x) - f(x)\}^2 / f(x)$ respectively. The resulting extended Newton-Cotes formula is thus

$$\begin{aligned} \int_{x_1}^{x_N} f(x) dx = & h \left[\frac{14}{45} g_1 + \frac{64}{45} g_2 + \frac{24}{45} g_3 + \frac{64}{45} g_4 \right. \\ & + \frac{28}{45} g_5 + \frac{64}{45} g_6 + \frac{24}{45} g_7 + \frac{64}{45} g_8 \\ & \dots \\ & \left. + \frac{28}{45} g_{N-4} + \frac{64}{45} g_{N-3} + \frac{24}{45} g_{N-2} + \frac{64}{45} g_{N-1} + \frac{14}{45} g_N \right] \\ & + O\left(\frac{|I|^7}{N^6} f^{(6)}\right). \end{aligned}$$

Finally we compute the ratios $4 \cdot HD(\hat{f}, f)$ and $2 \cdot KLL(\hat{f}, f)$ with respect to $VSISE(\hat{f}, f)$ respectively, and also $KLL(\hat{f}, f)$ relative to $2 \cdot HD(\hat{f}, f)$.

These procedures are repeated for the three types of kernel estimator one hundred times to obtain their averages and standard deviations for sample sizes of 100, 500, and 1,000.

We run the simulation for three types of histogram, two with a universal cell width whose numbers of cells are determined in slightly different ways and the other whose cell width is calculated according to equation (3.2) below except for the two outermost cells.

To construct any of these histograms, we first compute the cell width h_{EAIC}^{cell} in (1.5) that asymptotically optimizes the EKLL, MHD, and MVSISE as

$$\begin{aligned} (3.2) \quad h_{EAIC}^{\text{cell}} &= \left[\frac{12|I|}{\int \{f^{(1)}(x)^2 / f(x)\} dx} \right]^{1/3} n^{-1/3} \\ &= \begin{cases} 3.95763 \cdot n^{-1/3} & \text{on } [-1.960, 1.960], \\ 4.05856 \cdot n^{-1/3} & \text{on } [-2.576, 2.576], \\ 4.30740 \cdot n^{-1/3} & \text{on } [-3.291, 3.291]. \end{cases} \end{aligned}$$

Using these cell widths, we then compute the theoretical number of cells in these intervals as in table 1.

Of the two types of histograms with a universal cell width, in the first the number of cells is determined by rounding, and in the second the number of cells is determined by raising fractions to the nearest whole number. For example when $n=500$ for $[-2.576, 2.576]$ the first has ten cells of width $2.576 \times 2/10 \approx 0.5152$ while the second has eleven cells of width $2.576 \times 2/11 \approx 0.4684$. These two types of histograms differ in four cases in table 1, namely when $n=500$ for

Table 1. Theoretical number of cells for the interval. Pairs of numbers in parentheses are those after rounding and raising fractions to the nearest whole number.

n	$[-1.960, 1.960]$	$[-2.576, 2.576]$	$[-3.291, 3.291]$
100	4.59(5,5)	5.89(6,6)	7.09(7,8)
500	7.86(8,8)	10.07(10,11)	12.12(12,13)
1,000	9.90(10,10)	12.69(13,13)	15.28(15,15)

$[-2.576, 2.576]$, and $n=100, 500, 1,000$ for $[-3.291, 3.291]$.

The placement of center cells is determined in the following manner in order to avoid artificial asymmetry in $\hat{f}(x)$. When the number of cells is odd as in $2k-1$, where k is a natural number, we place the k -th cell centered about the origin. For example when $n=100$ for $[-1.960, 1.960]$, the number of cells is five after rounding or raising fractions. The third cell, which is the center cell, is placed symmetrically about zero. When it is even as in $2k$, we place the k -th and $k+1$ -th cells symmetrically about the origin. For instance when $n=500$ for $[-1.960, 1.960]$, the number of cells is eight after rounding or raising fractions. The fourth and fifth cells, which are the two center cells, are placed symmetrically about the origin.

The third histogram, whose center cells are placed similarly, differs from the first two only in whether we allow different cell widths. For instance when $n=100$ for $[-3.291, 3.291]$, the fourth and fifth cells of width $4.30740 \times 100^{-1/3} \approx 0.928$ are placed at $[-0.9280, 0]$ and $[0, 0.9280]$. When $n=500$ for $[-3.291, 3.291]$, the seventh cell of width $4.30740 \times 500^{-1/3} \approx 0.543$ is placed symmetrically at $[-0.2715, 0.2715]$. We then continue to place cells outward towards the boundaries and the two outermost cells exceeding boundaries are narrowed. For the former the two outermost cells of width 0.507 are placed at $[-3.291, -2.784]$ and $[2.784, 3.291]$, while for the latter the two outermost cells of width 0.3045 are placed at $[-3.291, -2.9865]$ and $[2.9865, 3.291]$. The height of those cells are recalculated so that the support I of $\hat{f}(x)$ matches that of $f(x)$ while $\hat{f}(x)$ is still integrated to unity on I . We use the theoretical number of cells whose fraction is raised to the nearest whole number as the number of cells for the third type of histogram to avoid artificially deflating the number of cells.

In summary the first and second histograms share the feature of one universal cell width, while the second and third histograms have a common number of cells, the cell width and the number of cells being the only two parameters that determine smoothness of the histogram.

Finally we compute $HD(\hat{f}, f)$, $KLL(\hat{f}, f)$, and $VSISE(\hat{f}, f)$ and their ratios as we do for the kernel estimator. These procedures are repeated for the three types of histogram one hundred times to obtain their averages and standard deviations for sample sizes of 100, 500, and 1,000.

Narrower h evaluates $f(x)$ and $\hat{f}(x)$ in finer mesh, but increases the amount of calculation in integration and thus can accumulate roundoff errors, while wider h makes estimates of $f(x)$ and $\hat{f}(x)$ less precise, but accumulation of roundoff errors less of a problem. To determine the desirable fixed-width h

for the interval used for integration according to Bode's rule in our simulation and to evaluate the extent of accuracy under h , we proceed as follows. For any $x \in I$ define

$$g(x) = \frac{\hat{f}(x) - f(x)}{f(x)}.$$

For any set on which $\log\{1 + g(x)\}$ is well defined, expanding the term with the remainder in integral form gives

$$\log\{1 + g(x)\} = g(x) - \frac{g(x)^2}{2} + \int_0^{g(x)} \{g(x) - t\}^2 (1+t)^{-3} dt.$$

Hence we have

$$(3.3) \quad 2 \int_A f(x) \log \frac{f(x)}{\hat{f}(x)} dx = \int_A \frac{\{\hat{f}(x) - f(x)\}^2}{f(x)} dx \\ + 2 \int_A \int_0^{g(x)} \{g(x) - t\}^2 (1+t)^{-3} dt \cdot f(x) dx.$$

We numerically evaluate both sides in (3.3) for the kernel estimators and histograms $\hat{f}(x)$ against the normal density $f(x)$ truncated at $[-1.960, 1.960]$ with different h s when $n=1,000$ or when the largest number of normal random deviates are crammed in the narrowest support in our simulation. We find that the kernel estimator $h=|I|/200$, or $N=200$, gives accuracy to the seventh decimal place, but narrower h does not improve the accuracy. For the histogram $h=|I|/10,000$, or $N=10,000$ the same accuracy is obtained, but further narrowing of h does not improve the accuracy.

According to Bode's rule, the true answer differs from the estimate by an amount that is the product of some numerical coefficient times N^{-6} times the value of the function's sixth derivative somewhere in the interval of integration. The desirable fixed-widths h for the interval used for integration being $h=|I|/200$ for the kernel estimator and $h=|I|/10,000$ for the histogram, we expect errors due to numerical integration to be orders of 10^{-14} and 10^{-24} respectively. These error figures of numerical integration along with double-precision accuracy of 2.22×10^{-16} for a typical IEEE-compliant personal computer is smaller than the accuracy of the seventh decimal place we obtain when we evaluate equation (3.3). We are reasonably confident that the numbers in the simulation are accurate up to the seventh decimal place and we use these h s throughout in our simulation.

The simulation programs are written in C++ and compiled with a Watcom C/C++ compiler Version 10.5 J. and run on a personal computer with a 266 Mhz Pentium Pro processor and 64 MB of memory. To evaluate the computer's floating-point arithmetic, we used the implementation of Cody (1988) embodied in the routine in Press et al. (pp. 892-894). We found the computer we use is IEEE-compliant.

The results are respectively presented in tables 2, 3, and 4 for kernel estimators with the Epanechnikov, Gaussian, and rectangular kernels and in

Table 2. The ratios $4HD(\hat{f}, f)/V\text{SISE}(\hat{f}, f)$, $2KLL(\hat{f}, f)/V\text{SISE}(\hat{f}, f)$, and $KLL(\hat{f}, f)/2HD(\hat{f}, f)$ after one hundred repetitions. The truncated normal on the respective intervals is used as f against the kernel estimator $\hat{f}$ using the Epanechnikov kernel.

ratio	n	[-1.960, 1.960]		[-2.576, 2.576]		[-3.291, 3.291]	
		mean	s.d.	mean	s.d.	mean	s.d.
$\frac{4HD(\hat{f}, f)}{V\text{SISE}(\hat{f}, f)}$	100	0.9470129	0.0469324	0.8774978	0.0843851	0.7373555	0.1164603
	500	0.9808879	0.0270548	0.9333151	0.0453505	0.8467559	0.0821754
	1,000	0.9954517	0.0254459	0.9472115	0.0363680	0.8739815	0.0740261
$\frac{2KLL(\hat{f}, f)}{V\text{SISE}(\hat{f}, f)}$	100	0.9337581	0.0603017	0.8512374	0.1091057	0.6889521	0.1306936
	500	0.9757532	0.0356945	0.9156733	0.0587316	0.8122579	0.0998430
	1,000	0.9948991	0.0340169	0.9323769	0.0466892	0.8434304	0.0908399
$\frac{KLL(\hat{f}, f)}{2HD(\hat{f}, f)}$	100	0.9852920	0.0146190	0.9674437	0.0301328	0.9291425	0.0338966
	500	0.9945232	0.0088857	0.9803869	0.0148906	0.9568569	0.0252890
	1,000	0.9992293	0.0085259	0.9838968	0.0116532	0.9631473	0.0228015

Table 3. The ratios $4HD(\hat{f}, f)/V\text{SISE}(\hat{f}, f)$, $2KLL(\hat{f}, f)/V\text{SISE}(\hat{f}, f)$, and $KLL(\hat{f}, f)/2HD(\hat{f}, f)$ after one hundred repetitions. The truncated normal on the respective intervals is used as f against the kernel estimator $\hat{f}$ using the Gaussian kernel.

ratio	n	[-1.960, 1.960]		[-2.576, 2.576]		[-3.291, 3.291]	
		mean	s.d.	mean	s.d.	mean	s.d.
$\frac{4HD(\hat{f}, f)}{V\text{SISE}(\hat{f}, f)}$	100	0.9449631	0.0510941	0.8643708	0.0769286	0.7566469	0.1367120
	500	0.9829234	0.0330932	0.9252219	0.0530362	0.8475299	0.0806539
	1,000	0.9947981	0.0273299	0.9419610	0.0360609	0.8719358	0.0720679
$\frac{2KLL(\hat{f}, f)}{V\text{SISE}(\hat{f}, f)}$	100	0.9309282	0.0652714	0.8332055	0.0948599	0.7136820	0.1622730
	500	0.9786747	0.0435803	0.9054614	0.0676243	0.8129877	0.0982806
	1,000	0.9940798	0.0364954	0.9256794	0.0464514	0.8405028	0.0878325
$\frac{KLL(\hat{f}, f)}{2HD(\hat{f}, f)}$	100	0.9842977	0.0159458	0.9618327	0.0242886	0.9356752	0.0431632
	500	0.9953172	0.0108334	0.9776891	0.0168412	0.9569059	0.0252646
	1,000	0.9990288	0.0091683	0.9822755	0.0116340	0.9621660	0.0219147

Table 4. The ratios $4HD(\hat{f}, f)/V\text{SISE}(\hat{f}, f)$, $2KLL(\hat{f}, f)/V\text{SISE}(\hat{f}, f)$, and $KLL(\hat{f}, f)/2HD(\hat{f}, f)$ after one hundred repetitions. The truncated normal on the respective intervals is used as f against the kernel estimator $\hat{f}$ using the rectangular kernel.

ratio	n	[-1.960, 1.960]		[-2.576, 2.576]		[-3.291, 3.291]	
		mean	s.d.	mean	s.d.	mean	s.d.
$\frac{4HD(\hat{f}, f)}{V\text{SISE}(\hat{f}, f)}$	100	0.9335204	0.0397401	0.8519851	0.0771487	0.6778051	0.0983408
	500	0.9846510	0.0303235	0.9224342	0.0517708	0.8233999	0.0810923
	1,000	0.9957844	0.0265766	0.9414410	0.0357663	0.8677843	0.0769613
$\frac{2KLL(\hat{f}, f)}{V\text{SISE}(\hat{f}, f)}$	100	0.9166502	0.0503402	0.8198560	0.0949215	0.6230735	0.1094559
	500	0.9810446	0.0401052	0.9023312	0.0666319	0.7849333	0.0971522
	1,000	0.9954344	0.0355687	0.9252696	0.0462715	0.8370060	0.0957381
$\frac{KLL(\hat{f}, f)}{2HD(\hat{f}, f)}$	100	0.9814216	0.0121421	0.9601416	0.0243861	0.9151817	0.0289775
	500	0.9960365	0.0098963	0.9772790	0.0167488	0.9509268	0.0245681
	1,000	0.9994128	0.0089347	0.9823863	0.0116416	0.9624117	0.0244435

Table 5. The ratios $4HD(\hat{f}, f)/VSISE(\hat{f}, f)$, $2KLL(\hat{f}, f)/VSISE(\hat{f}, f)$, and $KLL(\hat{f}, f)/2HD(\hat{f}, f)$ after one hundred repetitions. The truncated normal on the respective intervals is used as f against the histogram $\hat{f}$ with a universal cell width. Width of support $|I|$ is first divided by the theoretical cell width and the resulting theoretical number of cells is rounded to the nearest whole number to obtain the number of cells for $\hat{f}$.

ratio	n	on $[-1.960, 1.960]$		on $[-2.576, 2.576]$		on $[-3.291, 3.291]$	
		mean	s.d.	mean	s.d.	mean	s.d.
$\frac{4HD(\hat{f}, f)}{VSISE(\hat{f}, f)}$	100	0.9630913	0.1350741	0.8842269	0.1374695	0.7193663	0.1609367
	500	0.9694754	0.0614999	0.9608975	0.1286021	0.9039504	0.1181262
	1,000	0.9836915	0.0429809	0.9749960	0.0616040	0.9346517	0.0963016
$\frac{2KLL(\hat{f}, f)}{VSISE(\hat{f}, f)}$	100	0.9831560	0.1909135	0.8958493	0.1871205	0.7129331	0.1998913
	500	0.9700724	0.0818528	0.9745685	0.0905062	0.9060932	0.1533774
	1,000	0.9849329	0.0569407	0.9770101	0.0838168	0.9432863	0.1262867
$\frac{KLL(\hat{f}, f)}{2HD(\hat{f}, f)}$	100	1.0137820	0.0519489	1.0052950	0.0544773	0.9810010	0.0588819
	500	0.9993110	0.0210955	1.0109640	0.0367173	0.9975392	0.0410411
	1,000	1.0006510	0.0142378	1.0006980	0.0226168	1.0061310	0.0326919

Table 6. The ratios $4HD(\hat{f}, f)/VSISE(\hat{f}, f)$, $2KLL(\hat{f}, f)/VSISE(\hat{f}, f)$, and $KLL(\hat{f}, f)/2HD(\hat{f}, f)$ after one hundred repetitions. The truncated normal on the respective intervals is used as f against the histogram $\hat{f}$ with a universal cell width. Width of support $|I|$ is first divided by the theoretical cell width and the resulting theoretical number of cells is raised to the nearest whole number to obtain the number of cells for $\hat{f}$.

ratio	n	on $[-1.960, 1.960]$		on $[-2.576, 2.576]$		on $[-3.291, 3.291]$	
		mean	s.d.	mean	s.d.	mean	s.d.
$\frac{4HD(\hat{f}, f)}{VSISE(\hat{f}, f)}$	100	0.9630913	0.1350741	0.8842269	0.1374695	0.7168511	0.1584813
	500	0.9694754	0.0614999	0.9527972	0.0937409	0.8928009	0.1152200
	1,000	0.9836915	0.0429809	0.9749960	0.0616040	0.9266574	0.0942673
$\frac{2KLL(\hat{f}, f)}{VSISE(\hat{f}, f)}$	100	0.9831560	0.1909135	0.8958493	0.1871205	0.7093767	0.2050385
	500	0.9700724	0.0818528	0.9599009	0.1302375	0.8931010	0.1512372
	1,000	0.9849329	0.0569407	0.9770101	0.0838168	0.9291902	0.1231452
$\frac{KLL(\hat{f}, f)}{2HD(\hat{f}, f)}$	100	1.0137820	0.0519489	1.0052950	0.0544773	0.9786488	0.0609374
	500	0.9993110	0.0210955	1.0040010	0.0379945	0.9955194	0.0408154
	1,000	1.0006510	0.0142378	1.0006980	0.0226168	0.9996720	0.0337292

tables 5, 6, and 7 for three types of histogram $\hat{f}(x)$. Results for the Epanechnikov kernel estimators and for the histogram whose number of cells are determined by raising the theoretical number of cells to the next whole number are graphically summarized in figures following the corresponding tables. The figures 1 through 6 plot means and standard deviations of the three ratios against the corresponding sample sizes.

We first summarize the most important results. They are equally observed whether $\hat{f}(x)$ is the kernel estimator or the histogram.

- The ratios $4HD(\hat{f}, f)/VSISE(\hat{f}, f)$, $2KLL(\hat{f}, f)/VSISE(\hat{f}, f)$, and $KLL(\hat{f}, f)/2HD(\hat{f}, f)$ converge to unity as the number n of samples increases.

Table 7. The ratios $4HD(\hat{f}, f)/VSISE(\hat{f}, f)$, $2KLL(\hat{f}, f)/VSISE(\hat{f}, f)$, and $KLL(\hat{f}, f)/2HD(\hat{f}, f)$ after one hundred repetitions. The truncated normal on the respective intervals is used as f against the varying cell width histogram $\hat{f}$. The theoretical cell width is used except for the two outermost cells facing boundaries for $\hat{f}$.

ratio	n	on $[-1.960, 1.960]$		on $[-2.576, 2.576]$		on $[-3.291, 3.291]$	
		mean	s.d.	mean	s.d.	mean	s.d.
$\frac{4HD(\hat{f}, f)}{VSISE(\hat{f}, f)}$	100	0.9660338	0.1197109	0.8912677	0.1290322	0.6148133	0.1119335
	500	0.9720246	0.0607951	0.9538915	0.0760908	0.8640405	0.1021250
	1,000	0.9855243	0.0422619	0.9752948	0.0573177	0.9307278	0.0822681
$\frac{2KLL(\hat{f}, f)}{VSISE(\hat{f}, f)}$	100	0.9861332	0.1713458	0.9031914	0.1747174	0.5922517	0.1323772
	500	0.9693588	0.0801853	0.9597262	0.1040845	0.8609263	0.1257643
	1,000	0.9888618	0.0561294	0.9751281	0.0767958	0.9415178	0.1058661
$\frac{KLL(\hat{f}, f)}{2HD(\hat{f}, f)}$	100	1.0151140	0.0476992	1.0065790	0.0521547	0.9560272	0.0486380
	500	0.9960205	0.0204369	1.0039420	0.0286381	0.9931358	0.0312740
	1,000	1.0027950	0.0141518	0.9986794	0.0208402	1.0094440	0.0265545

Figure 1. The means and standard deviations of the three ratios $4HD/VSISE$, $2KLL/VSISE$ and $KLL/2HD$ after one hundred repetitions. The standard normal truncated on $[-1.960, 1.960]$ is used as true density against the kernel estimator using the Epanechnikov kernel.

Figure 2. The means and standard deviations of the three ratios $4HD/VSISE$, $2KLL/VSISE$ and $KLL/2HD$ after one hundred repetitions. The standard normal truncated on $[-2.576, 2.576]$ is used as true density against the kernel estimator using the Gaussian kernel.

- Convergence to unity requires larger numbers of samples as the support of $f(x)$ becomes wider.

The first result confirms the theorem. The second result obtains because the

Figure 3. The means and standard deviations of the three ratios $4HD/VSISE$, $2KLL/VSISE$ and $KLL/2HD$ after one hundred repetitions. The standard normal truncated on $[-3.291, 3.291]$ is used as true density against the kernel estimator using the rectangular kernel.

Figure 4. The means and standard deviations of the three ratios $4HD/VSISE$, $2KLL/VSISE$ and $KLL/2HD$ after one hundred repetitions. The standard normal truncated on $[-1.960, 1.960]$ is used as true density against the histogram with an universal cell width. Width of support $|I|$ is first divided by the theoretical cell width and the resulting theoretical number of cells is raised to the nearest whole number to obtain the number of cells for the histogram.

Figure 5. The means and standard deviations of the three ratios $4HD/VSISE$, $2KLL/VSISE$ and $KLL/2HD$ after one hundred repetitions. The standard normal truncated on $[-2.576, 2.576]$ is used as true density against the histogram with an universal cell width. Width of support $|I|$ is first divided by the theoretical cell width and the resulting theoretical number of cells is raised to the nearest whole number to obtain the number of cells for the histogram.

$\hat{f}(x)$ constructed with a smaller number of samples estimates $f(x)$ imprecisely when the support is wider and because $\hat{f}(x)$ in turn generates imprecise $4HD(\hat{f}, f)$, $2KLL(\hat{f}, f)$, and $VSISE(\hat{f}, f)$, resulting in slower convergence of their ratios to unity.

Figure 6. The means and standard deviations of the three ratios $4HD/VSISE$, $2KLL/VSISE$ and $KLL/2HD$ after one hundred repetitions. The standard normal truncated on $[-3.291, 3.291]$ is used as true density against the histogram with an universal cell width. Width of support $|I|$ is first divided by the theoretical cell width and the resulting theoretical number of cells is raised to the nearest whole number to obtain the number of cells for the histogram.

Second we summarize the results dependent on the estimator $\hat{f}(x)$ being the kernel estimator or histogram.

- The ratios $4HD(\hat{f}, f)/VSISE(\hat{f}, f)$, $2KLL(\hat{f}, f)/VSISE(\hat{f}, f)$, and $KLL(\hat{f}, f)/2HD(\hat{f}, f)$ fluctuate more for the estimator $\hat{f}(x)$ being the histogram than the kernel for the same density $f(x)$ with the same number of samples.
- All three ratios $4HD(\hat{f}, f)/VSISE(\hat{f}, f)$, $2KLL(\hat{f}, f)/VSISE(\hat{f}, f)$, and $KLL(\hat{f}, f)/2HD(\hat{f}, f)$ appear to converge to unity from below for the kernel estimator $\hat{f}$. In other words, both $4HD(\hat{f}, f)$ and $2KLL(\hat{f}, f)$ have a tendency to underestimate $VSISE(\hat{f}, f)$ and $KLL(\hat{f}, f)$ tends to underevaluate $HD(\hat{f}, f)$ for the kernel estimator $\hat{f}$.

Since the histogram is a coarser estimate of the underlying density between the two, the first result is understandable. The second result can be heuristically explained as follows. Expanding $4HD(\hat{f}, f)$ and $2KLL(\hat{f}, f)$ gives

$$(3.4) \quad 4HD(\hat{f}, f) \approx VSISE(\hat{f}, f) - \frac{1}{2} \int_I \frac{\{\hat{f}(x) - f(x)\}^3}{f(x)^2} dx + \frac{5}{16} \int_I \frac{\{\hat{f}(x) - f(x)\}^4}{f(x)^3} dx - \frac{7}{32} \int_I \frac{\{\hat{f}(x) - f(x)\}^5}{f(x)^4} dx,$$

and

$$(3.5) \quad 2KLL(\hat{f}, f) \approx VSISE(\hat{f}, f) - \frac{2}{3} \int_I \frac{\{\hat{f}(x) - f(x)\}^3}{f(x)^2} dx + \frac{2}{4} \int_I \frac{\{\hat{f}(x) - f(x)\}^4}{f(x)^3} dx - \frac{2}{5} \int_I \frac{\{\hat{f}(x) - f(x)\}^5}{f(x)^4} dx,$$

respectively. The sign of the leading term $\int_I \{\hat{f}(x) - f(x)\}^3 / f(x)^2 dx$ following $VSISE(\hat{f}, f)$ on the right-hand side in equations (3.4) and (3.5) could go either way, negative or positive, at first sight. For the kernel estimator $\hat{f}(x)$, a cumbersome but relatively straightforward calculation using the Edgeworth expansion for $\hat{f}(x)$ reveals that

$$(3.6) \quad E \int_I \frac{\{\hat{f}(x) - f(x)\}^3}{f(x)^2} dx \approx \frac{h^6}{8} \left\{ \int_I t^2 K(t) dt \right\}^3 \int_I \frac{f^{(2)}(x)^3}{f(x)^2} dx \\ + \frac{1}{(nh)^2} \int_I K(t)^3 dt \int_I f(x)^{-1/2} f(x)^{-1/2} dx,$$

where the first term on the right-hand side is on the order of magnitude $n^{-6/5}$ and the second term of $n^{-8/5}$. Therefore the sign of the term $\int_I f^{(2)}(x)^3/f(x)^2 dx$ within the first term on the right-hand side in equation (3.6) usually determines the sign of the expected value $E \int_I \{\hat{f}(x) - f(x)\}^3/f(x)^2 dx$ in equation (3.6). The term $\int_I f^{(2)}(x)^3/f(x)^2 dx$ takes value zero when $I = [-1.7758375, 1.7758375]$, is negative for narrower I , and positive for wider I for the normal $f(x)$. Since we run the simulation using the standard normal symmetrically truncated about the origin supported on $I = [-1.959964, 1.959964]$, $[-2.575829, 2.575829]$, and $[-3.290527, 3.290527]$ as our underlying densities $f(x)$, the leading term $\int_I \{\hat{f}(x) - f(x)\}^3/f(x)^2 dx$ following $VSISE(\hat{f}, f)$ must be positive on average for our simulation.

To confirm this, $E[\int_I \{\hat{f}(x) - f(x)\}^3/f(x)^2 dx]$ and the theoretically obtained term $h^6/8 \cdot \{\int_I t^2 K(t) dt\}^3 \int_I f^{(2)}(x)^3/f(x)^2 dx$ of largest order on the right hand side in equation (3.6) are computed for the standard normal symmetrically truncated about the origin supported on $I = [-1.96, 1.96]$ after one hundred repetitions. The result is summarized in table 8. Their signs are positive as expected and the simulation result tends to the theoretical value as the number of samples increases. Thus the ratios $4HD(\hat{f}, f)/VSISE(\hat{f}, f)$ and $2KLL(\hat{f}, f)/VSISE(\hat{f}, f)$ converge to unity from below for the kernel estimator $\hat{f}(x)$.

Assuming that expansions (3.4) and (3.5) hold, we know $2KLL(\hat{f}, f)$ oscillates more about $VSISE(\hat{f}, f)$ than $4HD(\hat{f}, f)$ does by virtue of the values of coefficients $-2/3$ against $-1/2$ for $\int_I \{\hat{f}(x) - f(x)\}^3/f(x)^2 dx$, $2/4$ against $5/16$ for $\int_I \{\hat{f}(x) - f(x)\}^4/f(x)^3 dx$, and $-2/5$ against $-7/32$ for $\int_I \{\hat{f}(x) - f(x)\}^5/f(x)^4 dx$ and so on. Since we know that the expected value of the leading term $\int_I \{\hat{f}(x) - f(x)\}^3/f(x)^2 dx$ is positive on average from the argument above, $KLL(\hat{f}, f)/2HD(\hat{f}, f)$ tends to converge to unity from below.

Finally expansions in (3.4) and (3.5) also reveal that the discrepancy between $KLL(\hat{f}, f)/2HD(\hat{f}, f)$ and unity must be smaller than those between $4HD(\hat{f}, f)/VSISE(\hat{f}, f)$ or $2KLL(\hat{f}, f)/VSISE(\hat{f}, f)$ and unity.

Third we compare the results among the three kernel estimators $\hat{f}(x)$.

Table 8. The effect of the leading term $\int_I \{\hat{f}(x) - f(x)\}^3/f(x)^2 dx$ following $VSISE(\hat{f}, f)$. The third column is the theoretically obtained term of largest order of magnitude for the term. The standard normal truncated on $[-1.960, 1.960]$ is used as f against the kernel estimator $\hat{f}$ using the Epanechnikov kernel.

n	$E \left[\int_I \frac{\{\hat{f}(x) - f(x)\}^3}{f(x)^2} dx \right]$	$\frac{h^6}{8} \left\{ \int_I t^2 K(t) dt \right\}^3 \int_I \frac{f^{(2)}(x)^3}{f(x)^2} dx$
100	0.0040456	0.0014538
500	0.0003659	0.0002107
1,000	0.0001179	0.0000917

- The ratios $4HD(\hat{f}, f)/VSISE(\hat{f}, f)$, $2KLL(\hat{f}, f)/VSISE(\hat{f}, f)$, and $KLL(\hat{f}, f)/2HD(\hat{f}, f)$ are marginally closer to 1 for the Epanechnikov than the Gaussian kernel when the support is narrow.
- The ratios $4HD(\hat{f}, f)/VSISE(\hat{f}, f)$, $2KLL(\hat{f}, f)/VSISE(\hat{f}, f)$, and $KLL(\hat{f}, f)/2HD(\hat{f}, f)$ are marginally closer to 1 for the Gaussian than the Epanechnikov kernel when the support is wide.
- The ratios $4HD(\hat{f}, f)/VSISE(\hat{f}, f)$, $2KLL(\hat{f}, f)/VSISE(\hat{f}, f)$, and $KLL(\hat{f}, f)/2HD(\hat{f}, f)$ are closer to 1 for both the Epanechnikov and Gaussian than for the rectangular kernel.

The first two results are explained as follows. The relative strength of the Epanechnikov kernel is that it is compactly supported, while that of the Gaussian is that it conforms to the shape of the underlying density $f(x)$ more closely except for the fact that $f(x)$ is truncated. As the support I increases from $[-1.960, 1.960]$ to $[-3.291, 3.291]$, the relative strength of the Epanechnikov kernel becomes more obscured while that of the Gaussian kernel becomes more apparent, generating the first two results. The third result comes from the fact that the kernel estimator $\hat{f}(x)$ using a rectangular kernel is coarser than the other two kernel estimators because the kernel itself is less smooth.

Lastly we find that whether the theoretical number of cells is rounded or raised to the nearest whole number or the theoretical cell width is used except for the two outermost cells facing boundaries does not seem to significantly affect the ratios $4HD(\hat{f}, f)/VSISE(\hat{f}, f)$, $2KLL(\hat{f}, f)/VSISE(\hat{f}, f)$, and $KLL(\hat{f}, f)/2HD(\hat{f}, f)$ for the histogram $\hat{f}(x)$.

4. Discussion

In density estimation, the integrated square error (ISE) is generally regarded as the *standard* measure of discrepancy between the density and estimator. There are several reasons. First, the ISE was the original measure of discrepancy used in Rosenblatt (1956), is mathematically tractable, and the mean square error is familiar to us. Second, the asymptotically mean ISE-optimal window width in (1.2) for the kernel estimator and cell width in (1.3) for the histogram do not contain the width of the support for the density. This makes the ISE especially attractive because the asymptotically optimal size of the smoothing parameter can be computed for infinitely-supported densities such as the Gaussian density. Third, a remarkable result of Stone (1984a) showed that least-squares cross-validation (LSCV) achieved asymptotically the best choice of window width for the kernel estimator in the sense of minimizing the ISE, assuming only f is bounded, and under very mild conditions on the kernel. See Silverman (1986, pp. 50–51) as well. A result on the histogram in Stone (1984b) showed, however, that asymptotically the best cell width was attained only if the LSCV was applied to compactly supported densities.

Arguments made against the use of the information-theoretic KLL and its accompanying data-based methods of likelihood cross-validation (LCV) for the kernel estimator and of AIC for the histogram are as follows. First, the KLL, though proposed relatively early in Kullback (1959), was purpose-built for

discrimination and thus is arguably an inappropriate measure of discrepancy because $\hat{f}$ and f are not interchangeable. Second, the asymptotically mean KLL-optimal window width in (1.4) for the kernel estimator and cell width in (1.5) for the histogram include width $|I|$ of the support for the density. The KLL should therefore be used only for densities with compact support, which excludes the Gaussian density. Third, the tail sensitive nature of the KLL has been criticized. Schuster and Gregory (1981) found that the LCV would not select consistent estimates of the density for long tailed distributions such as the double exponential and Cauchy distributions. Hall (1987) showed that the KLL's asymptotic properties were profoundly influenced by the tail properties of the kernel and of the density. Broniatowski, Deheuvels, and Devroye (1989) found that the LCV was non-consistent whenever the tails of f decrease as an exponential or slower.

We would argue, however, that the HD, KLL, or their asymptotic analog, the VSISE in (2.1), are a reasonable alternative to the more conventional ISE for the following reasons.

First, the argument that neither the KLL nor VSISE is mathematical distance and therefore not suitable for density estimation cannot be justified asymptotically in light of theorem 1, because $\hat{f}$ and f are interchangeable in the Hellinger distance.

The second reason, far more significant in our judgment, has to do with the question of whether one can afford to have the density estimator fluctuate more in one region than in the other. We shall make our case using the following least-squares estimation situation in which the variances in error terms vary with observations. It is of accepted practice to adjust the residual sum of squares by multiplying the inverse of its heteroscedastic, and thereby non-identity, variance-covariance matrix to obtain the best linear unbiased estimator. We believe that the same principle needs to be employed in density estimation, resulting in the use as our measure of the discrepancy

$$(4.1) \quad \int \frac{\{\hat{f}(x) - f(x)\}^2}{V(\hat{f}(x))} dx.$$

For the window width h , variance of the kernel estimator $\widehat{f_{\text{kernel}}}(x)$ is

$$V(\widehat{f_{\text{kernel}}}(x)) = \frac{f(x)}{nh} \int K(t)^2 dt + O(n^{-1}).$$

For the cell width h , variance of the histogram $\widehat{f_{\text{histogram}}}(x)$ is

$$V(\widehat{f_{\text{histogram}}}(x)) = \frac{f(x)}{nh} + O(n^{-1}).$$

If we assume the smoothing parameter does not vary with x , the variance $V(\hat{f}(x))$ in (4.1), whether $\hat{f}(x)$ is the kernel estimator or histogram, is proportional to $f(x)$. We obtain (2.1) by substituting $f(x)$ for $V(\hat{f}(x))$.

As a consequence, the density estimators constructed by LCV or AIC properly discount the contributions from high density regions where the varia-

bilities are also high by giving there the weight $1/f(x)$ to stabilize the variance over the support, a property from which the name “variance-stabilizing” in (2.1) originated.

The more conventional ISE, neglecting this essential weighting operation, fails in this respect. In other words, data-based methods of LSCV for the kernel estimator and histogram based on the ISE can be applied to infinitely supported densities precisely because the ISE neglects this weighting operation and concentrates on the high density region.

We believe that the seemingly limited applicability of data-based methods of LCV for the kernel estimator and of AIC for the histogram to compactly supported densities ought not to discourage us from using them as tools for exploring data because, as Chambers, Cleveland, Kleiner, and Tukey (1983) pointed out: “...Real data can never come from a genuine normal distribution, for that would require data... with the possibility of including arbitrarily large and small values. In practical terms,... they are bounded above and below...”

Finally we would like to point out the connection of the equivalence result with goodness of fit tests. In the test of goodness of fit of a distribution that assigns positive probability p_i to cell i where $i=1, \dots, m$, let $\hat{p}_i$ be the relative frequency in the i -th cell based on a sample of size n . Then it is well known that Pearson’s chi-square statistic

$$n \sum_{i=1}^m \frac{(\hat{p}_i - p_i)^2}{p_i},$$

likelihood ratio statistic

$$2n \sum_{i=1}^m \hat{p}_i \log\left(\frac{\hat{p}_i}{p_i}\right),$$

and rootgram test statistic

$$4n \sum_{i=1}^m (\sqrt{\hat{p}_i} - \sqrt{p_i})^2$$

are asymptotically equivalent. See Rice (1988, pp. 282-292) and Bishop, Fienberg, and Holland (1975, pp. 505). In light of the asymptotic relationship for the multinomials, the equivalence result should perhaps be expected when the density estimation is nonparametric.

5. Proofs

To prove the theorem we use the following lemma.

LEMMA 1 *Let K1-K4 and D1-D3 be satisfied. Suppose that $h \rightarrow 0$ and $(nh)^{-1} \log n \rightarrow 0$. Then*

$$\sup_{x \in I} \left| \frac{\hat{f}(x) - f(x)}{f(x)} \right| = o_p(1).$$

Proof of the lemma. There exists a constant C such that $f(x) \geq C > 0$ on I from D2. Hence

$$\sup_{x \in I} \left| \frac{\hat{f}(x) - f(x)}{f(x)} \right| \leq \frac{1}{C} \sup_{x \in I} |\hat{f}(x) - f(x)|.$$

It suffices to show $\sup_{x \in I} |\hat{f}(x) - f(x)| = o_p(1)$. For the kernel estimator $\hat{f}(x)$, this follows from equations (49) and (51) of Prakasa Rao (1983, pp. 50). For the histogram, this follows from theorem 2.3.1 (ibid. pp. 97-98) because D1 and D3 together imply f is uniformly continuous with bounded absolute moment. $\square$

Proof of the theorem. It suffices to show

$$(5.1) \quad HD(\hat{f}, f) = \frac{1}{4} VSISE(\hat{f}, f) + o_p(VSISE(\hat{f}, f)),$$

$$(5.2) \quad KLL(\hat{f}, f) = \frac{1}{2} VSISE(\hat{f}, f) + o_p(VSISE(\hat{f}, f)).$$

Let us first make some definitions. For any $x \in I$ define

$$g(x) = \frac{\hat{f}(x) - f(x)}{f(x)}.$$

For any $0 < \varepsilon < 1$ define the set A as $A = \{x \in I : g(x) \geq -\varepsilon\}$. Also denote the indicator function of a set A as $1_A(x)$.

Let us first prove (5.1). Since

$$\begin{aligned} A^c &= \{x \in I : g(x) < -\varepsilon\} \subseteq \{x \in I : |g(x)| > \varepsilon\} \\ &\subseteq \{x \in I : \sup_{x \in I} |g(x)| > \varepsilon\}, \end{aligned}$$

we have

$$1_{A^c}(x) \leq 1_{\{\sup_{x \in I} |g(x)| > \varepsilon\}}(x).$$

So

$$\begin{aligned} (5.3) \quad \int_{A^c \cap I} \{\hat{f}(x)^{1/2} - f(x)^{1/2}\}^2 dx &= \int_I 1_{A^c}(x) \cdot \{\hat{f}(x)^{1/2} - f(x)^{1/2}\}^2 dx \\ &\leq \int_I 1_{\{\sup_{x \in I} |g(x)| > \varepsilon\}}(x) \cdot \{\hat{f}(x)^{1/2} - f(x)^{1/2}\}^2 dx \\ &= 1_{\{\sup_{x \in I} |g(x)| > \varepsilon\}}(x) \cdot HD(\hat{f}, f) = o_p(HD(\hat{f}, f)) \end{aligned}$$

from lemma 1. Similarly we have

$$(5.4) \quad \int_{A^c \cap I} \frac{\{\hat{f}(x) - f(x)\}^2}{f(x)} dx = o_p(VSISE(\hat{f}, f)).$$

Therefore it suffices to show

$$(5.5) \quad \int_A \{\hat{f}(x)^{1/2} - f(x)^{1/2}\}^2 dx \stackrel{p}{=} \frac{1}{4} \int_A \frac{\{\hat{f}(x) - f(x)\}^2}{f(x)} dx.$$

Let $x \in A$ so that $\{1 + g(x)\}^{1/2}$ is well defined. Then expanding the term with the remainder in integral form gives

$$\{1+g(x)\}^{1/2}=1+\frac{g(x)}{2}-\frac{g(x)^2}{8}+\frac{3}{16}\int_0^{g(x)}\{g(x)-t\}^2(1+t)^{-5/2}dt.$$

Hence we have

$$\begin{aligned} & \int_A \{\hat{f}(x)^{1/2}-f(x)^{1/2}\}^2 dx \\ &= \int_A [\hat{f}(x)+f(x)-2\{1+g(x)\}^{1/2}f(x)] dx \\ &= \frac{1}{4} \int_A \frac{\{\hat{f}(x)-f(x)\}^2}{f(x)} dx \\ & \quad + \frac{3}{8} \int_A \int_0^{g(x)} \{g(x)-t\}^2(1+t)^{-5/2} dt \cdot f(x) dx. \end{aligned}$$

Therefore (5.5) follows because

$$\begin{aligned} & \left| \int_A \int_0^{g(x)} \{g(x)-t\}^2(1+t)^{-5/2} dt \cdot f(x) dx \right| \\ & \leq \int_A \sup_{0 \leq t \leq g(x)} \{(1+t)^{-5/2}\} \int_0^{g(x)} \{g(x)-t\}^2 dt \cdot f(x) dx \\ & \leq \frac{(1-\varepsilon)^{-5/2}}{3} \int_A \{g(x)\}^3 f(x) dx \\ & \leq \frac{(1-\varepsilon)^{-5/2}}{3} \cdot \sup_{x \in I} \left| \frac{\hat{f}(x)-f(x)}{f(x)} \right| \cdot VSISE(\hat{f}, f) \\ & = o_p(VSISE(\hat{f}, f)) \end{aligned}$$

from lemma 1. From (5.3), (5.4) and (5.5), we obtain (5.1).

Let us next prove (5.2). Similar to (5.3) it can be shown that

$$(5.6) \quad \int_{A^c \cap I} f(x) \log \frac{f(x)}{\hat{f}(x)} dx = o_p(KLL(\hat{f}, f)).$$

Therefore it suffices to show that

$$(5.7) \quad \int_A f(x) \log \frac{f(x)}{\hat{f}(x)} dx \stackrel{p}{=} \frac{1}{2} \int_A \frac{\{\hat{f}(x)-f(x)\}^2}{f(x)} dx.$$

Let $x \in A$ so that the $\log\{1+g(x)\}$ is well defined. Then expanding the term with the remainder in integral form gives

$$\log\{1+g(x)\}=g(x)-\frac{g(x)^2}{2}+\int_0^{g(x)}\{g(x)-t\}^2(1+t)^{-3}dt.$$

Hence we have

$$\begin{aligned} \int_A f(x) \log \frac{f(x)}{\hat{f}(x)} dx &= - \int_A f(x) \log\{1+g(x)\} dx \\ &= \frac{1}{2} \int_A \frac{\{\hat{f}(x)-f(x)\}^2}{f(x)} dx \\ & \quad + \int_A \int_0^{g(x)} \{g(x)-t\}^2(1+t)^{-3} dt \cdot f(x) dx. \end{aligned}$$

Therefore (5.7) follows because

$$\begin{aligned} & \left| \int_A \int_0^{g(x)} \{g(x) - t\}^2 (1+t)^{-3} dt f(x) dx \right| \\ & \leq \frac{(1-\varepsilon)^{-3}}{3} \cdot \sup_{x \in \gamma} \left| \frac{\hat{f}(x) - f(x)}{f(x)} \right| \cdot VSISE(\hat{f}, f) \\ & = o_p(VSISE(\hat{f}, f)). \end{aligned}$$

From (5.4), (5.6) and (5.7), we obtain (5.2).

REFERENCES

- [1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *2nd International Symposium on Information Theory* (eds. B. N. Petrov and P. Czaki), 267-281, Akademiai Kiadó, Budapest.
- [2] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*, MIT press, Cambridge, MA.
- [3] Broniatowski, M., Deheuvels, P. and Devroye, L. (1989). On the relationship between stability of extreme order statistics and convergence of the maximum likelihood kernel density estimate, *The Annals of Statistics*, **17**, 1070-1086.
- [4] Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical methods for data analysis*, Wadsworth International Group & Duxbury Press, Belmont, California & Boston, Massachusetts.
- [5] Cody, W. J. (1988). *ACM Transactions on Mathematical Software*, **14**, 303-311.
- [6] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The L_1 view*, John Wiley & Sons, New York.
- [7] Freedman, D. and Diaconis, P. (1981). On the histogram as a density estimator: l_2 theory, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, **57**, 453-476.
- [8] Hall, P. (1987). On Kullback-Leibler loss and density estimation, *The Annals of Statistics*, **15**, 1491-1519.
- [9] Hall, P. (1990). Akaike's information criterion and Kullback-Leibler loss for histogram density estimation, *Probability Theory and Related Fields*, **85**, 449-467.
- [10] Kanazawa, Y. (1988). An optimal variable cell histogram, *Communications in Statistics, Part A: Theory and Methods*, **17**, 1401-1422.
- [11] Kanazawa, Y. (1992). An optimal variable cell histogram based on the sample spacings, *The Annals of Statistics*, **20**, 291-304.
- [12] Kanazawa, Y. (1993a). Hellinger distance and Akaike's information criterion for the histogram, *Statistics & Probability Letters*, **17**(4), 293-298.
- [13] Kanazawa, Y. (1993b). Hellinger distance and Kullback-Leibler loss for the kernel density estimator, *Statistics & Probability Letters*, **18**(4), 315-321.
- [14] Kogure, A. (1987). Asymptotically optimal cells for a histogram, *The Annals of Statistics*, **15**, 1023-1030.
- [15] Kullback, S. (1959). *Information Theory and Statistics*, John Wiley & Sons, New York.
- [16] Parzen, E. (1962). On estimation of a probability density function and mode, *The Annals of Mathematical Statistics*, **33**, 1065-1076.
- [17] Prakasa Rao, B. L. S. (1983). *Nonparametric Functional Estimation*, Academic Press, Orlando, Florida.
- [18] Press, W. H., Teukolsky, S. A., Vetterling, W. T. and Flannery, B. P. (1992). *Numerical Recipes in C, The Art of Scientific Computing, Second Edition*, Cambridge University Press, Cambridge.
- [19] Rice, J. A. (1988). *Mathematical Statistics and Data Analysis*, The Wadsworth & Brooks, Monterey, Calif.
- [20] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function, *The Annals of Mathematical Statistics*, **27**, 832-837.
- [21] Schuster, E. F. and Gregory, G. G. (1981). On the nonconsistency of maximum likelihood nonparametric density estimators, *Computer Science and Statistics, Proceedings of the 13th Symposium on the Interface* (ed. W. F. Eddy), 295-298, Springer Verlag, New York.
- [22] Scott, D. W. (1979). On optimal and data-based histograms, *Biometrika*, **66**, 605-610.

- [23] Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.
- [24] Stone, C. J. (1984a). An asymptotically optimal histogram selection rule, *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (eds. L. M. Le Cam and R. A. Olshen), **2**, 513-520, Wadsworth, Monterey, California.
- [25] Stone, C. J. (1984b). An asymptotically optimal window selection rule for kernel density estimates, *The Annals of Statistics*, **12**, 1285-1297.
- [26] Taylor, C. C. (1987). Akaike's information criterion and the histogram, *Biometrika*, **74**, 636-639.