Probab. Th. Rel. Fields 85, 449–467 (1990)

Probability Theory and Related Fields

Akaike's information criterion and Kullback-Leibler loss for histogram density estimation

Peter Hall

Department of Statistics, The Australian National University, GPO Box 4, Canberra, ACT 2601, Australia

Received September 26, 1988; in revised form November 10, 1989

Summary. We take a detailed look at Akaike's information criterion (*AIC*) and Kullback-Leibler cross-validation (*KLCV*) in the problem of histogram density estimation. Two different definitions of "number of unknown parameters" in *AIC* are considered. A careful description is given of the influence of density tail properties on performance of both types of *AIC* and on *KLCV*. A number of practical conclusions emerge. In particular, we find that *AIC* will often give problems when used with heavy-tailed unbounded densities, but can perform quite well with compactly supported densities. In the latter case, both types of *AIC* produce similar results, and those results will sometimes be asymptotically equivalent to the ones obtained from *KLCV*. However, depending on the shape of the true density, the *KLCV* method can fail to balance "bias" and "variance" components of loss, with the result that *KLCV* and *AIC* may produce very different results.

1. Introduction

Methods founded on information theory have recently received considerable attention in statistics. One of the more commonly used definitions of "information" is the notion of Kullback-Leibler loss (Kullback [15]), which describes the amount of information in a single observation from one distribution for discriminating against another distribution. The associated method of Kullback-Leibler cross-validation is a popular tool in certain types of nonparametric density estimation, for discrimination and other problems [2, 3, 4, 7, 8, 9, 10, 11, 15, 17, 20]. Other definitions of information include Akaike's criterion [1] and Rissanen's [18] concepts of stochastic complexity and minimum description length. Akaike's criterion has recently been applied to the problem of selecting bin width for histogram density estimators [22]. The purpose of the present paper is to take a detailed look at Akaike's information criterion (*AIC*) in the setting of histogram density estimators, and to relate it to Kullback-Leibler loss and Kullback-Leibler cross-validation (*KLCV*).

AIC works by adding to the negative of the estimated loglikelihood a term equal to the "number of unknown parameters", and minimizing this quantity rather than the negative estimated likelihood. In the present paper we consider two different interpretations of "number of unknown parameters", J – one being the number of nonempty histogram bins, the other being the total number of bins (empty or nonempty) between the smallest and the largest data point. We show that each interpretation produces meaningful results provided the underlying distribution has sufficiently many finite moments. However, both will fail if certain low-order moments are infinite, and in those cases *AIC* is not workable.

We analyse two different classes of distribution – distributions with bounded support, and unbounded distributions with regularly varying tails. On a coarse scale, our main conclusions are the following. When the distribution is bounded, both definitions of J yield similar results, Kullback-Leibler loss is well-defined, and the *AIC* method usually produces results similar to *KLCV*. In the case of unbounded distributions, the different versions of J produce very different results, and neither Kullback-Leibler loss nor *KLCV* is well-defined in large samples.

In finer detail, our main conclusions are as follows.

(i) Bounded distributions

Assume that the true density f is supported on the interval $[0, 1]$, and that $f(x)$ and $f(1-x)$ behave like x^{α_1} and x^{α_2}, respectively, as $x \downarrow 0$, where $\alpha_1, \alpha_2 \geqq 0$. Divide $[0, 1]$ into m equal-width histogram bins. Both *AIC* and *KLCV* arrive at the choice of an optimal m by minimizing respective functions of m. These functions may each be divided into "bias" and "variance" components. When *AIC* and *KLCV* are both well-defined and finite, the respective bias and variance terms are essentially of the same size. In each case, minimization of the criterion "should" involve balancing bias against variance, and so both techniques "should" produce similar values of m. However, this argument fails in circumstances where *AIC* is well-defined but *KLCV* is not. The *AIC* function is well-defined for *all* possible choices of m, whereas *KLCV* is only defined when m is chosen sufficiently small for each bin to contain at least two data points. Sometimes the latter condition can only be fulfilled by choosing m so small that it is impossible to achieve balance between bias and variance components of *KLCV*. In this circumstance, *AIC* manages to balance bias against variance but *KLCV* does not. The value of m selected by *KLCV* is then very close to the largest one compatible with each histogram bin containing at least two data points, and is of a smaller order of magnitude than the m selected by *AIC*.

In particular, this type of behaviour occurs when the tail parameters α_1 and α_2 satisfy $\min(\alpha_1, \alpha_2) > 1$ and $\max(\alpha_1, \alpha_2) > 2$. On the other hand, there are many cases where the "low bin count" problem does not arise, for example when the optimal m is such as to ensure high counts in all bins. In such circumstances, *AIC* and *KLCV* produce asymptotically equivalent histogram estimators. A case in point is that where $0 < \min(\alpha_1, \alpha_2) < 1$ and $\max(\alpha_1, \alpha_2) < \min(\alpha_1, \alpha_2) + 1$.

In the case of bounded distributions, *KLCV* makes a very good attempt at selecting that value of m which minimizes Kullback-Leibler loss, L. In all cases, the ratio of the cross-validatory m to that which minimizes L converges weakly to a proper limit distribution having no atom at zero. The limit is unity when the "optimal" m is sufficiently small for there to be no problem with low bin counts.

(ii) Unbounded distributions

Assume that the true density f is positive on $(-\infty, \infty)$ and that its upper (lower) tails decrease like $x^{-\alpha_1}(|x|^{-\alpha_2})$ respectively. Define $\alpha = \min(\alpha_1, \alpha_2)$. If J is taken equal to the number of nonempty bins then the condition $\alpha > \alpha_0 \simeq 2.5$ (where α_0 is defined in Sect. 2) is necessary and sufficient for acceptable performance of *AIC*. Using the other criterion for J, $\alpha > 2$ is necessary and sufficient for acceptable performance. Here, "acceptable" means that the bin width h produced by *AIC* converges to zero at a rate between $n^{-1+\delta}$ and $n^{-\delta}$ for some $\delta > 0$. Should this condition fail, the pointwise mean squared error of the histogram estimator is of larger order than $n^{-\varepsilon}$ for each $\varepsilon > 0$, which is exceedingly poor. (Note too that the conditions $h \to 0$ and $nh \to \infty$ are necessary and sufficient for consistent density estimation by the histogram method.)

Our results also show that when both versions of J perform acceptably, the bin width selected using the "number of nonempty bins" criterion for J is of considerably smaller order of magnitude than that chosen using the other criterion. Neither criterion ever gives a smoothing parameter of the order which minimizes L^2 loss; they smooth much more than is optimal there.

Work on Kullback-Leibler loss and cross-validation for kernel-type density estimators includes Habbema et al. [9], Duin [4], Chow et al. [3], Bowman [2] and Hall [11, 12]. The present paper appears to be the first to take a detailed look at these methods for histogram estimators. Taylor [22] was the first to treat *AIC* in the context of histogram estimators.

Properties of Kullback-Leibler loss and *KLCV* for histogram estimators and bounded distributions are described in Sect. 2. *AIC* for bounded distributions is discussed in Sect. 3. The conclusions in (i) are drawn from Sect. 2 and 3. Section 4 describes *AIC* for unbounded distributions, and leads to the conclusions in (ii). Proofs are given in Sect. 5.

2. Bounded distributions: Kullback-Leibler loss and cross-validation

2.1. Introduction

Assume that the underlying density f is supported on interval (0, 1). In this setting we shall describe properties of Kullback-Leibler loss, L, as a criterion for selecting the number m of bins in a histogram density estimator (see Subsect. 2.2), and discuss performance of the cross-validation criterion CV as a data-driven device for choosing that value of m which minimizes Kullback-Leibler loss (see Subsect. 2.3). The functions L and CV are defined at (2.3) and (2.4), respectively.

Divide the interval $[0, 1]$ into m bins, or cells, the i'th being $\mathscr{B}_i \equiv [(i-1)m^{-1}, im^{-1}]$ where $1 \leqq i \leqq m$. Let $\mathscr{X} \equiv \{X_1, \ldots, X_n\}$ be a random n-sample from f, N_i be the number of X_j's in bin $\mathscr{B}_i$, and $p_i \equiv \int_{\mathscr{B}_i} f = n^{-1} E(N_i)$ be the chance that an arbitrary X_j is bin $\mathscr{B}_i$. The histogram density estimator based on these bins is defined by

$$\hat{f}(x) \equiv mn^{-1} N_i \quad \text{for } x \in \mathscr{B}_i, \qquad 1 \leqq i \leqq m. \tag{2.1}$$

When discussing cross-validation it is convenient to introduce the estimator $\hat{f}_j$ constructed from the $(n-1)$-sample $\mathscr{X}_j \equiv \mathscr{X} \backslash \{X_j\}$. If N_{ij} denotes the number of values from $\mathscr{X}_j$ in bin $\mathscr{B}_i$, then

$$\hat{f}_j(x) \equiv m(n-1)^{-1} N_{ij} \quad \text{for} \quad x \in \mathscr{B}_i, \quad 1 \leqq i \leqq m. \tag{2.2}$$

Kullback-Leibler loss or "discrimination information" is defined by

$$L = L(m) \equiv \int f \log (f/\hat{f}). \tag{2.3}$$

It equals the average amount of information present in a single observation from f for discriminating against $\hat{f}$. The cross-validation criterion is defined by

$$CV = CV(m) \equiv -n^{-1} \sum_{i=1}^{m} N_i \log(N_i - 1) - \log m. \tag{2.4}$$

We use this version, with the minus signs, so that the cross-validatory m is found by minimizing rather than maximizing CV.

A quantity identical to CV up to terms which do not depend on m is

$$CV_1(m) \equiv -\sum_{k=1}^{n} \log \hat{f}_j(X_j).$$

Therefore minimizing CV is equivalent to maximizing an estimator of log-likelihood. Both CV and CV_1 may be derived by following the prescription given by Titterington [23, 24]. Note that CV and CV_1 are either infinite or undefined if some $N_i \leqq 1$, and L is infinite or undefined if some $N_i = 0$.

Assume $f(x)$ and $f(1-x)$ behave like x^{α_1} and x^{α_2}, respectively, as $x \downarrow 0$. Asymptotic theory for L and for CV is governed by values taken by the tail parameters α_1 and α_2. Some aspects are a little unusual and complex, so to make life easier for the reader we summarize the main features here. Both Kullback-Leibler loss and the cross-validation criterion have "bias components" which are of size $m^{-\gamma}$ for some $1 < \gamma \leqq 2$, or of size $m^{-2} \log m$, depending on α_1 and α_2. There is also a variance component whose natural order is m/n. This means that in principle, the optimal number of bins is found by balancing m/n against $m^{-\gamma}$ (or $m^{-2} \log m$), resulting in the optimal m being approximately $n^{1/(\gamma+1)}$ (or $(n \log n)^{1/3}$). However, the fact that each N_i must be positive in the case of Kullback-Leibler loss, or that each $N_i - 1$ must be positive in the case of the cross-validation criterion, results in this balance being unachievable in certain circumstances.

For example, suppose $\alpha \equiv \min(\alpha_1, \alpha_2) > 1$. Then the bias component is of size m^{-2}. Balancing m^{-2} against m/n entails $m \simeq n^{1/3}$. Now, the minimum of the expected numbers of data points in bins 1 through m is $\min_i np_i \simeq nm^{-(\beta+1)} \simeq n^{(2-\beta)/3}$, where $\beta \equiv \max(\alpha_1, \alpha_2)$. If $\beta > 2$ then this number converges to zero, implying that

$$P(N_i = 0 \text{ for some } i) \to 1$$

as $n \to \infty$. This means that if we take $m \simeq n^{1/3}$, as suggested above, then both L and CV take infinite or undefined values with probability converging to one. Therefore with high probability, *balance between variance and bias components cannot be achieved.*

In circumstances such as the one just described, minimum loss and maximum CV are achieved not by effecting a balance between bias and variance, but by increasing m to a number close to the largest value compatible with the relevant criterion being well-defined and finite! For example when $\alpha > 1$ and $\beta > 2$, the requirement that loss be finite dictates that $nm^{-(\beta+1)}$ not decrease to zero, meaning that the optimal m is of size $n^{1/(\beta+1)}$, not $n^{1/3}$. In this circumstance the bias component is of size $m^{-2} \simeq n^{-2/(\beta+1)}$, whereas the variance component is negligible at $m/n \simeq n^{-\beta/(\beta+1)} = o(n^{-2/(\beta+1)})$. The operations of minimizing L and minimizing CV both produce values of m of size $n^{1/(\beta+1)}$, but the ratio of the two values is not asymptotic to unity. This comes about because in the case of Kullback-Leibler loss, the condition for finiteness is $\min_i N_i \geqq 1$, whereas in the case of cross-validation it is $\min_i N_i \geqq 2$. Therefore minimizing CV is not necessarily asymptotically equivalent to minimizing Kullback-Leibler loss. However, there are cases where the operations of minimizing CV and minimizing L do produce asymptotically equivalent results; see Sect. 2.3 for details.

It is worth reiterating here one of the conclusions of this paper, which is that AIC avoids all these difficulties. Since AIC is well-defined and finite even when some bins are empty, then using AIC is *in all cases* asymptotically equivalent to balancing a variance term of size m/n against a bias term of size $m^{-\gamma}$ (or $m^{-2}\log m$).

2.2. *Kullback-Leibler loss*

Recall the definition of L at (2.3). Notice that in the case of histogram estimators there is a nonzero probability that $\hat{f}$ vanishes within any given bin $\mathscr{B}_i$. Therefore $E(L) = +\infty$ for all $n \geqq 1$ and $m \geqq 2$. This means that we must work directly with raw loss L, not expected loss $E(L)$.

Write $L = B + V$, where

$$B \equiv \int f \log(f/E\hat{f}) \quad \text{and} \quad V \equiv \int f \log(E\hat{f}/\hat{f}) \tag{2.5}$$

denote bias and variance components respectively. Convexity ensures that $B \geqq 0$. Of course, B depends only on m, not on n, and is purely deterministic. It may be treated by routine analytical means, so we dispose of it first.

Given $\gamma > 0$, define $a_i = a_i(\gamma)$ by

$$-i^{-(\gamma+1)}a_i \equiv \gamma(\gamma+1)^{-1}(1-i^{-1})^{\gamma+1}\log(1-i^{-1}) + \gamma(\gamma+1)^{-2}\{1-(1-i^{-1})^{\gamma+1}\}$$
$$+ (\gamma+1)^{-1}\{1-(1-i^{-1})^{\gamma+1}\}\log[(\gamma+1)^{-1}i\{1-(1-i^{-1})^{\gamma+1}\}],$$

interpreting $0\log 0$ as 0 in the case of a_1. It may be shown that $|a_i| = O(i^{\gamma-2})$ as $i \to \infty$, so that $\sum |a_i| < \infty$ for $0 < \gamma < 1$. Define

$$b(d,\gamma) \equiv d \sum_{i=1}^{\infty} a_i(\gamma) \quad \text{for} \quad d > 0 \text{ and } 0 < \gamma < 1\,.$$

Assume that the support of f equals $[0,1]$ and:

$$\begin{aligned}&f \text{ is nonzero on } (0,1);\ f' \text{ exists and is continuous on } (0,1);\ \text{either}\\ &f(x) \to c_1 \text{ or } f'(x) \sim c_1\alpha_1 x^{\alpha_1 - 1} \text{ as } x \downarrow 0, \text{ where } c_1, \alpha_1 > 0; \text{ and}\\ &\text{either } f(1-x) \to c_2 \text{ or } f'(1-x) \sim c_2\alpha_2 x^{\alpha_2 - 1} \text{ as } x \downarrow 0, \text{ where } c_2, \alpha_2 > 0\,.\end{aligned} \tag{2.6}$$

In the cases $f(x) \to c_1$ and $f(1-x) \to c_2$, define $\alpha_1 \equiv 0$ and $\alpha_2 \equiv 0$ respectively. Then $f(x) \sim c_1 x^{\alpha_1}$ and $f(1-x) \sim c_2 x^{\alpha_2}$ as $x \to 0$, which is the "tail parameter" model discussed in Sect. 1 and Subsect. 2.1. Put $\alpha \equiv \min(\alpha_1, \alpha_2)$, $\beta \equiv \max(\alpha_1, \alpha_2)$, $c \equiv c_i$ if $0 < \alpha_i < \alpha_j$, $c \equiv c_j$ if $0 = \alpha_i < \alpha_j$, and $c \equiv c_1 + c_2$ if $\alpha_1 = \alpha_2$.

Theorem 2.1. *Assume condition* (2.6), *and that* $m \to \infty$. *If* $0 < \alpha < 1$ *then* $B(m) \sim b(c, \alpha) m^{-(\alpha+1)}$; *if* $0 = \alpha < \beta < 1$ *then* $B(m) \sim b(c, \beta) m^{-(\beta+1)}$; *if* $\alpha = 1$, *or if* $\alpha = 0$ *and* $\beta = 1$, *then* $B(m) \sim (c/24) m^{-2} \log m$; *and if* $\alpha > 1$, *or if* $\alpha = \beta = 0$, *or if* $\alpha = 0$ *and* $\beta > 1$, *then*

$$B(m) \sim m^{-2} (1/24) \int (f')^2 f^{-1}. \tag{2.7}$$

Except for the three cases $\alpha > 1, \alpha = \beta = 0$, and $\alpha = 0$ and $\beta > 1$, where $B(m)$ satisfies formula (2.7), the integral on the right-hand side of (2.7) is infinite.

We now turn to the variance term V, defined at (2.5). Observe that $V = +\infty$ unless $\min_i N_i \geqq 1$. The next theorem describes V when bin counts satisfy this condition, and uses the following weaker version of condition (2.6):

f is bounded away from zero on $[\delta, 1-\delta]$ for each $0 < \delta < \frac{1}{2}$, and satisfies

$f(x) \sim c_1 x^{\alpha_1}$ and $f(1-x) \sim c_2 x^{\alpha_2}$ as $x \downarrow 0$, where $c_1, c_2 > 0$ and $\alpha_1, \alpha_2 \geqq 0$. (2.8)

Theorem 2.2. *Assume condition* (2.8). *Then if* $0 < \varepsilon < \frac{1}{2}$,

$$V = \tfrac{1}{2} m n^{-1} + o_p(m n^{-1}) \tag{2.9}$$

uniformly in values m *such that* $n^\varepsilon \leqq m \leqq n^{1-\varepsilon}$ *and* $\min_i N_i \geqq 1$, *as* $n \to \infty$.

Next we combine results (2.7) and (2.9). Recall that $\alpha \equiv \min(\alpha_1, \alpha_2)$ and $\beta \equiv \max(\alpha_1, \alpha_2)$. It is convenient to ignore the circumstance when either $\alpha = 1$ or $\alpha = 0$ and $\beta = 1$, where $B \sim \text{const.}\, m^{-2} \log m$. This case is easily treated separately. In all other situations, $B \sim C m^{-\gamma}$ where $C > 0$ and $1 < \gamma \leqq 2$. Results (2.7) and (2.9) entail

$$L = B + V = (C m^{-\gamma} + \tfrac{1}{2} m n^{-1}) \{1 + o_p(1)\}.$$

In principle, the asymptotically optimal value of m should be obtained by minimizing $C m^{-\gamma} + \frac{1}{2} m n^{-1}$, giving $m \sim (2 C \gamma n)^{1/(\gamma+1)}$. However, remember that formula (2.9) holds only for values of m such that $\min_i N_i \geqq 1$. Since

$$\min_i E(N_i) = n \min_i p_i \sim \text{const.}\, n m^{-(\beta+1)},$$

then a necessary and sufficient condition for $P(\min_i N_i \geqq 1)$ to converge to one is that $n m^{-(\beta+1)} \to \infty$ as $n \to \infty$. If $m \sim (2 C \gamma n)^{1/(\gamma+1)}$ then the latter condition is equivalent to $\beta < \gamma$. Should it be true that either (a) $0 < \alpha < 1$ and $\beta \geqq \alpha + 1$, or (b) $\alpha = 0$ or $\alpha > 1$, and $\beta \geqq 2$, then the condition $\beta < \gamma$ will be violated. In this circumstance the value of m, m^* say, which minimizes L will be asymptotic to $\min(m_0, M_0)$, where $m_0 \equiv (2 C \gamma n)^{1/(\gamma+1)}$ and M_0 denotes the largest m such that $\min_i N_i \geqq 1$. Of course, M_0 is a random variable.

Asymptotic behaviour of M_0 is easily described in terms of Poisson processes, as follows. Let $\mathscr{P}$ be a Poisson process on $(0, \infty)$ with intensity function $\lambda(x) \equiv c_j x^\beta$, where j is the index such that $\beta = \alpha_j$. Let $Z_k(t)$ denote the number of points of $\mathscr{P}$ contained within the interval $I_k(t) \equiv [(k-1)t, kt)$, where $0 < t < \infty$ and $k \geqq 1$. Define the random variable S by

$$S^{-1} \equiv \inf\{t>0 : Z_k(t) \geqq 1 \text{ for each } k \geqq 1\}. \tag{2.10}$$

Now, $\varrho(t) \equiv P\{Z_k(t) \geqq 1$ for each $k \geqq 1\} = \Pi_k\{1-\exp(-\mu_k)\}$, where μ_k equals the expected number of points of $\mathscr{P}$ in $I_k(t)$ and is dominated by const. k^β. The series test for convergence of an infinite product shows that $\varrho(t) \to 1$ as $t \to \infty$. Therefore $0 < S < \infty$ with probability one. In the case $\alpha \neq \beta$ it may be shown that under condition (2.8), $n^{-1/(\beta+1)} M_0 \to S$ in distribution. The proof is by Poisson approximation to the binomial distribution.

We are now in a position to give a concise description of the optimal m, m^*, in the exceptional cases (a) and (b) mentioned two paragraphs earlier.

(a) $0 < \alpha < 1$ and $\beta \geqq \alpha+1$: Here $\gamma = \alpha+1$, $m_0 = \{2C(\alpha+1)n\}^{1/(\alpha+2)}$, and

$$\begin{aligned} n^{-1/(\alpha+2)} m^* &\sim n^{-1/(\alpha+2)} \min(m_0, M_0) \\ &\to \begin{cases} S & \text{if } \beta > \alpha+1 \\ \min[\{2C(\alpha+1)\}^{1/(\alpha+2)}, S] & \text{if } \beta = \alpha+1 \end{cases} \end{aligned} \tag{2.11}$$

in distribution;

(b) $\alpha = 0$ or $\alpha > 1$, and $\beta \geqq 2$: Here $\gamma = 2$, $m_0 = (4Cn)^{1/3}$, and

$$\begin{aligned} n^{-1/3}\, m^* &\sim n^{-1/3} \min(m_0, M_0) \\ &\to \begin{cases} S & \text{if } \beta > 2 \text{ and } \alpha \neq \beta \\ \min\{(4C)^{1/3}, S\} & \text{if } \beta = 2 \text{ and } \alpha \neq \beta \end{cases} \end{aligned} \tag{2.12}$$

in distribution. (The case $\alpha = \beta$ is only slightly different. There, both tails of f play a role, and S should be defined in terms of two independent Poisson processes $\mathscr{P}_1$ and $\mathscr{P}_2$ with respective intensities $c_1 x^\beta$ and $c_2 x^\beta$.)

These peculiar properties of Kullback-Leibler loss arise solely because loss is infinite whenever one or more bins are empty. As we shall show in Sect. 3, *AIC* circumvents this pathological behaviour.

2.3. *Kullback-Leibler cross-validation*

Recall that CV was defined at (2.4). Many of the characteristics of CV are also those of L. For example, up to terms which do not depend on m and so play no role in the operation of minimization, CV and L have virtually identical asymptotic properties, as the following theorem shows.

Theorem 2.3. *Assume condition* (2.6). *Then if* $0 < \varepsilon < \frac{1}{2}$,

$$\begin{aligned} CV + \log(n-1) + n^{-1} \sum_{j=1}^{n} \log f(X_j) &= L + o_p(B + mn^{-1}) \\ &= B + \tfrac{1}{2} mn^{-1} + o_p(B + mn^{-1}) \end{aligned}$$

uniformly in values of m such that $n^\varepsilon \leqq m \leqq n^{1-\varepsilon}$ and $\min_i N_i \geqq 2$.

The restriction "$n^\varepsilon \leqq m \leqq n^{1-\varepsilon}$" imposed in Theorem 2.3 hardly matters, for it is now clear that the value of m which we seek lies within this range if ε is sufficiently small. However, the constraint "$\min_i N_i \geqq 2$" is very important. Recall from Subsect. 2.2 that in certain circumstances, the value of m which minimizes

Kullback-Leibler loss is asymptotic to the largest m such that $\min_i N_i \geqq 1$. When we minimize CV in such cases, we obtain that value of m which is asymptotic to the largest m such that $\min_i N_i \geqq 2$. This difference can have a significant effect. For example it means that in cases (a) and (b) considered earlier, the value $m^\dagger$ which minimizes CV satisfies the following analogues of (2.11) and (2.12) respectively:

$$n^{-1/(\alpha+2)} m^\dagger \to \begin{cases} S' & \text{if} \quad \beta > \alpha + 1 \\ \min[\{2C(\alpha+1)\}^{1/(\alpha+2)}, S'] & \text{if} \quad \beta = \alpha + 1, \end{cases} \tag{2.13}$$

$$n^{-1/3} m^\dagger \to \begin{cases} S' & \text{if } \beta > 2 \text{ and } \alpha \neq \beta \\ \min\{(4C)^{1/3}, S'\} & \text{if } \beta = 2 \text{ and } \alpha \neq \beta, \end{cases} \tag{2.14}$$

where S' is defined as was S at (2.10) except that the inequality $Z_k(t) \geqq 1$ there should now be replaced by $Z_k(t) \geqq 2$.

We conclude from this discussion that in cases (a) and (b), the ratio $m^\dagger/m^*$ of the m chosen by cross-validation to the m which minimizes loss does not converge to unity. However, the ratio does converge weakly to a proper limiting distribution (defined by ratios of right-hand sides of (2.11) to (2.14)) with no atom at zero. Therefore at the very worst, cross-validation picks a bin width which is of the right order of magnitude from the point of view of minimizing loss. The ratio $m^\dagger/m^*$ does converge to unity in all the other cases considered in Subsect. 2.2 – there, cross-validation selects a bin width which is asymptotically optimal from the point of view of minimizing loss.

It is worth noting that the values of m which minimize L^1 and L^2 distance between $\hat{f}$ and f, are of size $n^{1/3}$; see Scott [19] and Freedman and Diaconis [5]. The only cases where CV would give an m of this size are $1 < \alpha \leqq \beta \leqq 2$; $\alpha = 0$ and $1 < \beta \leqq 2$; $0 < \alpha \leqq 1$ and $\beta = 2$; and $\alpha = \beta = 0$.

3. Bounded distributions: Akaike's information criterion

Adopt notation introduced in Sect. 2. In particular, assume f is supported on interval $(0, 1)$, let m be the number of histogram bins in the interval $[0, 1]$, and let N_i denote the number of data points in bin i. As shown by Taylor [22], Akaike's information criterion asks that m be selected so as to minimize

$$AIC = AIC(m) \equiv n^{-1} J - n^{-1} \sum_{i=1}^{m} N_i \log N_i - \log m, \tag{3.1}$$

where J, a random variable, may be taken to equal either the number of nonempty bins or the total number of bins between the smallest-index nonempty bin and the largest-index nonempty bin. Unlike either Kullback-Leibler loss L or the cross-validatory criterion CV, the function AIC is well-defined even when one or more N_i's are zero.

Let B, the bias component of both L and CV, be as at (2.5), and remember that the variance component of both L and CV is asymptotic to $\frac{1}{2} mn^{-1}$ in those cases where the respective function is well-defined. Our next theorem shows that, up to terms which do not depend on m and so play no role in the operation of minimization, the function AIC is also equivalent to $B + \frac{1}{2} mn^{-1}$.

Theorem 3.1. *Assume condition* (2.6). *Then if* $0<\varepsilon<\frac{1}{2}$,

$$AIC+\log n+n^{-1}\sum_{j=1}^{n}\log f(X_j)=B+\tfrac{1}{2}mn^{-1}+o_p(B+mn^{-1})$$

uniformly in values of m such that $n^{\varepsilon}\leqq m\leqq n^{1-\varepsilon}$.

Theorems 2.3 and 3.1 are directly comparable. The main feature which distinguishes them is that in Theorem 3.1, no conditions are imposed on values of the N_i's. Therefore when minimizing the function AIC we simply balance the bias term, B, against the variance term, $\frac{1}{2}mn^{-1}$, without being bothered by problems of empty or near-empty bins. For example, consider cases (a) and (b) from Subsect. 2.2, which caused so much trouble when we minimized L and CV. The value $\hat{m}$ of m which minimizes AIC satisfies $\hat{m}\sim\{2C(\alpha+1)n\}^{1/(\alpha+1)}$ in case (a) and $\hat{m}\sim(4Cn)^{1/3}$ in case (b).

4. Unbounded distributions: Akaike's information criterion

Assume that the underlying of f *satisfies*

$$\begin{aligned}&f>0 \text{ on } (-\infty,\infty),\ f' \text{ exists and is continuous on } (-\infty,\infty),\\ &\text{and for constants } c_1,c_2>0 \text{ and } \alpha_1,\alpha_2>1,\ f'(x)\sim -c_1\alpha_1x^{-\alpha_1-1}\\ &\text{and } f'(-x)\sim -c_2\alpha_2x^{-\alpha_2-1} \text{ as } x\to+\infty\,.\end{aligned}\tag{4.1}$$

(Thus, $f(x)\sim c_1x^{-\alpha_1}$ and $f(-x)\sim c_2x^{-\alpha_2}$ as $x\to+\infty$, so that upper and lower tails of f vary regularly with exponents α_1 and α_2 respectively.) Let h denote bin width for the histogram estimator, and let the i'th bin be $\mathscr{B}_i\equiv((i-1)h, ih]$ where $-\infty<i<\infty$. (Identical asymptotic results are obtainable for AIC applied to variable-centre bins, but we avoid that version so as to minimize technicalities.) Let N_i denote the number of values from a random sample $X_1,\ldots,X_n$ which fall into bin $\mathscr{B}_i$, so that $\sum N_i=n$. Our histogram estimator of the unknown density f of the X_j's is

$$\hat{f}(x)\equiv N_i/nh\,,\qquad \text{for } x\in\mathscr{B}_i\,.$$

Let $M(x)$ denote a Poisson-distributed random variable with mean x, and put

$$D(\alpha)\equiv\alpha^{-1}\int_0^{\infty}x^{-1/\alpha}E(\log[x^{-1}\{M(x)+1\}])dx\,,\qquad \alpha>1\,,$$

and $D_k\equiv c_k^{1/\alpha}D(\alpha_k)$. An equivalent definition of D_k is

$$D_k\equiv\int_0^{\infty}E[M(c_kx^{-\alpha_k})\log\{M(c_kx^{-\alpha_k})/c_kx^{-\alpha_k}\}]dx\,,$$

from which it follows by convexity that $D_k>0$.

Our first theorem in this section describes asymptotic behaviour of the estimated loglikehood. It is basic to elucidation of AIC. Given $\delta\in(0,\frac{1}{2})$, let $\mathscr{H}_n\equiv\{m^{-1}: m \text{ is an integer and } n^{\delta}\leqq m\leqq n^{1-\delta}\}$. As in Sect. 3, our bin widths h will be chosen from this class. Our results may be extended to any collection $\mathscr{H}_n$ such that $n^{-1+\delta}\leqq h\leqq n^{-\delta}$ for each $h\in\mathscr{H}_n$ and $\#\mathscr{H}_n=O(n^c)$ for some $c>0$, as in Stone [21]. However at several places in the proof, such as the derivation of (5.10), this requires calculation of high-order multinomial moments, and that takes very lengthy algebra. The restriction $n^{-1+\delta}\leq h\leq n^{-\delta}$ is commonly made in work of this type – see e.g. Marron [16] – and anyway, consistency of $\hat{f}$ demands $h\to 0$ and $nh\to\infty$. Define

$$b\equiv(1/24)\int(f')^2 f^{-1}>0\,.$$

Theorem 4.1. *Assume condition* (4.1). *Then*

$$\hat{L}\equiv n^{-1}\sum_{j=1}^{n}\log\hat{f}(X_j)=n^{-1}\sum_{j=1}^{n}\log f(X_j)+\sum_{k=1}^{2}D_k(nh)^{-1+(1/\alpha_k)}-h^2 b$$
$$+o_p\left\{\sum_{k=1}^{2}(nh)^{-1+(1/\alpha_k)}+h^2\right\} \tag{4.2}$$

uniformly in $h\in\mathscr{H}_n$.

A related result for the "leave-one-out" estimate of loglikehood, in the case of kernel estimators, was obtained in [11]. The term $n^{-1}\sum\log f(X_j)$ on the right-hand side of (4.2) does not depend on h, and so has no bearing on the operation of maximize $\hat{L}$.

Unfortunately, the sum of those non-negligible terms not depending on h on the right-hand side of $\hat{L}$, is a decreasing function of h. Therefore it is maximized by taking h to be the smallest value in $\mathscr{H}_n$. In fact, the over-all maximum of $\hat{L}$, $\hat{L}=+\infty$, is achieved at $h=0$. This means that maximizing $\hat{L}$ is not a practical procedure for selecting h. A similar difficulty was noted by Habbema et al. [9] in the context of kernel density estimation. Their solution, shown by Titterington [23, 24] to be cross-validatory, was to omit observation X_j when computing $\hat{f}$ for argument X_j. However, the corresponding version of $\hat{L}$ in the histogram case is not well-defined if one or more bins contain a single element. The probability that the latter event occurs converges to one if f satisfies (4.1), and so cross-validation cannot be used to remove the difficulties noted just above.

AIC attempts to remove the difficulties by maximizing not $\hat{L}$ but $\hat{L}-n^{-1}J$, where J denotes "number of unknown parameters". Two possible interpretations of J are J_1, the number of nonempty bins, and J_2, the total number of bins between the smallest- and largest-index nonempty bins. Asymptotic theory for J_2 is relatively easy, so we treat it first.

Let $\alpha\equiv\min(\alpha_1,\alpha_2)$. Recall from elementary extreme-value theory (e.g. Galambos [6, pp. 51 and 108] that $n^{-1/(\alpha-1)}$ times the range between largest and smallest sample values, converges weakly to a proper, nondegenerate limit. Let Z have the distribution of this limit. Then $n^{-1/(\alpha-1)}hJ_2\to Z$ in distribution. Since $n^{-1+1/(\alpha-1)}h^{-1}$ is of larger order than $(nh)^{-1+(1/\alpha)}$ then by Theorem 4.1,

$$A_2(h)\equiv-\hat{L}+n^{-1}J_2=n^{(2-\alpha)/(\alpha-1)}h^{-1}Z_n+h^2b+o_p(n^{(2-\alpha)/(\alpha-1)}h^{-1}+h^2)$$

uniformly in $h\in\mathscr{H}_n$, where Z_n does not depend on h and converges weakly to Z. Noting that $(2-\alpha)/(\alpha-1)\geq 0$ when $\alpha\leq 2$, we see that minimization of Akaike's information criterion $A_2(h)$ produces bin widths whose order exceeds $n^{-\delta}$ for any $\delta>0$ when $\min(\alpha_1,\alpha_2)\leq 2$. On the other hand, when $\min(\alpha_1,\alpha_2)>2$, which corresponds to the existence of finite mean, the value of h which minimizes $A_2(h)$ satisfies

$$n^{(\alpha-2)/3(\alpha-1)}h\to(Z/2b)^{1/3}$$

in distribution. Since $0<(\alpha-2)/3(\alpha-1)<1$ then this bin width gives "acceptable" performance according to the definition in Sect. 1.

Next we treat the option J_1 for J, for which we need the following theorem.

Theorem 4.2. *Assume condition* (4.1). *Then*

$$n^{-1}J_1=\sum_{k=1}^{2} c_k^{1/\alpha_k}\Gamma(1-\alpha_k^{-1})(nh)^{-1+(1/\alpha_k)}+o_p\left\{\sum_{k=1}^{2}(nh)^{-1+(1/\alpha_k)}\right\}$$

uniformly in $h\in\mathcal{H}_n$.

Combining Theorems 4.1 and 4.2, and remembering that $D_k=c_k^{1/\alpha_k}D(\alpha_k)$, we see that Akaike's information criterion has the form

$$A_1(h)\equiv-\hat{L}+n^{-1}J_1=\sum_{k=1}^{2} c_k^{1/\alpha_k}\{\Gamma(1-\alpha_k^{-1})-D(\alpha_k)\}(nh)^{-1+(1/\alpha_k)}+h^2b$$

$$+o_p\left\{\sum_{k=1}^{2}(nh)^{-1+(1/\alpha_k)}+h^2\right\}.$$

If we are to avoid having the minimum of $A_1(h)$ achieved at an unacceptably small h value, then it is essential that $\Gamma(1-\alpha^{-1})>D(\alpha)$, where $\alpha\equiv\min(\alpha_1,\alpha_2)$. Again, this is only true for large values of α, not for α close to 1. This is clear from the fact that $\Gamma(1-\alpha^{-1})\to1$ and $D(\alpha)\to\frac{1}{2}$ as $\alpha\to\infty$, whereas $\Gamma(1-\alpha^{-1})\sim(\alpha-1)^{-1}$ and $D(\alpha)\sim(\alpha-1)^{-2}$ as $\alpha\downarrow1$. If $\min(\alpha_1,\alpha_2)$ is close to one, minimization of Akaike's information criterion $A_1(h)$ will encounter precisely those problems which arise when we minimize $-\hat{L}$: we get a value h which converges to zero faster than n^{-c} for each $c>0$. On the other hand, if $\alpha=\min(\alpha_1,\alpha_2)$ is sufficiently large then minimization of $A_1(h)$ is tantamount to minimizing $C_1(nh)^{-1+(1/\alpha)}+h^2b$, where $C_1>0$, and produces $h\sim C_2n^{-(\alpha-1)/(3\alpha-1)}$. Since $0<(\alpha-1)/(3\alpha-1)<1$ then this bin width gives "adequate" performance. Numerical computation indicates that $\Gamma(1-\alpha^{-1})-D(\alpha)$ is a monotone function of α, equalling zero when $\alpha=\alpha_0\approx2.49$.

Comparing the sizes of bin widths selected by both criteria we see that, since

$$(\alpha-2)/3(\alpha-1)<(\alpha-1)/(3\alpha-1)$$

whenever $\alpha>1$, the "number of nonempty bins" criterion will give a bin width of smaller order of magnitude. However, since $(\alpha-1)/(3\alpha-1)<1/3$ for all $\alpha>1$ then both criteria produce bin widths of a larger order of magnitude than the $n^{-1/3}$ which is optimal in the sense of minimizing L^2 loss; see [5, 19] for accounts of the latter. The fact that *AIC* produces larger bin widths than usual is in line with experience in certain other problems, where *AIC* tends to underestimate the number of unknown parameters; see e.g. [14].

5. Proofs

All proofs are given in abbreviated form, with only the main points described in detail.

Proof of Theorem 2.1. Put $I=I(\eta)\equiv\int_{(0,\eta)}f\log(f/E\hat{f})$, where $0<\eta<1$. It suffices to show that (i) $I\sim b(c_1,\alpha_1)m^{-(\alpha_1+1)}$ if $0<\alpha_1<1$, (ii) $I\sim(c_1/24)m^{-2}\log m$ if $\alpha_1=1$, and (iii)

$$I \sim (1/24)m^{-2} \int_{(0,\eta)} (f')^2 f^{-1} \quad \text{if } \alpha_1 = 0 \quad \text{or} \quad \alpha_1 > 1 .$$

Result (iii) follows easily by Taylor expansion, on writing $\log(f/E\hat{f}) = (f - E\hat{f})(E\hat{f})^{-1} - \frac{1}{2}\{(f - E\hat{f})^{-1}\}^2 +$remainder. The proof of (ii) begins similarly, obtaining $I = \sum_{i \leqq \varepsilon m} d_i + O(m^{-2})$ for any $0 < \varepsilon < \eta$, where

$$d_i \equiv \tfrac{1}{2} \int_{\mathscr{B}_i} (E\hat{f} - f)^2 (E\hat{f})^{-1} .$$

If $\alpha_1 = 1$, $d_i = (1 + r_i)m^{-2}i^{-1}c_1/24$ where $\lim_{\varepsilon \to 0, \nu \to \infty} \limsup_{m \to \infty} \max_{\nu < i \leqq \varepsilon m} |r_i| = 0$. This gives (ii). To prove (i), choose $i_0 \to \infty$ such that $i_0/m \to 0$, notice that

$$p_i = c_1(1 + r_i) \int_{(i-1)/m}^{i/m} x^{\alpha_1} dx \quad \text{and} \quad f(x) = \{1 + r(x)\} c_1 x^{\alpha_1}$$

where $\Delta \equiv \max_{i \leqq i_0} |r_i| + \sup_{x \leqq i_0/m} |r(x)| \to 0$, choose $i_1 \to \infty$ such that $i_1 \leqq i_0$ and $i_1^{\alpha_1 + 1}\Delta \to 0$, and note that for such i_1, $I(\eta) - I(i_1/m) = o(m^{-(\alpha_1 + 1)})$ and

$$I(i_1/m) = c_1 m^{-(\alpha_1 + 1)} \sum_{i=1}^{i_1} \int_{i-1}^{i} x^{\alpha_1} \log((\alpha_1 + 1)(x/i)^{\alpha_i} [i\{1 - (1 - i^{-1})^{\alpha_1 + 1}\}]^{-1}) dx$$

$$+ o(m^{-(\alpha_1 + 1)}) ,$$

where the integral equals $a_i(\alpha_1)$. □

Proof of Theorem 2.3. Define $\hat{f}$ and $\hat{f}_j$ as at (2.1) and (2.2), respectively. Put $\mu(x) \equiv E\{\hat{f}(x)\} = E\{\hat{f}_j(x)\}$. Then

$$-nCV - n\log(n-1) - \sum \log f(X_j) = \sum \log\{\hat{f}_j(X_j)/\mu(X_j)\}$$

$$+ \sum \log\{\mu(X_j)/f(X_j)\} . \qquad (5.1)$$

The second series is a sum of independent variables with means $-B$. It is readily shown to equal $-nB + o_p(nB + m)$. To treat the first series, fix $\eta > 0$ very small; divide bins into two groups such that $\mathscr{B}_i$ is in group 1 if $np_i > n^\eta$ and in group 2 otherwise; let $\mathscr{A}$, $\tilde{\mathscr{A}}$ be unions of group 1, group 2 bins respectively; and write

$$\sum_{j=1}^{n} \log\{\hat{f}_j(X_j)/\mu(X_j)\} = \left(\sum_{X_j \in \mathscr{A}} + \sum_{X_j \in \tilde{\mathscr{A}}} \right) \log\{\hat{f}_j(X_j)/\mu(X_j)\} .$$

The second series is dominated by $C(\log n)(\# X_j \in \tilde{\mathscr{A}})$, which equals $o_p(m)$ if η is sufficiently small. The first series may be treated by Taylor expansion,

$$\log\{\hat{f}_j(X_j)/\mu(X_j)\} = \{\hat{f}_j(X_j) - \mu(X_j)\}\mu(X_j)^{-1}$$

$$- \tfrac{1}{2}[\{\hat{f}_j(X_j) - \mu(X_j)\}\mu(X_j)^{-1}]^2 + \cdots ,$$

noting that

$$\sum_{X_j\in\mathscr{A}} \{\hat{f}_j(X_j)-\mu(X_j)\}^2\mu(X_j)^{-2}=m+o_p(m),$$

$$\sum_{X_j\in\mathscr{A}} \{\hat{f}_j(X_j)-\mu(X_j)\}\mu(X_j)^{-1}=\sum_{j=1}^{n} \{\hat{f}_j(X_j)-\mu(X_j)\}\mu(X_j)^{-1}+o_p(m)$$
$$=o_p(m). \tag{5.2}$$

Uniformity in m may be achieved via the continuity argument, much as in Stone [21]. □

Proofs of Theorems 2.1 and 3.1 are similar. For example, to derive Theorem 3.1 note that in place of (5.1) we have

$$-nAIC-n\log n-\sum\log f(X_j)=\sum\log\{\hat{f}(X_j)/\mu(X_j)\}+\sum\log\{\mu(X_j)/f(X_j)\}-J.$$

Now argue as before. The only essentially different feature is that, observing that $J=m+o_p(m)$, $\sum\mu(X_j)^{-1}=n+o_p(n)$, $\hat{f}(X_j)=(1-n^{-1})\hat{f}_j(X_j)+mn^{-1}$ and

$$\sum\{\hat{f}(X_j)-\mu(X_j)\}\mu(X_j)^{-1}=(1-n^{-1})\sum\{\hat{f}_j(X_j)-\mu(X_j)\}\mu(X_j)^{-1}$$
$$+mn^{-1}\sum\mu(X_j)^{-1}-1,$$

we obtained instead of (5.2):

$$\sum_{X_j\in\mathscr{A}} \{\hat{f}(X_j)-\mu(X_j)\}\mu(X_j)^{-1}-J=o_p(m).\quad \square$$

Proof of Theorem 4.1. Put $\mu(x)\equiv E\hat{f}(x)$, and observe that

$$n^{-1}\sum_{j=1}^{n}\log\{\hat{f}(X_j)/f(X_j)\}=S_1+S_2, \tag{5.3}$$

where $S_1\equiv n^{-1}\sum_j\log\{\hat{f}(X_j)/\mu(X_j)\}$ and $S_2\equiv n^{-1}\sum_j\log\{\mu(X_j)/f(X_j)\}$ represent "variance" and "bias" terms respectively. We treat S_1 and S_2 separately.

Step (i): S_1. Let $0<r_k<1<s_k<\infty$, and decompose S_1 into six parts:

$$S_1=\sum_{k=1}^{2}\sum_{l=1}^{3}T_{kl},\qquad \text{where } T_{kl}\equiv n^{-1}\sum_{X_j\in\mathscr{A}_{kl}}\log\{\hat{f}(X_j)/\mu(X_j)\},$$

$A_{11}\equiv(0,r_1(nh)^{1/\alpha_1}]$, $A_{12}\equiv(r_1(nh)^{1/\alpha_1},s_1(nh)^{1/\alpha_1}]$, $A_{13}\equiv(s_1(nh)^{1/\alpha_1},\infty)$, $A_{21}\equiv(-r_2(nh)^{1/\alpha_2},0]$, $A_{22}\equiv(-s_2(nh)^{1/\alpha_2},-r_2(nh)^{1/\alpha_2}]$, $A_{23}\equiv(-\infty,-s_2(nh)^{1/\alpha_2}]$. We need $r_kh^{-1}(nh)^{1/\alpha_k}$ and $s_kh^{-1}(nh)^{1/\alpha_k}$ to be integers, so we shall take r_k,s_k to be $[rh^{-1}(nh)^{1/\alpha_k}]h(nh)^{-1/\alpha_k}$, $[sh^{-1}(nh)^{1/\alpha_k}]h(nh)^{-1/\alpha_k}$ respectively, where $0<r<1<s<\infty$ are fixed and $[\cdot]$ denotes the integer part function. Let $\sum_i^{(k,l)}$ denote summation over values i such that $\mathscr{B}_i\subseteq\mathscr{A}_{kl}$, and put $p_i\equiv\int_{\mathscr{B}_i}f$. Then

$$T_{kl}=n^{-1}\sum_i^{(k,l)}N_i\log(N_i/np_i)=n^{-1}\sum_i^{(k,l)}N_i(N_i-np_i)(np_i)^{-1}+R_{kl} \tag{5.4}$$

where, since $|\log(1+u)-u|\leqq u^2-\log(1+u)I(u<-\frac{1}{2})$,

$$|R_{kl}|\leqq n^{-1}\sum_i^{(k,l)}N_i[\{(N_i-np_i)(np_i)^{-1}\}^2-\log(N_i/np_i)I(N_i/np_i<\tfrac{1}{2})]$$

$$\leqq Cn^{-1}\sum_i^{(k,l)}[N_i\{(N_i-np_i)(np_i)^{-1}\}^2+(N_i-np_i)^2(np_i)^{-1}]\equiv CS_{k,l}$$

say.

Step (i.a): T_{11} and T_{21}. We begin by bounding S_{k1}. Observe that

$$E(S_{k1})=n^{-1}\sum_i^{(k,l)}\{(np_i-2p_i+1)(1-p_i)(np_i)^{-1}+(1-p_i)\}\leqq Cr(nh)^{-1+(1/\alpha_k)},$$

whence

$$\lim_{r\to 0}\limsup_{n\to\infty}\sup_{n^{-1+\delta}\leqq h\leqq n^{-\delta}}(nh)^{1-(1/\alpha_k)}E\{S_{k1}(h)\}=0. \qquad (5.5)$$

Suppose we show that for each $\lambda>0$, and some $\varepsilon=\varepsilon(\delta,\lambda)>0$,

$$\sup_{n^{-1+\delta}\leqq h\leqq n^{-\delta}}P\{|S_{k1}(h)-ES_{k1}(h)|>(nh)^{-1+(1/\alpha_k)}n^{-\varepsilon}\}=O(n^{-\lambda}). \qquad (5.6)$$

Since $\mathscr{H}_n$ contains only $O(n^c)$ elements then if we choose $\lambda>c$,

$$P\{|S_{k1}(h)-ES_{k1}(h)|>(nh)^{-1+(1/\alpha_k)}n^{-\varepsilon},\text{ some }h\in\mathscr{H}_n\}\to 0.$$

It then follows via (5.5) that for each $\varepsilon>0$,

$$\lim_{r\to 0}\limsup_{n\to\infty}P\left\{\max_{h\in\mathscr{H}_n}(nh)^{1-(1/\alpha_k)}|R_{k1}(h)|>\varepsilon\right\}=0. \qquad (5.7)$$

To prove (5.6), write $N_i-np_i=\sum_j U_{ij}$ where $U_{ij}\equiv I(X_j\in\mathscr{B}_i)-p_i$. Then $U_{i1},\ldots,U_{in}$ are independent random variables with zero means. It may be proved after some tedious algebra, as in Marron [16, pp. 1020–1022], that with

$$Z_1\equiv n^{-1}\sum_i^{(k,l)}\{(N_i-np_i)^2(np_i)^{-1}-E(N_i-np_i)^2(np_i)^{-1}\}$$

$$=n^{-1}\sum_{j=1}^{n}\sum_{l=1}^{n}\sum_i^{(k,l)}\{U_{ij}U_{il}-E(U_{ij}U_{il})\}(np_i)^{-1},$$

we have for sufficiently small $\varepsilon>0$,

$$E(Z_1^{2\nu})=O\left[n^{-2\nu}\sum_{j=0}^{2\nu}h^j\{h^{-1}(nh)^{1/\alpha_k}\}^{\nu+(j/2)}\right]$$

$$=O[\{(nh)^{-1+(1/\alpha_k)}\}^{2\nu}\{(n^{-1/\alpha_k}h^{1-(1/\alpha_k)})^\nu+h^{2\nu}\}]$$

$$=O[\{n^{-\varepsilon}(nh)^{-1+(1/\alpha_k)}\}^{2\nu}].$$

An identical bound holds for $E(Z_2^{2\nu})$, where

$$Z_2\equiv n^{-1}\sum_i^{(k,1)}\{(N_i-np_i)^3(np_i)^{-2}-E(N_i-np_i)^3(np_i)^{-2}\};$$

in each case, note that $np_i \geqq \varepsilon$, for some $\varepsilon > 0$, uniformly in $\mathscr{B}_i \subseteq \mathscr{A}_{k1}$. From these results and the fact that $S_{k1} - ES_{k1} = 2Z_1 + Z_2$ we see that for any $\lambda > 0$,

$$\sup_{n^{-1+\delta} \leqq h \leqq n^{-\delta}} E[n^\varepsilon (nh)^{1-(1/\alpha_k)} \{S_{k1}(h) - ES_{k1}(h)\}]^{2\nu} = O(n^{-\lambda}), \tag{5.8}$$

provided ε is sufficiently small and ν sufficiently large. Formula (5.6) follows via Markov's inequality.

Finally we treat the term

$$U_k(h) \equiv n^{-1} \sum_i^{(k,1)} N_i(N_i - np_i)(np_i)^{-1}$$

appearing in (5.4). Follow exactly the argument above, obtaining the following analogue of (5.8):

$$\sup_{n^{-1+\delta} \leqq h \leqq n^{-\delta}} E[n^\varepsilon (nh)^{1-(1/\alpha_k)} \{U_k(h) - EU_k(h)\}]^{2\nu} = O(n^{-\lambda}),$$

for sufficiently small ε and large ν. It follows that

$$\sup_{n^{-1+\delta} \leqq h \leqq n^{-\delta}} P\{|U_k(h) - EU_k(h)| > (nh)^{-1+(1/\alpha_k)} n^{-\varepsilon}\} = O(n^{-\lambda}),$$

whence $\max_{h \in \mathscr{H}_n} (nh)^{1-(1/\alpha_k)} |U_k(h) - EU_k(h)| \to 0$ in probability. Furthermore,

$$0 \leqq EU_k(h) = n^{-1} \sum_i^{(k,1)} (1 - p_i) \leqq Cr(nh)^{-1+(1/\alpha_k)},$$

and so for each $\varepsilon > 0$,

$$\lim_{r \to 0} \limsup_{n \to \infty} P\left\{\max_{h \in \mathscr{H}_n} (nh)^{1-(1/\alpha_k)} |U_k(h)| > \varepsilon\right\} = 0.$$

From this result, (5.4) and (5.5) we deduce that for all $\varepsilon > 0$,

$$\lim_{r \to 0} \limsup_{n \to \infty} P\left\{\max_{h \in \mathscr{H}_n} (nh)^{1-(1/\alpha_k)} |T_{k1}(h)| > \varepsilon\right\} = 0. \tag{5.9}$$

Step (i.b): T_{12} and T_{22}. Our initial goal is to prove that

$$\sup_{h \in \mathscr{H}_n} (nh)^{1-(1/\alpha_k)} |T_{k2}(h) - ET_{k2}(h)| \to 0 \tag{5.10}$$

in probability. For this, it suffices to show that for some $\eta > 0$, each $\delta \leqq a \leqq 1 - \delta$ and each $\varepsilon > 0$,

$$P\left\{\sup_{h \in \mathscr{H}_n : n^{-a-\eta} \leqq h \leqq n^{-a+\eta}} (nh)^{1-(1/\alpha_k)} |T_{k2}(h) - ET_{k2}(h)| > \varepsilon\right\} \to 0.$$

This in turn will follow if we prove that

$$\sum_{h \in \mathscr{H}_n : n^{-a-\eta} \leqq h \leqq n^{-a+\eta}} E\{(nh)^{1-(1/\alpha_k)} |T_{k2}(h) - ET_{k2}(h)|\}^2 \to 0. \tag{5.11}$$

If $i \neq j$ then, conditional on N_i, N_j is $Bi\{n-N_i, p_j(1-p_i)^{-1}\}$. From that observation and after some algebra it follows that

$$E\left[\sum_i^{(k,2)} \{N_i \log(N_i/np_i) - EN_i \log(N_i/np_i)\}\right]^2 = O\left(\sum_i^{(k,2)} 1\right)$$
$$= O\{h^{-1}(nh)^{1/\alpha_k}\},$$

whence

$$E[(nh)^{1-(1/\alpha_k)}\{T_{k2}(h) - ET_{k2}(h)\}]^2 = O\{(nh)^{1-(1/\alpha_k)} n^{-1}\}.$$

Since $\mathscr{H}_n$ contains $O(n^{a+\eta})$ values in the interval $[n^{-a-\eta}, n^{-a+\eta}]$ then the left-hand side of (5.11) is of the same order as

$$n^{a+\eta}(n^{1-a+\eta})^{1-(1/\alpha_k)} n^{-1} = n^{\gamma/\alpha_k},$$

where $\gamma \equiv a - 1 + \eta(2\alpha_k - 1) \leqq -\delta + \eta(2\alpha_k - 1) < 0$ if $\eta < \delta/(2\alpha_k - 1)$. Result (5.11) is immediate.

Finally, we compute $E(T_{k2})$:

$$E(T_{k2}) + n^{-1} \sum_{r_k(nh)^{1/\alpha_k} < (-1)^{k+1} hi < s_k(nh)^{1/\alpha_k}} E\{N_i \log(N_i/np_i)\}$$
$$= \{1 + o(1)\} n^{-1} \int_{r_k h^{-1}(nh)^{1/\alpha_k}}^{s_k h^{-1}(nh)^{1/\alpha_k}} E(N\{(nh^{1-\alpha_k})^{-1/\alpha_k} x\}$$
$$\times \log[N\{(nh^{1-\alpha_k})^{-1/\alpha_k} x\}/(nh^{1-\alpha_k} x^{-\alpha_k} c_k)]) dx,$$

where $N(x)$ has the $Bi(n, n^{-1}c_k x^{-\alpha_k})$ distribution. Let $W(x)$ be Poisson-distributed with mean $c_k x^{-\alpha_k}$. Changing variable in the above integral, and using the Poisson approximation to the Binomial, we see that $E(T_{k2}) = \{1 + o(1)\}(nh)^{-1+(1/\alpha_k)} A_k$, where

$$A_k \equiv \int_r^s E[W_k(x) \log\{W_k(x)/c_k x^{-\alpha_k}\}] dx.$$

From this result and (5.10) we conclude that

$$\max_{h \in \mathscr{H}_n} |(nh)^{1-(1/\alpha_k)} T_{k2}(h) - A_k| \to 0$$

in probability. The integral defining A_k converges absolutely, to D_k, as $r \to 0$ and $s \to \infty$. Therefore

$$\lim_{r \to 0, s \to \infty} \limsup_{n \to \infty} P\left\{\max_{h \in \mathscr{H}_n} |(nh)^{1-(1/\alpha_k)} T_{k2}(h) - D_k| > \varepsilon\right\} = 0. \qquad (5.12)$$

Step (i.c): T_{13} and T_{23}. An argument similar to that in Step (i.b) produces the following analogue of (5.10):

$$\sup_{h \in \mathscr{H}_n} (nh)^{1-(1/\alpha_k)} |T_{k3}(h) - ET_{k3}(h)| \to 0 \qquad (5.13)$$

in probability. To bound $E|T_{k3}|$, note that if N is $Bi(n, p)$ and $C_1 > 0$ then $E\{N|\log(N/np)|\} \leqq C_2 np |\log np|$ uniformly in $n \geqq 2$ and $0 < np < C_1$. Therefore by (5.4),

$$\begin{aligned} E|T_{k3}| &\leqq C_2 \sum_i{}^{(k,3)} p_i |\log np_i| \\ &\leqq C_3 \sum_{i > s_k h^{-1}(nh)^{1/\alpha_i}} h^{i-\alpha_k} i^{-\alpha_k} |\log(nh^{1-\alpha_k} i^{-\alpha_k})| \\ &\leqq C_4 (nh)^{-1+(1/\alpha_k)} \int_{s_k}^{\infty} x^{-\alpha_k} |\log x| \, dx . \end{aligned}$$

From this we deduce that

$$\lim_{s\to\infty} \limsup_{n\to\infty} \sup_{n^{-1+\delta} \leqq h \leqq n^{-\delta}} (nh)^{1-(1/\alpha_k)} E|T_3| = 0 ,$$

which together with (5.13) gives

$$\lim_{s\to\infty} \limsup_{n\to\infty} P\left\{ \max_{h \in \mathscr{H}_n} (nh)^{1-(1/\alpha_k)} |T_{k3}| > \varepsilon \right\} = 0 \tag{5.14}$$

for each $\varepsilon > 0$.

Combining (5.9), (5.12) and (5.14) we conclude that, uniformly in $h \in \mathscr{H}_n$,

$$S_1 = \sum_{k=1}^{2} \sum_{l=1}^{3} T_{kl} = \sum_{k=1}^{2} D_k (nh)^{-1+(1/\alpha_k)} + o_p\left\{ \sum_{k=1}^{2} (nh)^{-1+(1/\alpha_k)} \right\} . \tag{5.15}$$

Step (ii): S_2. Put $Y \equiv \log\{\mu(X_1)/f(X_1)\}$, $B \equiv -E(Y)$ and $Z \equiv Y - E(Y)$. Let $v \geqq 1$ be an integer. By Rosenthal's inequality [13, p. 23],

$$n^{2v} E(S_2 + B)^{2v} \leqq C[\{nE(Z^2)\}^v + nE(Z^{2v})] .$$

Since $|\mu(x) - f(x)| \leqq C_1 h x^{-\alpha_1 - 1}$ and $f(x) > C_2 x^{-\alpha_1}$ for $x > 1$, with analogous bounds for $x < -1$, then

$$\begin{aligned} E(Z^{2v}) &\leqq C_3 E(Y^{2v}) = C_3 E(\log[1 + \{\mu(X_1) - f(X_1)\} f(X_1)^{-1}])^{2v} \\ &\leqq C_4 \int \{\mu(x) - f(x)\}^{2v} f(x)^{-2v+1} dx \leqq C_5 h^{2v} . \end{aligned}$$

Consequently, $E(S_2 + B)^{2v} \leqq C_6 (n^{-1} h^2)^v$. For any $\alpha > 1$, $n^{-1} h^2 \leqq \frac{1}{2}\{(nh)^{-1+(1/\alpha)} + h^2\}^2 n^{-1/\alpha} h^{1-(1/\alpha)}$. Therefore if $\lambda > 0$ and v is chosen sufficiently large,

$$\sup_{n^{-1+\delta} \leqq h \leqq n^{-\delta}} E|\{(nh)^{-1+(1/\alpha_1)} + (nh)^{-1+(1/\alpha_2)} + h^2\}^{-1} (S_2 + B)|^{2v} = O(n^{-\lambda}) ,$$

from which it follows that

$$\max_{h \in \mathscr{H}_n} \{(nh)^{-1+(1/\alpha_1)} + (nh)^{-1+(1/\alpha_2)} + h^2\}^{-1} |S_2 + B| \to 0 \tag{5.16}$$

in probability. Elementary analysis shows that

$$B = -\int f \log\{1 + (\mu - f) f^{-1}\} \sim \tfrac{1}{2} \int (\mu - f)^2 f^{-1} \sim \tfrac{1}{24} h^2 \int (f')^2 f^{-1} .$$

This result and (5.16) give

$$S_2 = -\tfrac{1}{24} h^2 \int (f')^2 f^{-1} + o_p\{(nh)^{-1+(1/\alpha_1)} + (nh)^{-1+(1/\alpha_2)} + h^2\} \tag{5.17}$$

uniformly in $h \in \mathscr{H}_n$. Theorem 4.1 follows from (5.15) and (5.17). □

Proof of Theorem 4.2. The number of non-empty bins equals

$$J_1=\sum_i I(N_i\geqq 1)=\sum_{k=1}^{2}\sum_{l=1}^{3}U_{kl},\qquad \text{where } U_{kl}\equiv\sum_i{}^{(k,l)} I(N_i\geqq 1)$$

and $\sum_i{}^{(k,l)}$ is defined as in the proof of Theorem 4.1. Analogues of arguments in Steps (i.a)–(i.c) of that proof show that for small r and large s, U_{k1} and U_{k3} are negligible, and that

$$\max_{h\in\mathscr{H}_n}(nh)^{1-(1/\alpha_k)}|U_{k2}-EU_{k2}|\to 0$$

in probability. Now,

$$E(U_{k2})=\sum_i{}^{(k,2)}\{1-(1-p_i)^n\}=\{1+o(1)\}\sum_i{}^{(k,2)}\{1-\exp(-nh^{1-\alpha_k}c_k i^{-\alpha_k})\}$$
$$=\{1+o(1)\}n(nh)^{-1+(1/\alpha_k)}\int_r^s\{1-\exp(-c_k x^{-\alpha_k})\}\,dx.$$

Letting $r\to 0$ and $s\to\infty$ in the integral, integrating by parts and changing variable, the integral becomes $c_k^{1/\alpha_k}\Gamma\{1-(1/\alpha_k)\}$, whence the theorem. □

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Cźaki, P. (eds.) Second Internat. Symp. on Information Theory. pp. 267–281. Budapest: Akademiai Kiadó 1973
2. Bowman, A.W.: A comparative study of some kernel-based nonparametric density estimators. J. Stat. Comput. Simulation **21**, 313–327 (1985)
3. Chow, Y.S., Geman, S., Wu, L.-D.: Consistent cross-validated density estimation. Ann. Stat. **11**, 25–38 (1983)
4. Duin, R.P.W.: On the choice of smoothing parameters for Parzen estimators of probability density functions. IEEE Trans. Comput. C**25**, 1175–1179 (1976)
5. Freedman, D., Diaconis, P.: On the histogram as a density estimator: L^2 theory. Z. Wahrscheinlichkeitstheor. Verw. Geb. **57**, 453–476 (1981)
6. Galambos, J.:. The asymptotic theory of extreme order statistics. New York: Wiley 1978
7. Gregory, G.G., Schuster, E.F.: Contributions to non-parametric maximum likelihood methods of density estimation. In: Gentleman, J.F. (ed.) 12th Annual Symp. Interface Comput. Sci. Statist., pp. 427–431. Ontario: University of Waterloo, Canada 1979
8. Habbema, J.D.F., Hermans, J., Remme, J.: Variable kernel estimation in discriminant analysis. In: Corsten, L.C.A., Hermans, J. (eds.) Compstat 1978. pp. 178–185. Vienna: Physica 1978
9. Habbema, J.D.F., Hermans, J., Broek, K. van den: A stepwise discriminant analysis program using density estimation. In: Bruckman, G. (ed.) Compstat 1974. pp. 101–110. Vienna: Physica 1974
10. Habbema, J.D.F., Hermans, J., Broek, K. van den: Selection of variables in discriminant analysis by F-statistic and error rate. Technometrics **19**, 487–493 (1974)
11. Hall, P.: On Kullback-Leibler loss and density estimation. Ann. Stat. **15**, 1491–1519 (1987)
12. Hall, P.: On the estimation of probability densities using compactly supported kernels. J. Multivariate Anal. **23**, 131–158 (1987)
13. Hall, P., Heyde, C.C.: Martingale limit theory and its application. New York London: Academic Press 1980
14. Hannan, E.J., Rissanen, J.: The width of a spectral window. Adv. Appl. Probab. (1988)
15. Kullback, S.: Information theory and statistics. New York: Wiley 1959

16. Marron, J.S.: An asymptotically efficient solution to the bandwidth problem of kernel density estimation. Ann. Stat. **13**, 1011–1023 (1985)
17. Raatgever, J.W., Duin, R.P.W.: On the variable kernel method for multivariate nonparametric density estimation. In: Corsten, L.C.A., Hermans, J. (eds.) Compstat 1978. pp. 524–533. Vienna: Physica 1978
18. Rissanen, J.: Stochastic complexity. (With discussion.) J. R. Stat. Soc. **49**, 223–239 (1987)
19. Scott, D.W.: On optimal and data-based histograms. Biometrika **66**, 605–610 (1979)
20. Schuster, E.F., Gregory, G.G.: On the consistency of maximum likelihood nonparametric density estimators. In: Eddy, W.F. (ed.) 13th Annual Symp. Interface Comput. Sci. Statist., pp. 295–298. New York Berlin Heidelberg: Springer 1981
21. Stone, C.J.: An asymptotically optimal histogram selection rule. In: LeCam, L.M., Olshen, R.A. (eds.) Proc. Berkely Conf. in Honor of J. Neyman and J. Kiefer, vol. II, pp. 513–520. Belmont, Calif.: Wadsworth 1985
22. Taylor, C.C.: Akaike's information criterion and the histogram. Biometrika **74**, 636–639 (1987)
23. Titterington, D.M.: Contribution to discussion of paper by T. Leonard. J. R. Stat. Soc., Ser. B**40**, 139–140 (1978)
24. Titterington, D.M.: A comparative study of kernel-based density estimates for categorical data. Technometrics **22**, 259–268 (1980)