The Annals of Statistics
1987, Vol. 15, No. 4, 1491–1519

ON KULLBACK–LEIBLER LOSS AND DENSITY ESTIMATION

BY PETER HALL

Australian National University

"Discrimination information," or Kullback–Leibler loss, is an appropriate measure of distance in problems of discrimination. We examine it in the context of nonparametric kernel density estimation and show that its asymptotic properties are profoundly influenced by tail properties of the kernel and of the unknown density. We suggest ways of choosing the kernel so as to reduce loss, and describe the extent to which likelihood cross-validation asymptotically minimises loss. Likelihood cross-validation generally leads to selection of a window width of the correct order of magnitude, but not necessarily to a window with the correct first-order properties. However, if the kernel is chosen appropriately, then likelihood cross-validation does result in asymptotic minimisation of Kullback–Leibler loss.

1. Introduction and discussion. The purpose of this paper is to provide concise descriptions of discrimination information (Kullback–Leibler loss) for kernel probability density estimates and to explore the extent to which likelihood cross-validation leads to asymptotic minimisation of the information available for discriminating between true and estimated densities. The first section provides motivation, illustrative examples and a summary of results and conclusions.

1.1. *Kullback–Leibler loss and likelihood cross-validation.* Let f and g represent two probability density functions and X denote a single observation from the distribution with density f. The expected amount of information in X for discriminating against g is given by

$$L(f, g) \equiv \int f(x)\log\{f(x)/g(x)\}\,dx \tag{1.1}$$

[23, page 5] and is nonnegative. We call this Kullback–Leibler loss. If $g = \hat{f}$ is an estimate of f, then the expected Kullback–Leibler loss associated with this estimate is $E\{L(f, \hat{f})\}$.

The distance function L is not a metric and Kullback–Leibler loss is not an appropriate measure of distance in the goodness-of-fit sense. It is purpose-built for discrimination and in the context of density-estimation it may not have applications outside that context.

Likelihood cross-validation represents a data-driven attempt at constructing $\hat{f}$ so as to minimise $L(f, \hat{f})$. For example, if $X_1, \ldots, X_n$ is a random sample from f

Received January 1986; revised February 1987.
AMS 1980 *subject classifications*. Primary 62G99; secondary 62H99.
Key words and phrases. Density estimation, discrimination, kernel method, Kullback–Leibler loss, likelihood cross-validation.

and if $\hat{f}$ is the univariate kernel estimator

$$\hat{f}(x|h) \equiv (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\} \tag{1.2}$$

with window h and kernel K, then likelihood cross-validation recommends choosing h so as to maximise

$$CV(h) \equiv n^{-1} \sum_{i=1}^{n} \log \hat{f}_i(X_i|h),$$

where $\hat{f}_i(x|h) \equiv \{(n-1)h\}^{-1} \Sigma_{j \neq i} K\{(x - X_j)/h\}$. This procedure was proposed by Habbema, Hermans and van den Broek [13] and Duin [9] and shown by Titterington ([33], [34]) to be cross-validatory in the sense of Stone [29]–[31]. It forms part of a widely used program ALLOC for discrimination, and examples of its use may be found in [11], [12] and [25].

1.2. *Influence of tail behaviour on consistency.* One of the contributions of this paper is an explicit account of how Kullback–Leibler loss and likelihood cross-validation are influenced by interaction between tail properties of the kernel K and of the unknown density f. This interaction can result in infinite loss and inconsistent estimation as we show by example in the present subsection.

Suppose f is symmetric about the origin, bounded away from zero on compact intervals and satisfies $f(x) \sim c|x|^{-\alpha}$ as $|x| \to \infty$, where $c > 0$ and $\alpha > 1$. Examples include Cauchy and Student's distributions. Take the kernel to be

$$K(z) \equiv A_2 \exp(-A_1|z|^{\kappa}), \qquad -\infty < z < \infty, \tag{1.3}$$

where A_1, A_2 and κ are positive constants linked by the requirement that K integrate to unity. Examples include standard normal and double exponential kernels. Define the estimator $\hat{f}(\cdot|h)$ as in (1.2). We claim that (i) expected Kullback–Leibler loss is infinite if $\kappa \geq \alpha - 1$ and (ii) likelihood cross-validation selects a window $\hat{h}$ diverging to infinity and so leads to inconsistency if $\kappa > \alpha - 1$. These results follow from the inequalities

$$\begin{aligned} A_1 W_n(x) + \log\bigl(hA_2^{-1}\bigr) &\leq -\log \hat{f}(x|h) \\ &\leq A_1 W_n(x) + \log\bigl(nhA_2^{-1}\bigr), \\ n^{(1/\beta)-1} h^{-\kappa} A_1 Y_n + \log\bigl(hA_2^{-1}\bigr) &\leq -CV(h) \leq n^{(1/\beta)-1} h^{-\kappa} A_1 Z_n \\ &\quad + \log\bigl(hA_2^{-1}\bigr), \end{aligned} \tag{1.4}$$

respectively, where $W_n(x) \equiv \min_i |(x - X_i)/h|^{\kappa}$, $Y_n \equiv \{n^{-1/\beta\kappa}(X_{(2)} - X_{(1)})\}^{\kappa}$, $Z_n \equiv 2^{\kappa+1} n^{-1/\beta} \Sigma_i |X_i|^{\kappa}$, $\beta \equiv (\alpha - 1)/\kappa < 1$ and $X_{(1)} \leq \cdots \leq X_{(n)}$ are the order statistics of $X_1, \ldots, X_n$. Note that $EW_n(x) < \infty$ if and only if $\kappa < \alpha - 1$, $Y_n^{1/\kappa}$ converges weakly to a positive continuous limit which is the difference between extreme and penultimate extreme value limits, and Z_n has a positive stable-law weak limit with exponent β. Incidentally, result (1.4) plays an important role in several of our proofs.

1.3. *Influence of tail behaviour on loss.* The influence of tail properties of K and f on Kullback–Leibler loss and likelihood cross-validation is both complex and profound. In this subsection we explain the main features.

Define $\hat{f}(\cdot|h)$ and $L(f, \hat{f})$ as in (1.1) and (1.2), and put $l_n(h) \equiv E\{L(f, \hat{f})\}$. Then $l_n(h) = V + B$, where

$$V \equiv \int_{-\infty}^{\infty} f(x) E\left[\log\{E\hat{f}(x|h)/\hat{f}(x|h)\}\right] dx,$$

$$B \equiv \int_{-\infty}^{\infty} f(x) \log\{f(x)/E\hat{f}(x|h)\}\, dx$$

are variance and bias components, respectively, and are nonnegative. Analogues of V and B in the theory of squared-error loss are typically of orders $(nh)^{-1}$ and h^4, respectively, for a wide variety of positive K's and twice-differentiable f's. But in the case of Kullback–Leibler loss, orders of magnitude depend crucially on tail properties of K and f.

Usually, V can be decomposed into three parts: a "main-effect" part arising from "most" of the distribution and two other parts deriving from the extreme upper tail and extreme lower tail of the distribution, respectively. The first part can be significantly affected by tail properties of the underlying distribution, but not to the same extent as the other two. Only the two "tail-effect" terms in V are strongly influenced by properties of the tails of the kernel. The bias component B is generally simpler, having either one or two significant parts.

For example, consider a density f whose support equals $(0, a)$, where $0 < a < \infty$, and which is continuous and nonzero on $(0, a)$ and such that $f(x) \sim c_1 x^{\alpha_1}$ and $f(a - x) \sim c_2 x^{\alpha_2}$ as $x \downarrow 0$, where $c_1, c_2 > 0$ and $\alpha_1, \alpha_2 \geq 0$. Suppose K is given by (1.3) and h is chosen so that $h \to 0$ and $nh \to \infty$ as $n \to \infty$. Then the main-effect term in V is of order $(nh)^{-1}$, while the two tail-effect terms are of orders $n^{-1-\kappa/(\alpha_1+1)}h^{-\kappa}$ and $n^{-1-\kappa/(\alpha_2+1)}h^{-\kappa}$.

Overall loss $V + B$ can contain up to five terms, any one of which may dominate all the others. The aim is to select h so as to minimise the total of these contributions.

1.4. *Influence of tail behaviour on likelihood cross-validation.* The extent to which likelihood cross-validation minimises expected Kullback–Leibler loss is intimately bound up with the tail-effect terms in V. Those terms are due essentially to a small number of extreme order statistics from the sample. Since that number is so small, no "law of large numbers" applies. This means that if the tail-effect terms play a role in determining minimum Kullback–Leibler loss, then likelihood cross-validation *does not* lead to asymptotic minimisation of Kullback–Leibler loss. However, if those terms are not dominant (which can be achieved by correct choice of kernel, as we point out in the next subsection), then likelihood cross-validation *does* asymptotically minimise expected Kullback–Leibler loss. These comments apply no matter whether performance is measured in terms of expected Kullback–Leibler loss $E\{L(f, \hat{f})\}$ or in terms of "raw" Kullback–Leibler loss $L(f, \hat{f})$. Furthermore, if the tail-effect terms are dominant, then minimisation of expected loss is not asymptotically equivalent to minimisation of raw loss.

1.5. *Conclusions.* A major conclusion to be drawn from our work is the extraordinary influence which tail properties of f and K have on Kullback–Leibler loss and likelihood cross-validation. The theory of squared-error loss has no parallel for this phenomenon. From a practical point of view, it is important that K be chosen so that its tails are sufficiently thick for tail-effect terms in V to be negligible. As a general rule, "the thicker the tails of the underlying density, the thicker the tails required of the kernel." Tails of the standard normal kernel are too thin for most purposes and even the double exponential kernel is not always suitable. A practical alternative is the kernel $K_0(z) \equiv \text{const.} \exp[-\frac{1}{2}\{\log(1 + |z|)\}^2]$, $-\infty < z < \infty$, whose tails decrease more slowly than $\exp(-|z|^\kappa)$ for any $\kappa > 0$. This kernel renders negligible the tail-effect terms in all cases treated in this paper. Examples illustrating these points and also the kernel K_0 will be discussed in Sections 2.3 and 3.1.

Another conclusion is that the window h which minimises Kullback–Leibler loss can be quite inappropriate from a goodness-of-fit point of view. For example, consider estimating the Cauchy density $f(x) \equiv \pi^{-1}(1 + x^2)^{-1}$ and suppose we use the kernel K_0. Then expected Kullback–Leibler loss is

$$l_n(h) = C_1(nh)^{-1/2} + C_2h^4 + o\{(nh)^{-1/2} + h^4\}$$

as $n \to \infty$, $h \to 0$ and $nh \to \infty$, where C_1 and C_2 are positive constants. (See Sections 3.2 and 3.3.) This formula is minimised by taking $h \sim \text{const}\, n^{-1/9}$. By way of comparison, the window h which minimises L^2 distance between $\hat{f}$ and f is of order $n^{-1/5}$ and this is much smaller than $n^{-1/9}$. Therefore, minimising Kullback–Leibler loss leads us to smooth considerably more than is "optimal" in goodness-of-fit terms.

The basic conditions $n \to \infty$, $h \to 0$ and $nh \to \infty$, which are necessary and sufficient for pointwise and L^2 consistency of $\hat{f}$, are not sufficient for expected Kullback–Leibler loss to converge to zero. An example illustrating this point will be given in Remark 2.4.

For an appropriate choice of kernel (e.g., for K_0), likelihood cross-validation does deliver a window which asymptotically minimises Kullback–Leibler loss. In particular, $h_0 \sim \text{const}\, n^{-1/9}$ is "optimal" in a Kullback–Leibler sense for the Cauchy distribution, and the window $\hat{h}_0$ chosen by likelihood cross-validation satisfies $\hat{h}_0/h_0 \to 1$ in probability. (See Section 3.4.)

1.6. *Related work.* Numerical examples concerning likelihood cross-validation have been described in [3], [4], [10] and [27]. Inconsistency, usually with compactly supported kernels, has been reported in [10] and [27]. On the other hand, Chow, Geman and Wu [7] have shown that likelihood cross-validation produces consistent estimates when f and K are both compactly supported. We argue elsewhere [18] that despite consistency, the size of loss is unduly large in this circumstance. The present paper appears to be the first to give a general account of Kullback–Leibler loss and likelihood cross-validation which goes beyond the issue of consistency. We are not specifically concerned with con-

sistency, although sufficient conditions for consistency follow directly from our results.

Various versions of squared-error cross-validation have been treated in [3], [4], [16], [17], [19], [20], [24], [26] and [28]. Kullback–Leibler loss and likelihood cross-validation in the discrete case have been considered by several authors ([2], [5], [14], [32]), but no important parallels appear to exist with the continuous case which is the subject of the present paper.

Section 2 will state and prove our main theorems for densities with compact support, and Section 3 will present similar results for densities having regularly varying tails.

2. Densities with compact support. As we indicated in Section 1, the large-sample properties of Kullback–Leibler loss and likelihood cross-validation are rather complex. In this section we shall attempt to give a relatively detailed account in the case of compactly supported densities, so as to clearly demonstrate the impact of the "tail-effect" terms mentioned in Section 1.3.

Recall that expected Kullback–Leibler loss $l_n(h) \equiv E\{L(f, \hat{f})\}$ may be decomposed as $l_n(h) = V + B$, where V and B are the variance and bias components, respectively, and are defined in (1.5). It is convenient to treat these terms separately. That we shall do in Sections 2.1 and 2.2, combining them in Section 2.3. Likelihood cross-validation will be analysed in Section 2.4. Section 2.5 will examine raw Kullback–Leibler loss and proofs will be given in Section 2.6.

Throughout the remainder of this paper the symbols C, C_0, C_1, $C_2, \ldots$ will denote generic positive constants, different at different appearances.

2.1. *Variance component V.* Assume the following conditions on f:

(2.1) f is bounded away from zero and infinity on $(\varepsilon, a - \varepsilon)$ for each $\varepsilon > 0$, continuous almost everywhere, vanishes outside $[0, a]$ and satisfies $f(x) \sim c_1 x^{\alpha_1}$ and $f(a - x) \sim c_2 x^{\alpha_2}$ as $x \downarrow 0$, where $c_1, c_2 > 0$ and $\alpha_1, \alpha_2 \geq 0$.

Suppose the kernel K is bounded, integrates to unity and satisfies either

$$K(z) \equiv A_2 \exp(-A_1|z|^\kappa), \qquad -\infty < z < \infty, \tag{2.2}$$

or

$$K(z) \geq A_2 \exp(-A_1|z|), \qquad -\infty < z < \infty, \tag{2.3}$$

for positive constants A_1, A_2 and κ. In the case of (2.2) we may as well assume $\kappa > 1$, for otherwise the condition is subsumed by (2.3). By convention *we shall take* $\kappa = 1$ *if* K *satisfies* (2.3).

Next we define the coefficients of the tail-effect terms in an expansion of V. Given $c > 0$ and $\alpha \geq 0$, let $g = g(x, y)$ denote the solution of the equation

$$c \int_{\max(0,\, y-g)}^{y+g} u^\alpha \, du = x \tag{2.4}$$

for $x, y > 0$. Given A_1, c, α and κ, define

$$D = D(A_1, c, \alpha, \kappa) \equiv A_1 c \int_0^\infty \int_0^\infty g(x, y)^\kappa e^{-x} y^\alpha \, dx \, dy \tag{2.5}$$

if $\kappa > 1 + \alpha^{-1}$ and $D = 0$ if $\kappa \leq 1 + \alpha^{-1}$. The left-hand side of (2.4) is a continuous and strictly increasing function of g, increasing from zero to infinity as g increases from zero to infinity, and so $g(x, y)$ is well-defined. It may be shown that

$$g(x, y) \leq C\{x^{1/(\alpha+1)} \wedge (xy^{-\alpha})\}.$$

Substituting this estimate into (2.5) we deduce that D is finite and positive for $\kappa > 1 + \alpha^{-1}$.

If (2.1) and either (2.2) or (2.3) hold, set $D_i \equiv D(A_1, c_i, \alpha_i, \kappa)$. Our first theorem describes asymptotic properties of V.

THEOREM 2.1. *Assume* (2.1) *and either* (2.2) *or* (2.3) *and that* $h = h(n) \to 0$, $nh \to \infty$ *and* $(nh)^{-1}(\log n)^{\alpha_i} \to 0$ *in the special case* $\kappa = 1 + \alpha_i^{-1}$ ($i = 1$ *or* 2). *Then as* $n \to \infty$,

$$V = (nh)^{-1} \tfrac{1}{2} a \int_{-\infty}^\infty K^2(z) \, dz + \sum_{i=1}^2 D_i n^{-1-\kappa/(\alpha_i+1)} h^{-\kappa} + o\left\{(nh)^{-1} + \sum_{i=1}^2 n^{-1-\kappa/(\alpha_i+1)} h^{-\kappa}\right\}. \tag{2.6}$$

REMARK 2.1. To determine when the tail-effect terms may be ignored, form the ratio

$$n^{-1-\kappa/(\alpha_i+1)} h^{-\kappa} / (nh)^{-1} = \left(nh^{(\alpha_i+1)(1-\kappa^{-1})}\right)^{-\kappa/(\alpha_i+1)}.$$

Therefore, if $(\alpha_i + 1)(1 - \kappa^{-1}) \leq 1$ or, equivalently, if $\kappa \leq 1 + \alpha_i^{-1}$, then the condition $nh \to \infty$ dictates that $n^{-1-\kappa/(\alpha_i+1)} h^{-\kappa} / (nh)^{-1} \to 0$. In particular, if $\kappa \leq 1 + \alpha_i^{-1}$, then (2.6) holds regardless of the value of D_i. This observation motivated our definition $D(A_1, c, \alpha, \kappa) = 0$ in the case $\kappa \leq 1 + \alpha^{-1}$.

REMARK 2.2. In view of Remark 2.1, if

$$\kappa \leq 1 + \min(\alpha_1^{-1}, \alpha_2^{-1}), \tag{2.7}$$

then the tail-effect terms may be ignored and

$$V = (nh)^{-1} \tfrac{1}{2} a \int_{-\infty}^\infty K^2(z) \, dz + o\{(nh)^{-1}\}. \tag{2.8}$$

This expansion has an analogue in the case of squared-error loss, where the variance component is given by

$$V^* \equiv \int_{-\infty}^\infty E\{\hat{f}(x|h) - E\hat{f}(x|h)\}^2 \, dx$$

and satisfies

$$V^* = (nh)^{-1} \int_{-\infty}^{\infty} K^2(z)\, dz + o\{(nh)^{-1}\}.$$

REMARK 2.3. It is not difficult to generalise this result. In particular, if f satisfies (2.1) and if the kernel K is bounded, integrates to unity and satisfies

$$K(z) \geq A_2 \exp(-A_1|z|^\kappa), \qquad -\infty < z < \infty,$$

for $A_1, A_2 > 0$ and some κ with property (2.7), then V admits expansion (2.8).

REMARK 2.4. A necessary and sufficient condition for the right-hand side of (2.6) to converge to zero as $n \to \infty$ and $h \to 0$ is $nh^\beta \to \infty$, where

$$\beta \equiv \max\left[1, \left\{\kappa^{-1} + (\alpha_1 + 1)^{-1}\right\}^{-1}, \left\{\kappa^{-1} + (\alpha_2 + 1)^{-1}\right\}^{-1}\right].$$

For example, if we take $\kappa = 2$ (corresponding to the standard normal kernel) and $\alpha_1 = \alpha_2 = 5$, then we require $nh^{3/2} \to \infty$ in order to achieve mean consistency with respect to Kullback–Leibler loss. Mean consistency with respect to squared-error loss requires only $nh \to \infty$.

2.2. *Bias component B.* We shall describe the bias component under the assumption that f has two derivatives on $(0, a)$. The reader concerned with "weakest possible" assumptions will notice that our regularity conditions can be considerably weakened if $\max(\alpha_1, \alpha_2) < 3$, since there the bias component is of larger order than h^4 and so the full force of the second derivative condition is not required. However, refinements such as this relegate against the economies in complexity and length which we are trying to achieve.

Assume f satisfies the following condition:

f is bounded away from zero on $(\varepsilon, a - \varepsilon)$ for each $\varepsilon > 0$;
f'' exists and is almost everywhere continuous on $(0, a)$;
$f(x) \sim c_1 x^{\alpha_1}$ and $f(a - x) \sim c_2 x^{\alpha_2}$ as $x \downarrow 0$, where $c_1, c_2 > 0$ and $\alpha_1, \alpha_2 \geq 0$;
f'' satisfies

$$|f''(x)| \leq \begin{cases} Cx^{\alpha_1 - 2}, & \text{if } \alpha_1 \neq 0 \text{ or } 1, \\ C, & \text{if } \alpha_1 = 0 \text{ or } 1, \end{cases} \tag{2.9}$$

$$|f''(a - x)| \leq \begin{cases} Cx^{\alpha_2 - 2}, & \text{if } \alpha_2 \neq 0 \text{ or } 1, \\ C, & \text{if } \alpha_2 = 0 \text{ or } 1, \end{cases} \qquad \text{for } 0 < x \leq \tfrac{1}{2}a;$$

$$f''(x) \sim c_1\alpha_1(\alpha_1 - 1)x^{\alpha_1 - 2} \quad \text{as } x \downarrow 0 \text{ if } \alpha_1 = 3,$$

$$f''(a - x) \sim c_2\alpha_2(\alpha_2 - 1)x^{\alpha_2 - 2} \quad \text{as } x \downarrow 0 \text{ if } \alpha_2 = 3.$$

REMARK 2.5. The bounds on $|f''|$ in this assumption represent "second derivative versions" of bounds on f, except for the cases $\alpha_i = 0$ or 1. These cases are qualitatively different from the others, since they usually indicate expansions such as $f(x) = C_0 + C_1 x + C_2 x^2 + \cdots$ as $x \downarrow 0$. Hence the condition $|f''(x)| \leq C$ for $\alpha_1 = 0$ or 1.

Assume the kernel K is bounded, nonnegative, integrates to unity and satisfies

$$\int_{-\infty}^{\infty} |z|^{\max(\alpha_1, \alpha_2, 2)} K(z)\, dz < \infty \quad \text{and} \quad \int_{-\infty}^{\infty} zK(z)\, dz = 0. \tag{2.10}$$

Next we define the coefficients appearing in expansions of the bias component. Let b_1 and b_2 be the nonnegative functions given by

$$b_1(x) \equiv \int_{-\infty}^{x} (1 - x^{-1}z)^{\alpha_1} K(z)\, dz \quad \text{and} \quad b_2(x) \equiv \int_{-x}^{\infty} (1 + x^{-1}z)^{\alpha_2} K(z)\, dz,$$

both for $x > 0$. Define the positive constants

$$E_1 = E_1(\alpha_1, c_1)$$
$$\equiv c_1 \left[(\alpha_1 + 1)^{-1} \int_{-\infty}^{0} (-z)^{\alpha_1 + 1} K(z)\, dz + \int_0^{\infty} x^{\alpha_1} \{ b_1(x) - 1 - \log b_1(x) \}\, dx \right]$$

for $0 \le \alpha_1 < 3$,

$$E_2 = E_2(\alpha_2, c_2)$$
$$\equiv C_2 \left[(\alpha_2 + 1)^{-1} \int_0^{\infty} z^{\alpha_2 + 1} K(z)\, dz + \int_0^{\infty} x^{\alpha_2} \{ b_2(x) - 1 - \log b_2(x) \}\, dx \right]$$

for $0 \le \alpha_2 < 3$, and

$$E_i(3, c_i) \equiv 9c_i \left\{ \int_{-\infty}^{\infty} z^2 K(z)\, dz \right\}^2.$$

It may be proved that if $0 \le \alpha_i < 3$, then $b_i(x) = 1 + O(x^{-2})$ as $x \to +\infty$, so that

$$\int_0^{\infty} x^{\alpha_i} |b_i(x) - 1 - \log b_i(x)|\, dx < \infty.$$

THEOREM 2.2. *Assume conditions* (2.9) *and* (2.10). *Then if* $\min(\alpha_1, \alpha_2) < 3$,

$$B = \sum_{i=1}^{2} E_i h^{\alpha_i + 1} + o\left(\sum_{i=1}^{2} h^{\alpha_i + 1} \right);$$

if $\alpha_1 \ne \alpha_2$ *and* $\alpha_i \equiv \min(\alpha_1, \alpha_2) = 3$,

$$B = E_i h^4 \log h^{-1} + o(h^4 \log h^{-1});$$

if $\alpha_1 = \alpha_2 = 3$,

$$B = (E_1 + E_2) h^4 \log h^{-1} + o(h^4 \log h^{-1});$$

and if $\min(\alpha_1, \alpha_2) > 3$,

$$B = \tfrac{1}{8} h^4 \left\{ \int_{-\infty}^{\infty} z^2 K(z)\, dz \right\}^2 \int_0^a \{ f''(x) \}^2 \{ f(x) \}^{-1} dx + o(h^4), \tag{2.11}$$

all as $h \to \infty$.

REMARK 2.6. Expansion (2.11) has an analogue in the case of squared-error loss, where the bias component is given by

$$B^* \equiv \int_{-\infty}^{\infty} \{ E\hat{f}(x|h) - f(x) \}^2\, dx.$$

Assuming conditions (2.9) and (2.10) and that $\min(\alpha_1, \alpha_2) > 2$ (rather than > 3),

$$B^* = \tfrac{1}{4}h^4\Big\{\int_{-\infty}^{\infty} z^2K(z)\,dz\Big\}^2 \int_0^a \{f''(x)\}^2\,dx + o(h^4).$$

2.3. *Expected loss* $l_n(h) = V + B$. The formula for V involves three terms and that for B has up to two terms. Therefore, the formula for expected loss $l_n(h)$ contains up to five terms, of which those comprising V are decreasing in h and those comprising B are increasing in h. Minimising $l_n(h)$ involves balancing these components against one another. The overall balance may be achieved in various ways, depending on relative values of α_1, α_2 and κ. We mention here only one case and stress that the order of magnitude of Kullback–Leibler loss can be minimised by judicious choice of the kernel K.

Our example shows that the standard normal kernel can be inappropriate even for compactly supported densities. We pointed out in Subsection 1.2 that it can lead to inconsistency with thick-tailed densities. Take $\kappa = 2$ in definition (2.2) and assume f satisfies (2.9) with $3 < \min(\alpha_1, \alpha_2) < \alpha_j \equiv \max(\alpha_1, \alpha_2) > 9$. Then the bias component B is asymptotic to a constant multiple of h^4 (see Section 2.2), while the variance component V is asymptotic $C_1(nh)^{-1} + C_2n^{-1-2/(\alpha_j+1)}h^{-2}$ (see Section 2.1). Expected loss, $l_n(h) = V + B$, is minimised by taking $h \sim \text{const.}n^{-(\alpha_j+3)/6(\alpha_j+1)}$, which decreases more slowly than $n^{-1/5}$ and leads to a minimum loss whose order of magnitude is greater than $n^{-4/5}$. Had we chosen $\kappa \le 1$ (e.g., the double exponential kernel), then the optimal h would have been of order $n^{-1/5}$ and the minimum loss of order $n^{-4/5}$, which incidentally are the same orders as in the case of squared-error loss.

Recall from Remark 2.1 that if $\kappa \le 1$, then tail-effect terms make a negligible contribution to V, regardless of the values of α_1 and α_2.

2.4. *Likelihood cross-validation.* In the case of squared-error loss, the windows which minimise expected loss and raw loss, and which maximise the cross-validatory criterion, are asymptotically equivalent to one another. But in the case of Kullback–Leibler loss, no one of these windows is asymptotically equivalent to any one of the other two, in general, and so neither expected loss nor raw loss provides a natural vantage point for viewing cross-validation. In this subsection we shall examine likelihood cross-validation in the context of expected loss, which is in keeping with our work in Sections 2.1–2.3. The case of raw loss will be treated in Section 2.5.

Notice that the window $\hat{h}_0$ which maximises $CV(h)$ does not depend on whether we normalise by n^{-1} or by $(n-1)^{-1}$ in our definition of $\hat{f}_i$. (In contrast, choice of normalisation can have a slight effect in the case of squared-error cross-validation.) We choose to normalise by $(n-1)^{-1}$, and so

$$-E\{CV(h)\} + \int_{-\infty}^{\infty} f(x)\log f(x)\,dx = l_{n-1}(h), \tag{2.12}$$

where $l_{n-1}(h)$ is expected Kullback–Leibler loss for a sample of size $n-1$. Therefore, in maximising CV we are in effect minimising an unbiased estimate of expected Kullback–Leibler loss.

Formula (2.12) suggests that the extent to which the windows which maximise CV also minimise l_n can be explored by studying the stochastic process

$$CV(h) - E\{CV(h)\}, \qquad h > 0.$$

We shall confine attention to h-values in the range $n^{-1+\varepsilon} \le h \le n^{-\varepsilon}$ for arbitrarily small $\varepsilon > 0$. This restriction is slightly stronger than the minimum required for pointwise consistency. However, Theorems 2.1 and 2.2 show that the window h_0 which minimises $l_n(h)$ is asymptotic to const. $n^{-\beta}(\log n)^{\gamma}$ for some $0 < \beta < 1$, and so restriction to $n^{-1-\varepsilon} \le h \le n^{-\varepsilon}$ for $\varepsilon < \min(\beta, 1-\beta)$ seems appropriate.

Let $CV^*(h)$ denote the criterion analogous to $CV(h)$ in the case of squared-error loss (see, for example, [3], [4] and [26]). Define $l_n^*(h)$ to equal mean integrated squared error in the sample of size n. It is known that in quite general circumstances,

$$CV^*(h) - E\{CV^*(h)\} = Q^*(n) + o_p\{l_n^*(h)\} \tag{2.13}$$

uniformly in h, where $Q^*(n)$ denotes a random variable *not depending on* h. (See [16], [17] and [28].) This result, together with the observation that

$$-E\{CV^*(h)\} + \int_{-\infty}^{\infty} f^2(x)\,dx = l_{n-1}^*(h) = l_n^*(h) + o\{l_n^*(h)\}$$

[the squared-error analogue of (2.12)], indicates that the window which maximises CV^* is asymptotically equivalent to the window which minimises l_n^*. A key result in the case of likelihood cross-validation is that the analogue of (2.13) *fails to hold*. The process $CV(h) - E\{CV(h)\}$ can contain terms which depend on h and which are not negligible relative to $l_n(h)$. Therefore, the windows which maximise CV and minimise l_n are not always asymptotic to one another.

The offending terms in an expansion $CV - E(CV)$ are connected with the tail-effect terms in an expansion of the variance component of V (see Section 2.1). If the kernel K is chosen so that the tail-effect terms are negligible in comparison to the main-effect term in V, then it will be true that

$$CV(h) - E\{CV(h)\} = Q(n) + o_p\{l_n(h)\} \tag{2.14}$$

uniformly in h, where $Q(n)$ does not depend on h. In this case, likelihood cross-validation will produce a window which is asymptotic to that which minimises l_n.

To describe the terms in CV which can cause difficulties, we introduce the random variables

$$T_i \equiv \min_{1 \le j \le n,\, j \ne i} |X_i - X_j|, \qquad 1 \le i \le n,$$

$$W_{n1} \equiv \begin{cases} n^{\kappa/(\alpha_1+1)} \sum_{i=1}^{n} \left[T_i^{\kappa} I\left(X_i \le \tfrac{1}{2}a\right) - E\left\{T_i^{\kappa} I\left(X_i \le \tfrac{1}{2}a\right)\right\}\right], & \text{if } \kappa > 1 + \alpha_1^{-1}, \\ 0, & \text{if } \kappa \le 1 + \alpha_1^{-1}, \end{cases}$$

and

$$W_{n2} \equiv \begin{cases} n^{\kappa/(\alpha_2+1)} \sum_{i=1}^{n} \left[T_i^{\kappa} I\left(X_i > \tfrac{1}{2}a\right) - E\left\{ T_i^{\kappa} I\left(X_i > \tfrac{1}{2}a\right)\right\}\right], & \text{if } \kappa > 1 + \alpha_2^{-1}, \\ 0, & \text{if } \kappa \le 1 + \alpha_2^{-1}. \end{cases}$$

It may be proved that if $\kappa > 1 + \alpha_1^{-1}$ and if f satisfies (2.1), then

$$E\left\{ T_1^{\kappa} I\left(X_1 \le \tfrac{1}{2}a\right)\right\} = O(n^{-1-\kappa/(\alpha_1+1)})$$

as $n \to \infty$, so that $W_{n1} = O_p(1)$. Similarly, $W_{n2} = O_p(1)$. Note particularly that neither W_{n1} nor W_{n2} depends on h. The remainder term $Q(n)$ appearing in expansions such as (2.14) is given by

$$Q(n) = n^{-1} \sum_{i=1}^{n} \{\log f(X_i) - E \log f(X_i)\}.$$

Recall from Sections 2.1 and 2.2 that the variance and bias components of expected loss contain terms of orders $(nh)^{-1}$, $n^{-1-\kappa/(\alpha_1+1)}h^{-\kappa}$, $n^{-1-\kappa/(\alpha_2+1)}h^{-\kappa}$, h^{α_1+1} and h^{α_2+1}, among others. Therefore, the quantity

$$J(n, h) \equiv (nh)^{-1} + \sum_{i=1}^{2} (n^{-1-\kappa/(\alpha_i+1)}h^{-\kappa} + h^{\alpha_i+1}) + B(h)$$

is of the same order of magnitude as expected Kullback–Leibler loss. By using this expansion to describe expected loss, we avoid having to impose unnecessary smoothness assumptions on the density f; note the remarks in the first paragraph of Section 2.2.

Assume the following condition on f:

(2.15) f vanishes outside $[0, a]$, and satisfies $C_1 x^{\alpha_1} \le f(x) \le C_2 x^{\alpha_1}$ and $C_1 x^{\alpha_2} \le f(a - x) \le C_2 x^{\alpha_2}$ for $0 < x \le \frac{1}{2}a$, where $C_1, C_2 > 0$ and $\alpha_1, \alpha_2 \ge 0$.

Assume the following condition on K:

(2.16) K integrates to unity, and either

$$K(z) \equiv A_2 \exp(-A_1|z|^{\kappa}), \qquad -\infty < z < \infty,$$

for constants $A_1, A_2 > 0$ and $\kappa > 1$, or

$$K(z) \ge A_2 \exp(-A_1|z|), \qquad -\infty < z < \infty,$$

for constants $A_1, A_2 > 0$. In the latter case, assume in addition that K is of bounded variation and Hölder continuous on $(-\infty, \infty)$ and satisfies

$$\int_{-\infty}^{\infty} |z|^{1+\max(\alpha_1, \alpha_2)} \left\{ K(z)\, dz + |dK(z)| \right\} < \infty.$$

We again adopt the convention that $\kappa = 1$ if $K(z) \ge A_2 \exp(-A_1|z|)$, $-\infty < z < \infty$. Hölder continuity of K means that for some C, $s > 0$, $|K(u) - K(v)| \le C|u - v|^s$ for all $-\infty < u, v < \infty$.

THEOREM 2.3. *Assume* (2.15) *and* (2.16). *Then*

$$CV(h) - E\{CV(h)\} = Q(n) - A_1 \sum_{i=1}^{2} n^{-1-\kappa/(\alpha_i+1)} h^{-\kappa} W_{ni} + R_1(n, h), \tag{2.17}$$

where for any $0 < \varepsilon < 1$,

$$\sup_{n^{-1+\varepsilon} \le h \le n^{-\varepsilon}} J(n, h)^{-1} |R_1(n, h)| \to 0$$

in probability as $n \to \infty$.

REMARK 2.7. If $\kappa \le 1 + \min(\alpha_1^{-1}, \alpha_2^{-1})$ then $W_{n1} = W_{n2} = 0$ and so (2.17) may be written in the form

$$CV(h) - E\{CV(h)\} = Q(n) + o_p\{J(n, h)\},$$

which is the analogue of (2.13) in the case of Kullback–Leibler loss. But in general the "tail-effect" terms cannot be ignored.

Next we shall show that when $\kappa > 1 + \alpha_i^{-1}$, W_{ni} has a proper nondegenerate limit. We begin by defining this limit. Let $Z_1, Z_2, \ldots,$ denote independent negative-exponential random variables. Given $d > 0$ and $\alpha \ge 0$, set

$$V_{i1} = V_{i1}(d, \alpha) \equiv d \exp\left[-(\alpha + 1)^{-1} \left\{ \sum_{j=1}^{\infty} (Z_j - 1) j^{-1} - \sum_{j=1}^{i-1} j^{-1} + \gamma \right\} \right], \quad i \ge 1,$$

where γ is Euler's constant. Then $V_{i+1,1} \ge V_{i1}$ and so

$$V_{i2} = V_{i2}(d, \alpha, \kappa) \equiv (V_{i+1,1} - V_{i1})^{\kappa}, \qquad i \ge 1,$$

is well defined. Set $V_{02} \equiv +\infty$ and

$$V = V(d, \alpha, \kappa) \equiv \sum_{i=1}^{\infty} \min(V_{i2}, V_{i-1,2}).$$

It may be shown that $E(V_{i2}) \le C i^{-\alpha\kappa/(\alpha+1)}$ for $i \ge 1$ and so $\sum_1^\infty E(V_{i2}) < \infty$ if $\kappa > 1 + \alpha^{-1}$. This proves that $V < \infty$ almost surely and $E(V) < \infty$ for $\kappa > 1 + \alpha^{-1}$. Define

$$W(d, \alpha, \kappa) \equiv V(d, \alpha, \kappa) - E\{V(d, \alpha, \kappa)\}.$$

THEOREM 2.4. *Assume condition* (2.1). *If* $\kappa > 1 + \alpha_i^{-1}$, *then*

$$W_{ni} \to W\left[\{c_i^{-1}(\alpha_i + 1)\}^{1/(\alpha_i+1)}, \alpha_i, \kappa \right] \tag{2.18}$$

in distribution as $n \to \infty$. *If* $\kappa > 1 + \max(\alpha_1^{-1}, \alpha_2^{-1})$, *then* W_{n1} *and* W_{n2} *are asymptotically independent.*

REMARK 2.8. Let h_0 and $\hat{h}_0$ be the values of h which minimise $l_n(h)$ and maximise $CV(h)$, respectively, chosen from within the range $n^{-1+\varepsilon} \le h \le n^{-\varepsilon}$ for very small $\varepsilon > 0$. Combining Theorems 2.1–2.4 and noting the identity (2.12), we see that $\hat{h}_0/h_0 \to 1$ in probability if and only if the tail-effect terms make an

asymptotically negligible contribution to $l_n(h)$ when $h = h_0$. If the tail-effect terms make a significant contribution, then $\hat{h}_0/h_0$ has a proper, nondegenerate limiting distribution with no atom at the origin. In this sense, likelihood cross-validation is guaranteed to produce a window $\hat{h}_0$ of the same order of magnitude as h_0, but not necessarily a window asymptotic to h_0.

REMARK 2.9. If the kernel K satisfies

$$K(z) \geq A_2 \exp(-A_1|z|), \qquad -\infty < z < \infty; \tag{2.19}$$

then the tail-effect terms are guaranteed to be asymptotically negligible (see Section 2.1) and in that case, $\hat{h}_0/h_0 \to 1$ in probability. It may be shown that this result continues to hold if (2.19) is weakened to

$$K(z) \geq A_2 \exp(-A_1|z|^\kappa), \qquad -\infty < z < \infty,$$

where $\kappa \leq 1 + \min(\alpha_1^{-1}, \alpha_2^{-1})$. [The conditions on K in the last sentence of (2.16) should also be assumed.]

2.5. *Minimisation of raw Kullback–Leibler loss.* Raw loss is defined by

$$L_n(h) \equiv \int_{-\infty}^{\infty} f(x) \log\{ f(x)/\hat{f}(x|h) \}\, dx. \tag{2.20}$$

Let $\tilde{h}_0$ denote the value of h which minimises $L_n(h)$, restricted to the interval $n^{-1+\varepsilon} \leq h \leq n^{-\varepsilon}$ for very small $\varepsilon > 0$. Then h_0, $\tilde{h}$ and $\hat{h}_0$ are all of the same order of magnitude and if one of $\hat{h}_0/h_0$, $\tilde{h}_0/h_0$ and $\hat{h}_0/\tilde{h}_0$ has a nondegenerate weak limit, then so do the other two. But if the kernel K is correctly chosen [for example, if K satisfies (2.19)], then all three ratios converge to unity in probability. In this case, $L_n(\tilde{h}_0)/l_n(h_0) \to 1$ and

$$L_n(\hat{h}_0)/L_n(\tilde{h}_0) \to 1 \tag{2.21}$$

in probability, so that $\hat{h}_0$ achieves asymptotic minimisation of raw Kullback–Leibler loss.

We shall outline the main points of this theory. Define

$$\tilde{D}_1 \equiv \begin{cases} A_1 n^{1+\kappa/(\alpha_1+1)} \int_0^{a/2} \Big\{ \min_{1 \leq i \leq n} |X_i - x|^\kappa \Big\} f(x)\, dx, & \text{if } \kappa > 1 + \alpha_1^{-1}, \\ 0, & \text{if } \kappa \leq 1 + \alpha_1^{-1} \end{cases}$$

and

$$\tilde{D}_2 \equiv \begin{cases} A_1 n^{1+\kappa/(\alpha_2+1)} \int_{a/2}^{a} \Big\{ \min_{1 \leq i \leq n} |X_i - x|^\kappa \Big\} f(x)\, dx, & \text{if } \kappa > 1 + \alpha_2^{-1}, \\ 0, & \text{if } \kappa \leq 1 + \alpha_2^{-1}, \end{cases}$$

and recall the definitions of D_1 and D_2 given in Section 2.1.

THEOREM 2.5. *Assume conditions* (2.15) *and* (2.16). *Then*

$$L_n(h) - l_n(h) = \sum_{i=1}^{2} n^{-1-\kappa/(\alpha_i+1)} h^{-\kappa} \big(\tilde{D}_i - D_i \big) + R_2(n, h), \tag{2.22}$$

where for any $0 < \varepsilon < \frac{1}{2}$,

$$\sup_{n^{-1+\varepsilon} \le h \le n^{-\varepsilon}} J(n, h)^{-1} |R_2(n, h)| \to 0$$

in probability as $n \to \infty$.

REMARK 2.10. Result (2.21) in the case of a kernel satisfying (2.19) follows from Theorems 2.1–2.3 and 2.5 and the fact that $\tilde{D}_i = O_p(1)$.

REMARK 2.11. The random variable $\tilde{D}_i - D_i$ in (2.22) is related to the term $A_1 W_{ni}$ in the expansion (2.17). In fact, a version of Theorem 2.4 may be proved for $\tilde{D}_i - D_i$, declaring that under condition (2.1) and for $\kappa > 1 + \alpha_i^{-1}$, $\tilde{D}_i - D_i$ has a proper nondegenerate limit distribution having no atom at the origin. The limit is different from that in Theorem 2.4. Therefore, result (2.21) will fail to hold if the tail-effect terms make a significant contribution to $l_n(h_0)$.

2.6. *Proofs.* The proofs are tedious rather than difficult and so are given only in outline.

PROOF OF THEOREM 2.1. We begin with two lemmas. Let $T = T(x) \equiv \min_{1 \le i \le n} |x - X_i|$, $\tau = \tau(x) \equiv n^{-1/(\alpha_1+1)}$ if $0 < x \le n^{-1/(\alpha_1+1)}$, $\tau \equiv (nx^{\alpha_1})^{-1}$ if $x > n^{-1/(\alpha_1+1)}$ and $U = U(x) \equiv \{\hat{f}(x|h) - E\hat{f}(x|h)\}/E\hat{f}(x|h)$.

LEMMA 2.6. *If* f *satisfies* (2.1), *then for any* $\varepsilon \in (0, a)$ *and* $\beta > 0$, *there exists a positive constant* $C(\varepsilon, \beta)$ *such that* $E(T^\beta) \le C(\varepsilon, \beta)\tau^\beta$ *for* $0 < x < a - \varepsilon$.

PROOF. The distribution function G of $|x - X|$ is given by

$$G(y) = P(x - y < X < x + y) = y\int_{-1}^{1} f(x + ty)\, dt,$$

from which it follows that $G(y) \ge C_1 y(x \vee y)^{\alpha_1}$ for small x, y. Therefore,

$$G^{-1}(u) \le \sup\{y: C_1 y(x \vee y)^{\alpha_1} \le u\} \le C_2(ux^{-\alpha_1} \wedge u^{1/(\alpha_1+1)}),$$

whence if $0 < x \le 1$,

$$\begin{aligned} E(T^\beta) &= n\int_0^1 \{G^{-1}(u)\}^\beta (1 - u)^{n-1}\, du \\ &\le nC_2^\beta \left\{ x^{-\alpha_1\beta} \int_0^{x^{\alpha_1+1}} u^\beta (1 - u)^{n-1}\, du + \int_{x^{\alpha_1+1}}^1 u^{\beta/(\alpha_1+1)} (1 - u)^{n-1}\, du \right\}. \end{aligned}$$

Noting that $(1 - u)^{n-1} \le e \cdot e^{-nu}$ for $0 < u \le 1$ and changing the variable to $v \equiv nu$, we get

$$E(T^\beta) \le nC_3 \left(x^{-\alpha_1\beta} n^{-\beta-1} \int_0^{nx^{\alpha_1+1}} v^\beta e^{-v}\, dv + n^{-\beta/(\alpha_1+1)-1} \int_{nx^{\alpha_1+1}}^\infty v^{\beta/(\alpha_1+1)} e^{-v}\, dv \right).$$

If $nx^{\alpha_1+1} > 1$, we bound this by

$$nC_4 \left\{ x^{-\alpha_1\beta} n^{-\beta-1} + n^{-\beta/(\alpha_1+1)-1} (nx^{\alpha_1+1})^{\beta/(\alpha_1+1)+1} \exp(-nx^{\alpha_1+1}) \right\} \le C_5 \tau^\beta,$$

while if $nx^{\alpha_1+1} \le 1$, we bound it by

$$nC_4\{x^{-\alpha_1}\beta n^{-\beta-1}(nx^{\alpha_1+1})^{\beta+1} + n^{-\beta/(\alpha_1+1)-1}\} \le 2C_4\tau^\beta.$$

This gives the desired result. (The case $x > 1$ is trivial.) □

LEMMA 2.7. *If f satisfies* (2.1), *then for any $\varepsilon \in (0, a)$ there exists a positive constant $C(\varepsilon)$ such that $P(U < -\frac{1}{2}) \le \exp\{-C(\varepsilon)nhx^{\alpha_1}\}$ for $0 < x < a - \varepsilon$.*

PROOF. First we show that with $W \equiv \hat{f}(x|h)$ and

$$\sigma^2 \equiv \operatorname{var}[K\{(x - X)/h\}/EK\{(x - X)/h\}]$$

we have

$$P(W/EW < \tfrac{1}{2}) \le \exp\{-(3n/4)(6\sigma^2 + 1)^{-1}\}. \tag{2.23}$$

Notice that

$$W/EW = 1 - n^{-1}\sum_{i=1}^{n}[1 - \mu_1^{-1}K\{(x - X_i)/h\}],$$

where $\mu_j \equiv E[K^j\{(x - X)/h\}]$. Result (2.23) now follows directly from the one-sided version of Bernstein's inequality (Hoeffding [21], page 17, with $b = 1$, $t = \frac{1}{2}$ and $\sigma^2 = \mu_1^{-2}\mu_2 - 1$).

A little algebra shows that $\mu_1 \ge C_1h(x + h)^{\alpha_1}$ and $\mu_2 \le C_2h(x + h)^{\alpha_2}$, so that $\sigma^2 \le C_3h^{-1}x^{-\alpha_1}$. Lemma 2.7 follows from this estimate and (2.23). □

Decompose V as

$$\begin{aligned} V &= V_1 + V_2 + V_3 \\ &= \left(\int_0^\varepsilon + \int_\varepsilon^{a-\varepsilon} + \int_{a-\varepsilon}^{a}\right) f(x)E\left[\log\{E\hat{f}(x|h)/\hat{f}(x|h)\}\right] dx. \end{aligned} \tag{2.24}$$

To handle V_2, note that

$$\left|\log(1 + u) - u + \tfrac{1}{2}u^2\right| \le C\{|u|^3 - \log(1 + u)I(u < -\tfrac{1}{2})\}, \qquad u > -1,$$

so that

$$\begin{aligned} D &\equiv \left|V_2 - \tfrac{1}{2}\int_\varepsilon^{a-\varepsilon} f \operatorname{var}(U)\, dx\right| \\ &\le C\int_\varepsilon^{a-\varepsilon} f\left[E(|U|^3) + E\{|\log(1 + U)|I(U < -\tfrac{1}{2})\}\right] dx. \end{aligned}$$

From this inequality, the bounds

$$\begin{aligned} E(|U|^3) &\le (EU^4)^{3/4} = O\{(nh)^{-3/2}\}, \\ -\log(1 + U) &\le A_1(T/h)^\kappa + \log(nh) + C \qquad [\text{see } (1.4)], \\ E\{T^\kappa I(U < -\tfrac{1}{2})\} &\le (ET^{2\kappa})^{1/2}P(U < -\tfrac{1}{2})^{1/2} \end{aligned}$$

and Lemmas 2.6 and 2.7, we may deduce that $D = o\{(nh)^{-1}\}$. Furthermore,

$nh \operatorname{var}(U) \to f\int K^2$ at continuity points of f and is uniformly bounded. Therefore,

$$V_2 = (nh)^{-1}\tfrac{1}{2}(a - 2\varepsilon)\int_{-\infty}^{\infty} K^2(z) + o\{(nh)^{-1}\}. \quad . \tag{2.25}$$

The term V_1 is comparatively easy to treat if we assume $\kappa \le 1 + \alpha_1^{-1}$, since that entails $n^{-1-\kappa/(\alpha_1+1)}h^{-\kappa} = o\{(nh)^{-1}\}$. Therefore, we assume $\kappa > 1 + \alpha_1^{-1}$. Noting that $|\log(1+u) - u - \log(1+u)I(u < -\frac{1}{2})| \le Cu^2$, $u > -1$, we find that

$$\left|V_1 + E\left\{\int_0^{\varepsilon} f \log(\hat{f}/E\hat{f})\, I(\hat{f}/E\hat{f} < \tfrac{1}{2})\, dx\right\}\right| \le C\varepsilon(nh)^{-1}. \tag{2.26}$$

Using the bounds $A_1(T/h)^{\kappa} - C\log n \le -\log(\hat{f}/E\hat{f}) \le A_1(T/h)^{\kappa} + C\log(nh)$ [see (1.4)], Lemmas 2.6 and 2.7 and the fact that $\kappa > 1 + \alpha_1^{-1}$ implies $n^{\delta}(nh)^{-1-(1/\alpha_1)} = o\{(nh)^{-1} + n^{-1-\kappa/(\alpha_1+1)}h^{-\kappa}\}$ for some $\delta > 0$, we obtain from (2.26),

$$\begin{aligned}\left|V_1 - A_1\int_0^{\varepsilon} fE(T/h)^{\kappa}\, dx\right| &\le C\varepsilon(nh)^{-1} + o\left\{(nh)^{-1} + n^{-1-\kappa/(\alpha_1+1)}h^{-\kappa}\right\} \\ &\quad + Ch^{-\kappa}\int_0^{\varepsilon} x^{\alpha_1}\tau^{\kappa}P(\hat{f}/E\hat{f} > \tfrac{1}{2})^{1/2}\, dx.\end{aligned} \tag{2.27}$$

To treat the last-written integral, observe that for $0 < x < \varepsilon$,

$$\begin{aligned}P(\hat{f}/E\hat{f} > \tfrac{1}{2}) &\le P\left[\sum_{i=1}^{n} \exp\left\{-A_1|(x - X_i)/h|^{\kappa}\right\} > C_1 n^{-\alpha_1}\right] \\ &\le nP\left\{|x - X_1| \le C_2 h(\log n)^{1/\kappa}\right\} \\ &\le C_3\left[nh(\log n)^{1/\kappa}x^{\alpha_1} + n\left\{h(\log n)^{1/\kappa}\right\}^{\alpha_1+1}\right].\end{aligned}$$

Using this bound it may be shown that the right-hand side of (2.27) equals $C\varepsilon(nh)^{-1} + o\{(nh)^{-1} + n^{-1-\kappa/(\alpha_1+1)}h^{-\kappa}\}$. A little analysis shows that

$$\lim_{n\to\infty} n^{1+\kappa/(\alpha_1+1)}\int_{rn^{-1/(\alpha_1+1)}}^{sn^{-1/(\alpha_1+1)}} x^{\alpha_1}E(T^{\kappa})\, dx = \int_r^s y^{\alpha_1}\, dy\int_0^{\infty} g(x, y)^{\kappa}e^{-x}\, dx$$

for $0 < r < s < \infty$. This result and judicious use of Lemma 2.6 give us

$$\lim_{n\to\infty} n^{1+\kappa/(\alpha_1+1)}A_1\int_0^{\varepsilon} f(x)E(T^{\kappa})\, dx = D_1$$

and so by (2.27),

$$\left|V_1 - n^{-1-\kappa/(\alpha_1+1)}h^{-\kappa}D_1\right| \le C\varepsilon(nh)^{-1} + o\left\{(nh)^{-1} + n^{-1-\kappa/(\alpha_1+1)}h^{-\kappa}\right\}. \tag{2.28}$$

Theorem 2.1 follows from (2.24), (2.25), (2.28) and an analogue of (2.28) for V_3. □

PROOF OF THEOREM 2.2. Write

$$\begin{aligned}B + \int_0^a (E\hat{f} - f)\, dx &= B_1 + B_2 + B_3 \\ &= \left(\int_0^{\varepsilon} + \int_{\varepsilon}^{a-\varepsilon} + \int_{a-\varepsilon}^{a}\right)\{f\log(f/E\hat{f}) + E\hat{f} - f\}\, dx\end{aligned} \tag{2.29}$$

and notice that if F is the distribution function derived from f,

$$-\int_0^a (E\hat{f} - f)\,dx = \int_{-\infty}^{\infty} K(z)[F(-hz) + \{1 - F(a - hz)\}]\,dz$$

$$\sim c_1(\alpha_1 + 1)^{-1}h^{\alpha_1+1}\int_{-\infty}^{0}(-z)^{\alpha_1+1}K(z)\,dz \tag{2.30}$$

$$+c_2(\alpha_2 + 1)^{-1}h^{\alpha_2+1}\int_0^{\infty} z^{\alpha_2+1}K(z)\,dz.$$

Taylor expansion shows that B_2 admits the formula (2.11), but with $\int_\varepsilon^{a-\varepsilon}$ replacing $\int_0^a$. The terms in h^{α_i+1} in B derive from B_1, B_3 and (2.30). We shall prove only that if $\alpha_1 < 3$, then $B_1 \sim E_1' h^{\alpha_1+1}$, where

$$E_1' \equiv c_1\int_0^{\infty} x^{\alpha_1}\{b_1(x) - 1 - \log b_1(x)\}\,dx.$$

Let $I_1(x) \equiv E\{\hat{f}(hx|h)\}$, $J_1(x) \equiv f(hx)$, $J_2(x) \equiv c_1(hx)^{\alpha_1}$ and

$$I_2(x) \equiv c_1 h^{\alpha_1}\int_{-\infty}^{x}(x - z)^{\alpha_1}K(z)\,dz.$$

Define $R = R(x)$ by $I_1/J_1 = (I_2/J_2)(1 + R)$. If $h \to 0$, $r = r(h) \to \infty$ and $hr \to 0$, then $\sup_{0 \le x \le r}|R(x)| \to 0$. Given $r_1 \to \infty$ such that $hr_1 \to 0$, choose $r \le r_1$ such that $r \to \infty$ and $r^{\alpha_1+1}\sup_{0 \le x \le r_1}|R(x)| \to 0$. For this r,

$$\int_0^{hr}\{f \log(f/E\hat{f}) + E\hat{f} - f\}\,dx$$

$$= -h\int_0^{r} f(hx)\left[\log\{I_1(x)/J_1(x)\} - \{I_1(x) - J_1(x)\}J_1(x)^{-1}\right]dx$$

$$\sim -c_1\int_0^{\infty} x^{\alpha_1}[\log b_1(x) - \{b_1(x) - 1\}]\,dx.$$

For any $r \to \infty$ such that $hr \to 0$ and $\varepsilon > 0$,

$$\int_{hr}^{\varepsilon}|f \log(f/E\hat{f}) + E\hat{f} - f|\,dx = o(h^{\alpha_1+1}).$$

Combining these two estimates we conclude that $B_1 \sim E_1' h^{\alpha_1+1}$. □

PROOF OF THEOREM 2.3. A key trick is the observation that several terms are more easily handled if $\mu \equiv E(\hat{f})$ is replaced by $\xi \equiv \mu + (nh)^{-1}$. Therefore, we write $CV - E(CV) = Q(n) + S_1 + S_2$, where

$$S_1 \equiv n^{-1}\sum_{i=1}^{n}\{\log \rho(X_i) - E\log \rho(X_i)\},$$

$$S_2 \equiv n^{-1}\sum_{i=1}^{n}\{\log \rho_i(X_i) - E\log \rho_i(X_i)\},$$

$\rho \equiv \xi/f$ and $\rho_i \equiv \hat{f}_i/\xi$. An argument based on subsequences shows that it is

sufficient to prove that for each $\lambda \in (0,1)$ and sufficiently small $\delta = \delta(\lambda) > 0$,

$$\sup_{n^{-\lambda-\delta} \le h \le n^{-\lambda+\delta}} J(n,h)^{-1} |S_1(h)| \to_p 0, \tag{2.31}$$

$$\sup_{n^{-\lambda-\delta} \le h \le n^{-\lambda+\delta}} J(n,h)^{-1} \left| S_2(h) + A_1 \sum_{i=1}^{2} n^{-1-\kappa/(\alpha_i+1)} h^{-\kappa} W_{ni} \right| \to_p 0. \tag{2.32}$$

The remainder of the proof of Theorem 2.3 consists of derivations of (2.31) and (2.32).

Proof of (2.31). First we bound the moments of $S_1(h)$.

LEMMA 2.8. *Under the prescribed conditions on f and K and for any integer $t \ge 1$ and $\varepsilon \in (0, \frac{1}{2})$,*

$$E\{S_1(h)^{2t}\} \le C(t)(h^t + nh^{2t}) J(n,h)^{2t}$$

uniformly in $n^{-1+\varepsilon} \le h \le n^{-\varepsilon}$, where $C(t)$ does not depend on n or h.

PROOF. Let $y \equiv \log \rho(X|h) - E\{\log \rho(X|h)\}$. By Rosenthal's inequality (Burkholder [6, page 40])

$$E\{S_1(h)^{2t}\} \le Cn^{-2t}\{nE(Y^{2t}) + (n \operatorname{var} Y)^t\}. \tag{2.33}$$

Since $C_1 \le \rho(x|h) \le C_2/f(x)$, then $E(Y^{2t}) \le C_3 E\{\log \rho(X|h)\}^{2t} \le C_4 < \infty$. Also, $\operatorname{var}(Y) \le E\{\log \rho(X|h)\}^2 \le 2(I_1 + I_2)$, where

$$I_1 \equiv E\left[\log\{\mu(X|h)/\xi(X|h)\}\right]^2, \qquad I_2 \equiv E\left[\log\{f(X)/\mu(X|h)\}\right]^2.$$

We shall prove in the following text that $I_1 \le C(nh)^{-1}$ and $I_2 \le C(h^{\alpha_1+1} + h^{\alpha_2+1} + B)$. It then follows that $\operatorname{var}(Y) \le CJ$ and so by (2.33),

$$E\{S_1(h)^{2t}\} \le C\{n^{-2t+1} + (n^{-1}J)^t\} \le C(nh^{2t} + h^t)J^{2t},$$

as had to be shown.

Bound for I_1. Put $I_{11} \equiv E([\log\{\mu(X|h)/\xi(X|h)\}]^2 I(X \le \frac{1}{2}a))$. Notice that $C_1(x+h)^{\alpha_1} \le \mu(x|h) \le C_2(x+h)^{\alpha_1}$ for $0 < x \le \frac{1}{2}a$. If $nh^{\alpha_1+1} \ge 1$, then $1 \le \xi(x|h)/\mu(x|h) \le 1 + C_3(nh)^{-1}\{(nh)^{-1} + (x+h)^{\alpha_1}\}^{-1}$ for $0 < x \le \frac{1}{2}a$, whence it follows that $I_{11} \le C_4(nh)^{-1}$. If $nh^{\alpha_1+1} < 1$, then $1 \le \xi(x|h)/\mu(x|h) \le 1 + C_3\{nh(x+h)^{\alpha_1}\}^{-1}$ for $0 < x \le \frac{1}{2}a$ and so

$$\begin{aligned} I_{11} &\le C_4 h^{\alpha_1+1} \int_0^{a/2h} \left(\log\left[1 + C_3\{nh^{\alpha_1+1}(1+y)^{\alpha_1}\}^{-1}\right]\right)^2 y^{\alpha_1}\, dy \\ &\le C_5 (\log n)^2 h^{\alpha_1+1} \int_0^{(nh^{\alpha_1+1})^{-1/\alpha_1}} y^{\alpha_1}\, dy \\ &\quad + C_5 h^{\alpha_1+1} \int_{(nh^{\alpha_1+1})^{-1/\alpha_1}}^{a/h} (nh^{\alpha_1+1} y^{\alpha_1})^{-2} y^{\alpha_1}\, dy \\ &\le C_6 (nh)^{-1}. \end{aligned}$$

Therefore, $I_{11} \le C(nh)^{-1}$, no matter what the value of nh^{α_1+1}. An identical argument supplies the same bound to $I_{12} \equiv I_1 - I_{11}$ and so $I_1 \le C(nh)^{-1}$.

Bound for I_2. Write

$$I_2 = \left(\int_0^h + \int_h^{a-h} + \int_{a-h}^a\right)\left[\log\{f(x)/\mu(x|h)\}\right]^2 f(x)\,dx = I_{21} + I_{22} + I_{23}.$$

Since $C_1 \le \mu(x|h)/f(x) \le C_2(x+h)^{\alpha_1}/x^{\alpha_1}$ and $f(x) \le C_3 x^{\alpha_1}$ for $0 < x \le h$, then $I_{21} \le Ch^{\alpha_1+1}$. Likewise, $I_{23} \le Ch^{\alpha_2+1}$. Since $C_1 \le \mu(x|h)/f(x) \le C_2$ for $h < x \le a - h$, then

$$I_{22} \le C\int_h^{a-h}\{f(x) - \mu(x|h)\}^2 f(x)^{-1}\,dx.$$

The inequality $u - \log(1+u) \ge C(\eta)u^2$, valid for any $0 < \eta < 1$ and all $-1 + \eta \le u \le \eta^{-1}$, entails

$$\begin{aligned}
&\int_h^{a-h}\{f\log(f/\mu) + (\mu - f)\}\,dx \\
&\quad = \int_h^{a-h} f\left[(\mu - f)/f - \log\{1 + (\mu - f)/f\}\right]dx \\
&\quad \ge C_1\int_h^{a-h}(\mu - f)^2(f)^{-1}\,dx.
\end{aligned}$$

Therefore,

$$\begin{aligned}
B = \int_0^a f\log(f/\mu)\,dx &\ge C_1\int_h^{a-h}(\mu - f)^2(f)^{-1} + \int_0^a (f - \mu)\,dx \\
&\quad - \left(\int_0^h + \int_{a-h}^h\right)\{f|\log(f/\mu)| + \mu + f\}\,dx \\
&\ge C_2 I_{22} - C_3(h^{\alpha_1+1} + h^{\alpha_2+1}).
\end{aligned}$$

Combining these bounds we deduce that $I_2 \le C(h^{\alpha_1+1} + h^{\alpha_2+1} + B)$. □

Now we prove (2.31). Let $0 < \delta < \min(\lambda, 1-\lambda)$ and given $p > 0$, define $\mathscr{H} = \mathscr{H}(p)$ to be the set of all pairs (h_1, h_2) with $n^{-\lambda-\delta} \le h_i \le n^{-\lambda+\delta}$ $(i = 1,2)$ and $|h_1 - h_2| \le n^{-p}$. If $|K(u) - K(v)| \le C_1|u - v|^s$ $(s > 0)$, then

$$\begin{aligned}
|\mu(x|h_1) - \mu(x|h_2)| &= \left|\int_{-\infty}^{\infty}\{h_1^{-1}K(z/h_1) - h_2^{-1}K(z/h_2)\}f(x-z)\,dz\right| \\
&\le |h_1h_2^{-1} - 1|\sup f + C_1h_2^{-1}|h_1^{-1} - h_2^{-1}|^s\int_{x-a}^x |z|^s f(x-z)\,dz.
\end{aligned}$$

Therefore, if p is sufficiently large, $|\mu(x|h_1)\mu(x|h_2)^{-1} - 1| \le Cn^{-1}$ uniformly in $0 < x < a$ and $(h_1, h_2) \in \mathscr{H}$. It follows that for large p, $|S_1(h_1) - S_1(h_2)| \le Cn^{-1}$ uniformly in samples $X_1, \ldots, X_n$ and pairs $(h_1, h_2) \in \mathscr{H}$. Therefore, if $h_1, \ldots, h_{m+1}$ represent lattice points spaced n^{-p} apart and satisfying $n^{-\lambda-\delta} = h_1 < h_2 < \cdots < h_m \le n^{-\lambda+\delta} < h_{m+1}$, then

$$S \equiv \sup_{n^{-\lambda-\delta} \le h \le n^{-\lambda+\delta}} J(n,h)^{-1}|S_1(h)| \le \sup_{1 \le i \le m} J(n,h_i)^{-1}|S_1(h_i)| + C_n^{-\lambda+\delta}.$$

Consequently, by Markov's inequality,

$$P\{S > \eta + Cn^{-\lambda+\delta}\} \le \eta^{-2t} m \sup_{1 \le i \le m} J(n, h_i)^{-2t} E\{S_1(h_i)^{2t}\}$$

for any $\eta > 0$ and integer $t \ge 1$. Take η arbitrary but fixed and use Lemma 2.8 to show that if t is sufficiently large, then the right-hand side converges to zero, proving (2.31).

Proof of (2.32). Let $\lambda > 0$, $\mathscr{S}_1 \equiv (n^{-\lambda}, a - n^{-\lambda})$, $\mathscr{S}_2 \equiv (0, a) \setminus \mathscr{S}_1$,

$$\Delta_i \equiv \{\hat{f}_i(X_i|h) - \xi(X_i|h)\}/\xi(X_i|h), \qquad S_{21} \equiv n^{-1} \sum_{i=1}^{n} (\Delta_i - E\Delta_i),$$

$$S_{22} \equiv n^{-1} \sum_{i=1}^{n} \left[E(\Delta_i^2|X_i) I(X_i \in \mathscr{S}_1) - E\{\Delta_i^2 I(X_i \in \mathscr{S}_1)\}\right],$$

$$S_{23} \equiv n^{-1} \sum_{i=1}^{n} \{\Delta_i^2 - E(\Delta_i^2|X_i)\} I(X_i \in \mathscr{S}_1),$$

$$S_{26} \equiv n^{-1} \sum_{i=1}^{n} \left[\log(1 + \Delta_i) I(\Delta_i < -\tfrac{1}{2}) - E\{\log(1 + \Delta_i) I(\Delta_i < -\tfrac{1}{2})\}\right].$$

Then $S_2 = S_{21} - \frac{1}{2}S_{22} - \frac{1}{2}S_{23} + S_{24} + S_{25} + S_{26}$, where

$$|S_{24}| \le Cn^{-1} \sum_{i=1}^{n} \{|\Delta_i|^3 + E(|\Delta_i|^3|X_i)\} I(X_i \in \mathscr{S}_1) + CE\{|\Delta_1|^3 I(X_1 \in \mathscr{S}_1)\},$$

$$|S_{25}| \le C_n^{-1} \sum_{i=1}^{n} \{\Delta_i^2 + E(\Delta_i^2|X_i)\} I(X_i \in \mathscr{S}_2) + CE\{\Delta_1^2 I(X_1 \in \mathscr{S}_2)\}.$$

These results, together with the method in [17] based on a lattice argument and the Kolmós–Major–Tusnády [22] approximation, give after some tedious analysis, for sufficiently small δ,

$$\sup_{n^{-\lambda-\delta} \le h \le n^{-\lambda+\delta}} J(n, h)^{-1} \sum_{i=1}^{5} |S_{2i}| \to_p 0.$$

It remains to treat the term S_{26}. This has two facets, describing behaviour in the lower and upper tails, respectively. We consider only the lower tail, assume $\kappa > 1 + \alpha_1^{-1}$ and show how to prove that for small $\delta > 0$,

$$\sup_{n^{-\lambda-\delta} \le h \le n^{-\lambda+\delta}} J(n, h)^{-1} |n^{-1} \sum_{i=1}^{n} \left[\log(1 + \Delta_i) I(X_i \le \tfrac{1}{2}a, \Delta_i < -\tfrac{1}{2}) - E\{\log(1 + \Delta_i) I(X_i \le \tfrac{1}{2}a, \Delta_i < -\tfrac{1}{2})\}\right] + n^{-1-\kappa/(\alpha_1+1)} h^{-\kappa} A_1 W_{n1}| \to_p 0.$$

Using the definition (2.2) of K, deduce that $|\log(1 + \Delta_i) + A_1(T_i/h)^\kappa| \le C \log n$ uniformly in $n^{-\lambda-\delta} \le h \le n^{-\lambda+\delta}$. Therefore, the desired result follows if for

some $\eta > 0$,

$$\eta^{\eta} \sup_{n^{-\lambda-\delta} \le h \le n^{-\lambda+\delta}} h \sum_{i=1}^{n} I\big(\Delta_i < -\tfrac{1}{2}\big) \to_p 0, \tag{2.34}$$

$$n^{\kappa/(\alpha_1+1)} \sup_{n^{-\lambda-\delta} \le h \le n^{-\lambda+\delta}} \sum_{i=1}^{n} T_i^{\kappa} I\big(X_i \le \tfrac{1}{2}a, \Delta_i > -\tfrac{1}{2}\big) \to_p 0. \tag{2.35}$$

These formulas are established in Lemmas 2.9 and 2.10.

LEMMA 2.9. *Under the prescribed conditions on f and K and for sufficiently small δ and η, (2.34) holds.*

PROOF. Choose $\delta < \min(\lambda, 1-\lambda)$ so small that for some $\beta_1, \beta_2 > 0$,

$$(\alpha_i + 1)\beta_i > 1 - \lambda + \delta \quad \text{and} \quad \alpha_i\beta_i < 1 - \lambda - \delta \quad \text{for } i = 1, 2. \tag{2.36}$$

The event $\Delta_i < -\frac{1}{2}$ is equivalent to $\hat{f}_i(X_i|h) < \frac{1}{2}\{\mu(X_i|h) + (nh)^{-1}\}$. Using Hölder continuity of K, we deduce that for some $p > 0$ and all large n,

$$\big|\hat{f}_i(x|h_1) - \hat{f}_i(x|h_2)\big| + \big|\mu(x|h_1) - \mu(x|h_2)\big| \le \tfrac{1}{2}(nh_2)^{-2} \tag{2.37}$$

uniformly in $1 \le i \le n$, samples $X_1, \ldots, X_n$, $-\infty < x < \infty$, $n^{-\lambda-\delta} \le h_1 \le h_2 \le n^{-\lambda+\delta}$ and $h_2 - h_1 \le n^{-p}$. Suppose $h_1, \ldots, h_{m+1}$ represent lattice points spaced n^{-p} apart and satisfying $n^{-\lambda-\delta} = h_1 < h_2 < \cdots < h_m \le n^{-\lambda+\delta} \le h_{m+1}$. Given $h \in [n^{-\lambda-\delta}, n^{-\lambda+\delta}]$, let $j = j(h)$ be that index out of $1, \ldots, m$ which minimises $|h - h_j|$. Then $|h - h_{j(h)}| \le n^{-p}$ and so by (2.37), the event $\Delta_i < -\frac{1}{2}$ implies $\hat{f}_i(X_i|h_{j(h)}) \le \frac{1}{2}\mu(X_i|h_{j(h)})$. Consequently, with β_1 and β_2 as in (2.36),

$$\begin{aligned}
\sup_{n^{-\lambda-\delta} \le h \le n^{-\lambda+\delta}} \sum_{i=1}^{n} I\big(\Delta_i < -\tfrac{1}{2}\big) &\le \max_{1 \le j \le m} \sum_{i=1}^{n} I\big\{\hat{f}_i(X_i|h_j) \le \tfrac{1}{2}\mu(X_i|h_j)\big\} \\
&\le \sum_{i=1}^{n} I\big(X_i \le n^{-\beta_1} \text{ or } X_i > a - n^{-\beta_2}\big) \\
&\quad + \sum_{i=1}^{n} \sum_{j=1}^{m} I\big\{\hat{f}_i(X_i|h_j) \le \tfrac{1}{2}\mu(X_i|h_j), \\
&\qquad\qquad n^{-\beta_1} \le X_i \le a - n^{-\beta_2}\big\}.
\end{aligned} \tag{2.38}$$

The mean of the first series on the right-hand side equals

$$n\left(\int_0^{n^{-\beta_1}} + \int_{a-n^{-\beta_2}}^{\infty}\right) f(x)\,dx \le Cn\big(n^{-(\alpha_1+1)\beta_1} + n^{-(\alpha_2+1)\beta_2}\big),$$

while the mean of the second series equals

$$\begin{aligned}
\gamma &\equiv n \sum_{j=1}^{m} \int_{n^{-\beta_1}}^{a-n^{-\beta_2}} P\big\{\hat{f}_1(x|h_j) \le \tfrac{1}{2}\mu(x|h_j)\big\} f(x)\,dx \\
&\le C_1 n \sum_{j=1}^{m} \int_{n^{-\beta_1}}^{a-n^{-\beta_2}} \exp\big[-C_2 n h_j \min\{x^{\alpha_1}, (a-x)^{\alpha_2}\}\big]\,dx,
\end{aligned}$$

using the argument leading to Lemma 2.7. Since $m = O(n^p)$ and each $h_j \geq n^{-\lambda-\delta}$, then it follows from (2.36) that $\gamma = O(n^{-\kappa})$ for all $\kappa > 0$. Combining these estimates we see that the mean of the left-hand side of (2.38) equals

$$O\left(\sum_{i=1}^{2} n^{1-(\alpha_i+1)\beta_i}\right) = O(n^{\lambda-\delta-2\eta})$$

for some $\eta > 0$. Result (2.34) follows from this estimate and Markov's inequality. □

LEMMA 2.10. *Assume* $K(z) \leq A_2 \exp(-A_1|z|^\kappa)$ *for* $-\infty < z < \infty$, *where* $\kappa > 1 + \alpha_1^{-1}$, *and that* $\delta < \min(\lambda, 1-\lambda)$ *is so small that* $(\alpha_1+1)\lambda - 1 - 2\delta(\alpha_1+1)^2 > 0$. *Then* (2.35) *holds under the prescribed conditions on* f.

PROOF. Since $\mu(x|h) \geq C(x+h)^{\alpha_1}$ for $0 < x \leq \frac{1}{2}a$, then for $nh \geq 1$ the event $\{X_i \leq \frac{1}{2}a, \Delta_i > -\frac{1}{2}\}$ implies

$$\sum_{j \neq i} K\{(X_i - X_j)/h\} > C_1(n-1)h \cdot h^{\alpha_1} \geq C_2 n^{-\alpha_1},$$

which in turn implies $(|X_i - X_j|/h)^\kappa \leq C_3 \log n$ for some $j \neq i$. Consequently,

$$\{X_i \leq \tfrac{1}{2}a, \Delta_i > -\tfrac{1}{2}\} \subseteq \{X_i \leq \tfrac{1}{2}a\} \cap \left[\bigcup_{1 \leq i \leq n,\, j \neq i} \{|X_i - X_j| \leq Cn^{-\lambda+2\delta}\}\right],$$

from which it follows that the left-hand side of (2.35) is dominated by $n^{1+\kappa/(\alpha_1+1)}U$, where

$$U \equiv n^{-1} \sum_{i=1}^{n} T_i^\kappa I\big(|X_i - X_j| \leq Cn^{-\lambda+2\delta}, \text{ some } j \neq i;\ X_i \leq \tfrac{1}{2}a\big).$$

For any $0 < x_0 \leq \frac{1}{2}a$, the mean of U is dominated by

$$\int_0^{x_0} \{E(T_1^{2\kappa}|X_1 = x) nP(|x - X| \leq Cn^{-\lambda+2\delta})\}^{1/2} f(x)\,dx + \int_{x_0}^{a/2} E(T_1^\kappa|X_1 = x) f(x)\,dx.$$

It may be shown after a little algebra that $P(|x - X| \leq Cn^{-\lambda+2\delta}) \leq C_1 n^{2(\alpha_1+1)\delta}(n^{-\lambda}x^{\alpha_1} + n^{-(\alpha_1+1)\lambda})$ and it follows from Lemma 2.6 that $E(T_1^\beta|X = x) \leq C(\beta)\tau(x)^\beta$ $(\beta > 0)$, where $\tau(x)$ is defined just prior to Lemma 2.6. Combining these estimates and taking $x_0 \equiv (C_1 n^{1-\lambda+4(\alpha_1+1)\delta})^{-1/\alpha_1}$, we may prove that $E(U) = o(n^{-1-\kappa/(\alpha_1+1)})$. The lemma now follows via Markov's inequality. □

PROOF OF THEOREM 2.4. Let $X_{(1)} \leq \cdots \leq X_{(n)}$ denote the order statistics of the sample $X_1, \ldots, X_n$. Rényi's representation ([8], page 21) may be used to prove that for any fixed $m \geq 1$, $n^{1/(\alpha_1+1)}(X_{(1)}, \ldots, X_{(m)}) \to (V_{11}, \ldots, V_{m1})$ in distribution, where we take $\alpha = \alpha_1$ and $d = \{c_1^{-1}(\alpha_1+1)\}^{1/(\alpha_1+1)}$ in the definition of V_{i1}, $i \geq 1$. Observe that

$$\sum_{i=1}^{n} T_i^\kappa I(X_i \leq \tfrac{1}{2}a) = \sum_{i=1}^{n} \min\{(X_{(i)} - X_{(i-1)})^\kappa, (X_{(i+1)} - X_{(i)})^\kappa\} I(X_{(i)} \leq \tfrac{1}{2}a),$$

where we define $X_{(0)} = -\infty$ and $X_{(n+1)} = +\infty$, and that

$$n^{\kappa/(\alpha_1+1)} \sum_{i=1}^{m} \min\left\{\left(X_{(i)} - X_{(i-1)}\right)^{\kappa}, \left(X_{(i+1)} - X_{(i)}\right)^{\kappa}\right\} I\left(X_{(i)} \le \tfrac{1}{2}a\right)$$
$$\to \sum_{i=1}^{m} \min(V_{i2}, V_{i-1,2})$$

in distribution as $n \to \infty$. Techniques used to establish Lemma 2.6 may be employed to prove that

$$\lim_{m\to\infty} \limsup_{n\to\infty} n^{\kappa/(\alpha_1+1)} \sum_{i=m}^{\infty} E\left\{\left(X_{(i)} - X_{(i-1)}\right)^{\kappa} I\left(X_{(i)} \le \tfrac{1}{2}a\right)\right\} = 0.$$

Theorem 2.4 follows from these two results. □

The proof of Theorem 2.5 is similar to that of Theorem 2.3 and so will not be given.

3. Densities with regularly varying tails. In the previous section we showed how choice of kernel K can profoundly affect properties of Kullback–Leibler loss and likelihood cross-validation. Selection of K has the same influence in the case of an f with regularly varying tails. However, rather than dwell on details we shall summarise the main features in Section 3.1 and then concentrate on the case where K is chosen to minimise order of magnitude of Kullback–Leibler loss.

Recall that the variance and bias components, V and B, of Kullback–Leibler loss were defined in (1.5).

3.1. *Properties of Kullback–Leibler loss and likelihood cross-validation.* Assume the following conditions on f:

(3.1) f is bounded away from zero and infinity on $(-\lambda, \lambda)$ for each $\lambda > 0$, and $f(x) \sim c_1 x^{-\alpha_1}$ and $f(-x) \sim c_2 x^{-\alpha_2}$ as $x \to +\infty$, where $c_1, c_2 > 0$ and $\alpha_1, \alpha_2 > 1$.

Suppose the kernel K satisfies (2.2). The argument in Section 1.2 may be reworked to show that in the present circumstance, expected Kullback–Leibler loss $l_n(h)$ is finite if and only if $\kappa < \min(\alpha_1, \alpha_2) - 1$. Assuming the latter condition, the main-effect term in an expansion of the variance component V is of order $(nh)^{-1+(1/\alpha_1)} + (nh)^{-1+(1/\alpha_2)}$ and the tail-effect terms are of orders $n^{-1+\kappa/(\alpha_1-1)}h^{-\kappa}$ and $n^{-1+\kappa/(\alpha_2-1)}h^{-\kappa}$. [To be precise, the order $n^{-1+\kappa/(\alpha_i-1)}h^{-\kappa}$ of the ith tail-effect term should be increased by a logarithmic factor in the special case $\kappa = 1 - (1/\alpha_i)$ for $i = 1$ or 2.] Notice that

$$n^{-1+\kappa/(\alpha_i-1)}h^{-\kappa} = (nh)^{-1+(1/\alpha_i)}\left(nh^{-(\alpha_i-1)}\right)^{(\alpha_i\kappa-\alpha_i+1)/\alpha_i(\alpha_i-1)}$$

and so the tail-effect terms are negligible if and only if $\alpha_i\kappa - \alpha_i + 1 < 0$ for each i; that is, if and only if $\kappa < 1 - \max(\alpha_1^{-1}, \alpha_2^{-1})$. This condition excludes both standard normal and double exponential kernels and is strictly stronger than $\kappa < \min(\alpha_1, \alpha_2) - 1$.

More generally, it may be shown that a kernel K satisfying $K(z) \ge A_2\exp(-A_1|z|^{\kappa})$, $-\infty < z < \infty$, for positive constants A_1 and A_2 and a positive

constant $\kappa < 1 - \max(\alpha_1^{-1}, \alpha_2^{-1})$ will result in a variance component V containing no significant tail-effect terms. One such kernel is

$$K_0(z) \equiv \{(8\pi e)^{1/2}\Phi(1)\}^{-1}\exp\left[-\tfrac{1}{2}\{\log(1 + |z|)\}^2\right], \qquad -\infty < z < \infty,$$

where Φ denotes the standard normal distribution function and $\{(8\pi e)^{1/2}\Phi(1)\}^{-1} \doteq 0.1438$. This has the property that for all $\kappa > 0$, $K_0(z) \geq C_1\exp(-|z|^\kappa)$ and $K_0(z) \leq C_2(1 + |z|)^{-\kappa}$, $-\infty < z < \infty$, where C_1 and C_2 depend only on κ.

Properties of the bias component B are simpler than those of V. Assuming the kernel K to be symmetric and the density f to be twice differentiable, the bias component B is asymptotic to a constant multiple of h^4, just as in the case of squared-error loss.

If tail-effect terms dominate in the formula for V then likelihood cross-validation does not asymptotically minimise expected Kullback–Leibler loss, and minimisation of expected loss is not asymptotically equivalent to minimisation of "raw" loss, as discussed in Sections 2.4 and 2.5 for the case of compactly supported densities. Both these negative findings are reversed if tail-effect terms are insignificant in V.

In combination, these properties lend emphasis to the importance of correctly choosing the kernel function so as to reduce tail-effect terms. Throughout the remainder of this section we shall treat only those kernels, such as K_0 defined previously, for which tail-effect terms in V are negligible. Sections 3.2–3.4 treat separately the variance component V, the bias component B and likelihood cross-validation. Proofs are given in Section 3.5. Recall that V and B were defined in (1.5) and that expected loss is $l_n(h) = V + B$.

3.2. *Variance component V.* Assume f satisfies condition (3.1) and K satisfies:

K is symmetric about the origin and nonincreasing on $[0, \infty)$;

K integrates to unity;

$$\int_{-\infty}^{\infty} |z|^{\max(\alpha_1, \alpha_2)}K(z)\,dz < \infty; \tag{3.2}$$

and for positive constants A_1, A_2 and κ with $\kappa < 1 - \max(\alpha_1^{-1}, \alpha_2^{-1})$,

$$K(z) \geq A_2\exp(-A_1|z|^\kappa), \qquad -\infty < z < \infty.$$

The kernel K_0 defined in (1.7) satisfies (3.2) for all $\alpha_1, \alpha_2 > 1$.

Next we define the coefficient of the main-effect term in an expansion of V. A portion of the proof of Theorem 3.2 consists in showing that for each fixed $v > 0$ and for $(\alpha, c) = (\alpha_i, c_i)$ $(i = 1 \text{ or } 2)$, the random variable

$$\sum_{j=1}^{n} K\left[\left\{(nh)^{1/\alpha}v - X_j\right\}/h\right]$$

has a proper limiting distribution with characteristic function

$$\zeta(t) = \zeta(t|v, \alpha, c) \equiv \exp\left[-2cv^{-\alpha}\int_0^\infty \{1 - e^{itK(z)}\}\,dz\right], \tag{3.3}$$

$$-\infty < t < \infty.$$

A random variable $Z(v) = Z(v|\alpha, c)$ having this characteristic function has a continuous distribution with support confined to the positive half-line and has mean $E\{Z(v)\} = cv^{-\alpha}$. Define

$$D(\alpha, c) \equiv \int_0^\infty E\{Z(v)\} E\left[\log\{EZ(v)/Z(v)\}\right] dv$$

for $\alpha > 1$ and $c > 0$. The following proposition declares that this integral is well-defined.

PROPOSITION 3.1. *Assume* $c_1, c_2 > 0$, $\alpha_1, \alpha_2 > 1$ *and* K *satisfies* (3.2). *Then the infinite integral defining* $D(\alpha, c)$ *converges absolutely when* $(\alpha, c) = (\alpha_i, c_i)$ *for* $i = 1$ *and* 2 *and* $0 < D(\alpha_i, c_i) < \infty$.

THEOREM 3.2. *Assume* (3.1) *and* (3.2) *and that* $h = h(n) \to 0$ *and* $nh \to \infty$. *Then*

$$V = \sum_{i=1}^{2} (nh)^{-1+(1/\alpha_i)} D(\alpha_i, c_i) + o\left\{ \sum_{i=1}^{2} (nh)^{-1+(1/\alpha_i)} \right\} \tag{3.4}$$

as $n \to \infty$.

REMARK 3.1. Expansion (3.4) is the analogue of (2.8) in the case of densities with regularly varying tails.

3.3. *Bias component, B.* Assume the following condition on f:

$$\begin{aligned}
&f'' \text{ exists and is bounded and almost everywhere continuous} \\
&\text{on } (-\infty, \infty); \\
&f \text{ is bounded away from zero on compact intervals;} \\
&\text{and for constants} \\
&\alpha_1, \alpha_2 > 1, \qquad C_1 > 0 \quad \text{and} \quad C_2 < \infty, \\
&C_1 x^{-\alpha_1} \le f(x), \qquad |f''(x)| \le C_2 x^{-\alpha_1 - 2}, \\
&C_1 x^{-\alpha_2} \le f(-x) \quad \text{and} \quad |f''(-x)| \le C_2 x^{-\alpha_2 - 2} \quad \text{for } x > 1.
\end{aligned} \tag{3.5}$$

Assume the following condition on K:

$$\begin{aligned}
&K \ge 0, \qquad \int_{-\infty}^{\infty} |z|^{\max(\alpha_1, \alpha_2)+2} K(z)\, dz < \infty, \\
&\int_{-\infty}^{\infty} K(z)\, dz = 1 \quad \text{and} \quad \int_{-\infty}^{\infty} zK(z)\, dz = 0.
\end{aligned} \tag{3.6}$$

Define

$$E_0 \equiv \tfrac{1}{8}\left\{ \int_{-\infty}^{\infty} z^2 K(z)\, dz \right\}^2 \int_{-\infty}^{\infty} \{f''(x)\}^2 \{f(x)\}^{-1}\, dx.$$

THEOREM 3.3. *Assume conditions* (3.5) *and* (3.6). *Then* $B = h^4 E_0 + o(h^4)$ *as* $h \to 0$.

Theorems 3.2 and 3.3 together give us an expansion of expected Kullback–Leibler loss,

$$l_n(h) = V + B = \sum_{i=1}^{2} (nh)^{-1+(\alpha/\alpha_i)} D(\alpha_i, c_i) + h^4 E_0 + o\left\{ \sum_{i=1}^{2} (nh)^{-1+(1/\alpha_i)} + h^4 \right\}$$

as $h \to 0$ and $nh \to \infty$. Therefore, the window h_0 which minimises $l_n(h)$ satisfies $h_0 \sim \text{const.}\, n^{-(\alpha_i - 1)/(5\alpha_i - 1)}$, where $\alpha_i \equiv \min(\alpha_1, \alpha_2)$. For example, $h_0 \sim \text{const.}\, n^{-1/9}$ in the case of the Cauchy density and so the window which is asymptotically optimal from the point of view of minimising expected Kullback–Leibler loss is of a much larger order of magnitude than that which minimises expected squared-error loss.

3.4. *Likelihood cross-validation.* Our choice of kernel in Section 3.1 ensures that the tail-effect terms in V are negligible and, hence, that likelihood cross-validation asymptotically minimises Kullback–Leibler loss. To outline the relevant theory, assume condition (3.1) on f, condition (3.2) on K and that K satisfies

$$\int_{-\infty}^{\infty} |z|^{3+\max(\alpha_1, \alpha_2)} |\, dK(z)| < \infty,$$

$$K(u) \le C(1 + u)^{-(\alpha_j + 2)}$$

and

$$|K(u) - K(v)| \le C|u - v|^s \left\{ (1 + u)^{-(\alpha_j+2)} + (1 + v)^{-(\alpha_j+2)} \right\}$$

for some $0 < s \le 1$, $\alpha_j \equiv \max(\alpha_1, \alpha_2)$ and all $0 \le u < v < \infty$. Then for arbitrarily small $\varepsilon > 0$,

$$\sup_{n^{-1+\varepsilon} \le h \le n^{-\varepsilon}} \left\{ \sum_{i=1}^{2} (nh)^{-1+(1/\alpha_i)} + B(h) \right\}^{-1} \left[|CV(h) - E\{CV(h)\}| + |L_n(h) - l_n(h)| \right] \to 0$$

in probability as $n \to \infty$, where $L_n(h)$ denotes raw Kullback–Leibler loss [defined in (2.20)]. Therefore, if $\hat{h}_0$ and $\tilde{h}_0$ maximise CV and minimise L_n, respectively, within the range $[n^{-1+\varepsilon}, n^{-\varepsilon}]$, then $L_n(\hat{h}_0)/L_n(\tilde{h}_0) \to 1$ in probability. In other words, maximising CV is asymptotically equivalent to minimising L_n.

3.5. *Proofs.*

PROOFS OF PROPOSITION 3.1 AND THEOREM 3.2. Let $0 < r < 1$ and define

$$\begin{aligned} V_1 &= V_{11} + V_{12} + V_{13} \\ &\equiv \left(\int_0^{r(nh)^{1/\alpha_1}} + \int_{r(nh)^{1/\alpha_1}}^{r^{-1}(nh)^{1/\alpha_1}} + \int_{r^{-1}(nh)^{1/\alpha_1}}^{\infty} \right) f E\left\{ \log\left(E\hat{f}/\hat{f} \right) \right\} dx. \end{aligned} \tag{3.7}$$

We shall show that $V_1 \sim (nh)^{-1+(1/\alpha_1)} D(\alpha_1, c_1)$.

Define $U = U(x) \equiv (\hat{f} - E\hat{f})/E\hat{f}$, and to bound V_{11}, note that for any $u > -1$,

$$|\log(1+u) - u| \le C\{u^2 - \log(1+u)I(u < -\tfrac{1}{2})\}. \tag{3.8}$$

Both $E\hat{f}$ and $nh \operatorname{var} \hat{f}$ lie within the range $[C_1(1+x)^{-\alpha_1}, C_2(1+x)^{-\alpha_1}]$ for all $x > 0$ and so

$$\int_0^{r(nh)^{1/\alpha_1}} E(U^2)\,dx = \int_0^{r(nh)^{1/\alpha_1}} \operatorname{var}(\hat{f})(E\hat{f})^{-2}\,dx \le Cr(nh)^{-1+(1/\alpha_1)}.$$

The argument leading to Lemma 2.7 gives

$$P(U < -\tfrac{1}{2}) \le \exp\{-C_1 nh(1+x)^{-\alpha_1}\}, \qquad x > 0,$$

and so for any $0 < \varepsilon < 1$ and $x > 0$,

$$\begin{aligned} &E\{|\log(1+U)|I(U < -\tfrac{1}{2})\} \\ &\quad \le \left[E\{|\log(1+U)|^{1+\varepsilon} I(U < -\tfrac{1}{2})\}\right]^{1/(1+\varepsilon)} P(U < -\tfrac{1}{2})^{\varepsilon/(1+\varepsilon)} \\ &\quad \le C_2 \exp\{-\varepsilon C_1 nh(1+x)^{-\alpha_1}\}\left[h^{-\kappa'}E\Big(\min_{1\le i\le n}|x - X_i|^{\kappa'}\Big)\right. \\ &\qquad\qquad \left. + \left|\log\{nh(1+x)^{-\alpha_1}\}\right|^{1+\varepsilon} + 1\right], \end{aligned}$$

where $\kappa' = \kappa(1+\varepsilon)$. [Note (1.4).] Choose ε so small that $\kappa' < 1 - \max(\alpha_1^{-1}, \alpha_2^{-1})$ and use the argument leading to Lemma 2.6 to obtain the bound

$$h^{-\kappa'}E\Big(\min_{1\le i\le n}|x - X_i|^{\kappa'}\Big) \le C\{(nh)^{-\kappa'}(1+x)^{\alpha_1\kappa'} + 1\}$$

for $0 < x \le (nh)^{1/\alpha_1}$. Taking $u = U$ in (3.8) and combining the estimates from there down, we conclude that

$$\lim_{r\to 0} \limsup_{n\to\infty} (nh)^{1-(1/\alpha_1)} E(V_{1i}) = 0 \tag{3.9}$$

for $i = 1$. A similar argument establishes (3.9) for $i = 3$.

It remains to estimate V_{12}. Observe that $P[K\{(x - X)/h\} > y] = P\{|x - X| < hK^{-1}(y)\}$, where $K^{-1}(y) \equiv \inf\{z > 0\colon K(z) \ge y\}$. Take $x = (nh)^{1/\alpha_1}v$, where $v > 0$. Then for fixed v and y,

$$P[K\{(x - X)/h\} > y] \sim 2hK^{-1}(y)f(x) \sim n^{-1}2c_1 v^{-\alpha_1}K^{-1}(y)$$

as $n \to \infty$. Also, the left-hand side is dominated by a constant multiple of the right-hand side, uniformly in $0 < y < K(0)$, for fixed v. Therefore,

$$\begin{aligned} \Psi_n(t) &\equiv n\{1 - E(\exp[itK\{(x - X)/h\}])\} \\ &= -n\int_0^{K(0)} P[K\{(x - X)/h\} > y]\,ite^{ity}\,dy \\ &\to -2c_1 v^{-\alpha_1}\int_0^{K(0)} K^{-1}(y)\,ite^{ity}\,dy \\ &= 2c_1 v^{-\alpha_1}\int_0^{\infty} [1 - \exp\{itK(z)\}]\,dz, \end{aligned}$$

whence

$$E\left(\exp\left[it\sum_{j=1}^{n}K\{(x-X_j)/h\}\right]\right)=\{1-n^{-1}\psi_n(t)\}^n\to\zeta(t|v,\alpha_1,c_1)$$

as $n\to\infty$. This establishes that $Y(x)\equiv\Sigma_j K\{(x-X_j)/h\}$ converges weakly to $Z(v)$ as $n\to\infty$. If $\varepsilon>0$ is so small that $\kappa(1+\varepsilon)<1-\max(\alpha_1^{-1},\alpha_2^{-1})$, then $E\{|\log Y(x)|^{1+\varepsilon}\}$ is bounded uniformly in n and $r(nh)^{1/\alpha_1}\le x\le r^{-1}(nh)^{1/\alpha_1}$. Therefore, by weak convergence and dominated convergence,

$$(nh)^{1-(1/\alpha_1)}\int_{\tau(nh)^{1/\alpha_1}}^{r^{-1}(nh)^{1/\alpha_1}}f(x)E\{\log Y(x)\}\,dx\to c_1\int_r^{r^{-1}}v^{-\alpha_1}E\{\log Z(v)\}\,dv$$

and similarly,

$$(nh)^{1-(1/\alpha_1)}V_{12}\to\int_r^{r^{-1}}E\{Z(v)\}E\left[\log\{EZ(v)/Z(v)\}\right]dv. \tag{3.10}$$

[Note that $E\{Z(v)\}=c_1v^{-\alpha_1}$.] A slight variant of the techniques used to establish (3.9) for $i=1$ and 3 shows that the integral on the right-hand side of (3.10) remains bounded as $r\to 0$. It follows by convexity that the integrand is positive. These observations, together with (3.7) and (3.9), show that $V_1\sim(nh)^{-1+(1/\alpha_1)}D(\alpha_1,c_1)$ and that Proposition 3.1 is true in the case $(\alpha,c)=(\alpha_1,c_1)$. □

Theorem 3.3 is relatively easy to prove, and so will not be derived here.

Acknowledgments. The presentation of results and of technical arguments has benefited from very helpful comments by a referee and an Associate Editor.

REFERENCES

[1] AITCHISON, J. and AITKEN, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika* **63** 413–420.

[2] BOWMAN, A. W. (1980). A note on consistency of the kernel method for the analysis of categorical data. *Biometrika* **67** 682–684.

[3] BOWMAN, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** 353–360.

[4] BOWMAN, A. W. (1985). A comparative study of some kernel-based non-parametric density estimators. *J. Statist. Comput. Simul.* **21** 313–327.

[5] BOWMAN, A. W., HALL, P. and TITTERINGTON, D. M. (1984). Cross-validation in nonparametric estimation of probabilities and probability densities. *Biometrika* **71** 341–351.

[6] BURKHOLDER, D. L. (1973). Distribution function inequalities for martingales. *Ann. Probab.* **1** 19–42.

[7] CHOW, Y. S., GEMAN, S. and WU, L. D. (1983). Consistent cross-validated density estimation. *Ann. Statist.* **11** 25–38.

[8] DAVID, H. A. (1980). *Order Statistics*, 2nd ed. Wiley, New York.

[9] DUIN, R. P. W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans. Comput.* **C25** 1175–1179.

[10] GREGORY, G. G. and SCHUSTER, E. F. (1979). Contributions to non-parametric maximum likelihood methods of density estimation. In *Computer Science and Statistics: Proc. 12th Symp. on the Interface* (J. F. Gentleman, ed.) 427–431. Univ. Waterloo, Ontario.

[11] HABBEMA, J. D. F. and HERMANS, J. (1977). Selection of variables in discriminant analysis by F-statistic and error rate. *Technometrics* **19** 487–493.

[12] HABBEMA, J. D. F., HERMANS, J. and REMME, J. (1978). Variable kernel estimation in discriminant analysis. In *Compstat 1978* (L. C. A. Corsten and J. Hermans, eds.) 178–185. Physica, Vienna.

[13] HABBEMA, J. D. F., HERMANS, J. and VAN DEN BROEK, K. (1974). A stepwise discriminant analysis program using density estimation. In *Compstat 1974* (G. Bruckmann, ed.) 101–110. Physica, Vienna.

[14] HALL, P. (1981). On nonparametric multivariate binary discrimination. *Biometrika* **68** 287–294.

[15] HALL, P. (1982). Cross-validation in density estimation. *Biometrika* **69** 383–390.

[16] HALL, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11** 1156–1174.

[17] HALL, P. (1985). Asymptotic theory of minimum integrated square error for multivariate density estimation. In *Proc. Sixth Internat. Symp. Multivariate Anal.* (P. R. Krishnaiah, ed.) 289–309 North-Holland, Amsterdam.

[18] HALL, P. (1985). On the estimation of probability densities using compactly supported kernels. *J. Multivariate Anal.* To appear.

[19] HALL, P. and MARRON, J. S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab. Theory Related Fields* **74** 567–581.

[20] HALL, P. and MARRON, J. S. (1987). On the amount of noise inherent in bandwidth selection for a kernel density estimator. *Ann. Statist.* **15** 163–181.

[21] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

[22] KOMLÓS, J., MAJOR, P. and TSUNÁDY, G. (1975). An approximation of partial sums of independent RV's and the sample DF. I. *Z. Wahrsch. verw. Gebiete* **32** 111–131.

[23] KULLBACK, S. (1959). *Information Theory and Statistics*. Wiley, New York.

[24] MARRON, J. S. (1985). An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *Ann. Statist.* **13** 1011–1023.

[25] RAATGEVER, J. W. and DUIN, R. P. W. (1978). On the variable kernel method for multivariate nonparametric density estimation. In *Compstat 1978* (L. C. A. Corsten and J. Hermans, eds.) 524–533. Physica, Vienna.

[26] RUDEMO, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65–78.

[27] SCHUSTER, E. F. and GREGORY, G. G. (1981). On the nonconsistency of maximum likelihood nonparametric density estimators. In *Computer Science and Statistics: Proc. 13th Symp. on the Interface* (W. F. Eddy, ed.) 295–298. Springer, New York.

[28] STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** 1285–1297.

[29] STONE, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *J. Roy. Statist. Soc. Ser. B* **36** 111–147.

[30] STONE, M. (1974). Cross-validation and multinomial prediction. *Biometrika* **61** 509–515.

[31] STONE, M. (1977). Asymptotics for and against cross-validation. *Biometrika* **64** 29–35.

[32] TITTERINGTON, D. M. (1977). Analysis of incomplete multivariate data by the kernel method. *Biometrika* **64** 455–460.

[33] TITTERINGTON, D. M. (1978). Contribution to discussion of paper by T. Leonard. *J. Roy. Statist. Soc. Ser. B* **40** 139–140.

[34] TITTERINGTON, D. M. (1980). A comparative study of kernel-based density estimates for categorical data. *Technometrics* **22** 259–268.

DEPARTMENT OF STATISTICS
AUSTRALIAN NATIONAL UNIVERSITY
GPO BOX 4
CANBERRA, ACT 2601
AUSTRALIA