Ann. Inst. Statist. Math.
Vol. 45, No. 1, 69–88 (1993)

ON THE ESTIMATION OF ENTROPY

PETER HALL[1] AND SALLY C. MORTON[2]

[1]*Centre for Mathematics and its Applications, Australian National University, G.P.O. Box 4, Canberra A.C.T. 2601, Australia and CSIRO Division of Mathematics and Statistics*
[2]*Statistical Research and Consulting Group, The RAND Corporation, 1700 Main Street, P.O. Box 2138, Santa Monica, CA 90407-2138, U.S.A.*

(Received May 7, 1991; revised April 14, 1992)

Abstract. Motivated by recent work of Joe (1989, *Ann. Inst. Statist. Math.*, **41**, 683–697), we introduce estimators of entropy and describe their properties. We study the effects of tail behaviour, distribution smoothness and dimensionality on convergence properties. In particular, we argue that root-n consistency of entropy estimation requires appropriate assumptions about each of these three features. Our estimators are different from Joe's, and may be computed without numerical integration, but it can be shown that the same interaction of tail behaviour, smoothness and dimensionality also determines the convergence rate of Joe's estimator. We study both histogram and kernel estimators of entropy, and in each case suggest empirical methods for choosing the smoothing parameter.

Key words and phrases: Convergence rates, density estimation, entropy, histogram estimator, kernel estimator, projection pursuit, root-n consistency.

1. Introduction

This paper was motivated by work of Joe (1989) on estimation of entropy. Our work has three main aims: elucidating the role played by Joe's key regularity condition (A); developing theory for a class of estimators whose construction does not involve numerical integration; and providing a concise account of the influence of dimensionality on convergence rate properties of entropy estimators. Our main results do not require Joe's (1989) condition (A), which asks that tail properties of the underlying distribution be ignorable. We show concisely how tail properties influence estimator behaviour, including convergence rates, for estimators based on both kernels and histograms. We point out that histogram estimators may be used to construct root-n consistent entropy estimators in $p = 1$ dimension, and that kernel estimators give root-n consistent entropy estimators in $p = 1, 2$ and 3 dimensions, but that neither type generally provides root-n consistent estimation beyond this range, unless (for example) the underlying distribution is compactly

supported, or is particularly smooth and bias-reduction techniques are employed. Joe (1989) develops theory for a root-n consistent estimator in the case $p = 4$, but he makes crucial use his condition (A). Even for $p = 1$, root-n consistency of our estimators or of that estimator suggested by Joe (1989) requires certain properties of the tails of the underlying distribution. Goldstein and Messer (1991) briefly mention the problem of entropy estimation, but like Joe they work under the assumption (A).

To further elucidate our results it is necessary to introduce a little notation. Let $X_1, X_2, \ldots, X_n$ denote a random sample drawn from a p-variate distribution with density f, and put $I = \int f \log f$, where the integral is assumed to converge absolutely. Then $-I$ denotes the entropy of the distribution determined by f. We consider estimation of I. Our estimators are motivated by the observation that $\bar{I} = n^{-1} \sum_{i=1}^{n} \log f(X_i)$ is unbiased for I, and is root-n consistent if $\int f(\log f)^2 < \infty$. Of course, since f is not known then $\bar{I}$ is not a practical estimator. However, if f may be estimated nonparametrically by $\hat{f}$, say, then

$$\hat{I}_1 = n^{-1} \sum_{i=1}^{n} \log \hat{f}(X_i) \tag{1.1}$$

might be an appropriate alternative to $\bar{I}$.

Since $\hat{f}$ would typically depend on each X_i then the expected value of $E(\hat{I}_1)$ might differ significantly from $\int f E(\log \hat{f})$. This observation motivates an alternative estimator,

$$\hat{I}_2 = n^{-1} \sum_{i=1}^{n} \log \hat{f}_i(X_i), \tag{1.2}$$

where $\hat{f}_i$ has the same form as $\hat{f}$ except that it is computed for the $(n-1)$-sample which excludes X_i. We develop a version of $\hat{I}_1$ when $\hat{f}$ is a histogram estimator, and a version of $\hat{I}_2$ when $\hat{f}$ is a kernel estimator. A version of $\hat{I}_1$ for kernel estimators is also discussed. In both cases we prove that, under appropriate regularity conditions, $\hat{I} = \bar{I} + o_p(n^{-1/2})$. Then, a central limit theorem and other properties of $\hat{I}$ follow immediately from their counterparts for $\bar{I}$.

The estimator $\hat{I}_2$ is sensitive to outliers, since the density estimator can be very close to zero when evaluated at outlying data values. This is one way of viewing the effects noted by Hall (1987). There it is shown that the adverse effects of tail behaviour, or equivalently of outliers, may be alleviated by using a kernel with heavier tails. Depending on their extent, outliers can be problematic when using entropy estimators in exploratory projection pursuit.

For both histogram and kernel methods, choice of smoothing parameter determines the performance of the entropy estimator. We suggest practical methods for smoothing parameter choice. For the histogram estimator of type (1.1) we propose that a penalty term be subtracted from $\hat{I}_1$, that the penalized version be maximized with respect to histogram binwidth, and that the resulting binwidth be used to compute $\hat{I}_1$. A version of this technique may also be developed for the kernel estimator of type (1.1). For the kernel estimator of type (1.2) we suggest

that $\hat{I}_2$ be maximized with respect to bandwidth, without regard for any penalty. The performance of each approach is assessed by both theoretical and numerical analyses. Section 2 treats the case of histogram estimators, and kernel estimators are studied in Section 3.

Our theoretical account is based on arguments in Hall (1987, 1989), which analyse empirical properties of Kullback-Leibler loss. We develop substantial generalizations of that work which include, in the case of histogram estimators, a study of $\hat{I}_1$ for a wide class of densities f having unbounded support.

Entropy estimators may be employed to effect a test for normality (see e.g. Vasicek (1986)) and to construct measures of "interestingness" in the context of projection pursuit (see e.g. Huber (1985), Friedman (1987), Jones and Sibson (1987), Hall (1989) and Morton (1989)). If one-dimensional projections are used then the first step of exploratory projection pursuit is often to choose that orientation θ_0 which maximizes $I(\theta) = \int f_\theta \log f_\theta$, where θ is a unit p-vector, f_θ denotes the density of the projected scalar random variable $\theta \cdot X$, and $u \cdot v$ denotes the dot (i.e. scalar) product of vectors u and v. The results in Sections 2 and 3 show that, under appropriate regularity conditions,

$$\hat{I}(\theta) = \bar{I}(\theta) + o_p\big(n^{-1/2}\big) \tag{1.3}$$

for each θ, where $\hat{I}(\theta)$ denotes the version of $\hat{I}$ computed from the univariate sample $\theta \cdot X_1, \ldots, \theta \cdot X_n$, and $\bar{I}(\theta) = n^{-1} \sum_{i=1}^n \log f_\theta(X_i)$. Result (1.3) may readily be proved to hold uniformly in unit p-vectors θ, by using Bernstein's inequality and noting that the class of all unit p-vectors is compact. Arguing thus we may show that if $\hat{\theta}$, $\bar{\theta}$ are chosen to maximize $\hat{I}(\theta)$, $\bar{I}(\theta)$ respectively, then $\hat{\theta} - \bar{\theta} = o_p\big(n^{-1/2}\big)$. A limit theorem for $\bar{\theta}$, of the form $n^{1/2}\big(\bar{\theta} - \theta_0\big) \to Z$ (where Z is a normal random variable with mean zero) in distribution, is readily established. It then follows immediately that $n^{1/2}\big(\hat{\theta} - \theta_0\big) \to Z$.

2. Histogram estimators

2.1 *Summary*

Subsection 2.2 proposes a histogram estimator, $\hat{I}$, of negative entropy, $I = \int f \log f$. Properties of the estimator in the p-dimensional case are outlined, and it is shown that the estimator can only be root-n consistent when $p = 1$ or 2. Furthermore, only in the case $p = 1$ can binwidth be chosen so that $\hat{I}$ is identical to the unbiased estimator $\bar{I} = n^{-1} \sum_{i=1}^n \log f\big(X_i\big)$ up to terms of smaller order than $n^{-1/2}$; when $p = 2$, any root-n consistent histogram estimator of I has significant bias, of size $n^{-1/2}$. (For reasons of economy, detailed proofs of some of these results are omitted; they are very similar to counterparts in the case of kernel estimators, treated in Section 3.)

Thus, only for $p = 1$ dimension is the histogram estimator particularly attractive. Subsections 2.3 *et seq* confine attention to that context. In particular, Subsection 2.3 suggests an empirical rule for bandwidth selection when $p = 1$, and describes its performance in the case of densities with regularly varying tails. The rule is based on maximizing a penalized version of $\hat{I}$, and is related to techniques

derived from Akaike's information criterion. Subsection 2.4 describes properties of $\hat{I}$, and of the empirical rule, when the underlying density has exponentially decreasing rather than regularly varying tails. A numerical study of the rule is presented in Subsection 2.5, and proofs are outlined in Subsection 2.6.

2.2 *Outline of general properties*

We first introduce notation, then we define a histogram estimator of the entropy of a p-variate density, and subsequently we describe its properties.

Let Z^p denote the set of all integer p-vectors $i = (i_1, \ldots, i_p)^T$, let $v = (v_1, \ldots, v_p)^T$ denote any fixed p-vector, and let

$$B_i = \left\{ x = (x_1, \ldots, x_p)^T : |x_j - (v_j + i_j h)| \le \frac{1}{2} h, 1 \le j \le p \right\}$$

represent the histogram bin centred at $v + ih$. Here h, a number which decreases to zero as sample size increases, represents the binwidth of the histogram. Write N_i for the number of data values which fall into bin B_i. Then for $x \in B_i$, $N_i/(nh^p)$ is an estimator of $f(x)$, and

$$\hat{I} = n^{-1} \sum_i N_i \log \{N_i/(nh^p)\} = n^{-1} \sum_i N_i \log N_i - \log (nh^p)$$

is an estimator of $I = \int f \log f$.

Let $\| \cdot \|$ denote the usual Euclidean metric in p-dimensional Euclidean space. It may be shown that if $f(x)$ has tails which decrease like a constant multiple of $\|x\|^{-\alpha}$ as $\|x\| \to \infty$, for example if $f(x) = c_1 (c_2 + \|x\|)^{-\alpha}$ for positive constants c_1 and c_2, then $\hat{I}$ admits an expansion of the form

$$\hat{I} = \bar{I} + a_1 (nh^p)^{-1+(p/\alpha)} - a_2 h^2 + o_p\{(nh^p)^{-1+(p/\alpha)} + h^2\}, \tag{2.1}$$

where a_1, a_2 are positive constants and $\bar{I} = n^{-1} \sum \log f(X_i)$. The term of size $(nh^p)^{-1+(p/\alpha)}$ in (2.1) comes from a "variance component" of $\hat{I}$, and the term of size h^2 comes from a "bias component". The constraint that f be integrable dictates that $\alpha > p$.

In the multivariate case, our assumption that the density's tails decrease like $\|x\|^{-\alpha}$ serves only to illustrate, in a general way, the manner in which tail weight affects convergence rate. We do not claim that this "symmetric" distribution might be particularly useful as a practical model for real data. However, when $p > 1$ it is not possible to give a simple description of the impact of tail weight under realistic models, because of the very large variety of ways in which tail weight may vary in different directions. Variants of result (2.1) are available for a wide range of multivariate densities, that decrease like $\|x\|^{-\alpha}$ in one or more directions but decrease more rapidly in other directions. The nature of the result does not change, but the power (p/α) does alter. There also exists an analogue of (2.1) in the case of a multivariate normal density, where the quantity $(nh^p)^{-1+(p/\alpha)}$ is replaced by $(nh^p)^{-1}$ multiplied by a logarithmic term.

However, in the case of $p = 1$ dimension, our model is amply justified by practical considerations. See for example Hill (1975) and Zipf (1965). We employ the model for general p in order to show that, at least in simple distributions and for the histogram estimator, optimal convergence rates may only be obtained when $p = 1$. Then we focus on the latter case, where the model may be strongly motivated.

If $\sigma^2 = \int f(\log f)^2 - I^2 < \infty$ then $\bar{I}$ is root-n consistent for I, and in fact $n^{1/2}(\bar{I} - I)$ is asymptotically normal $N(0, \sigma^2)$. The extent to which $\hat{I}$ achieves the same rate of convergence, and the same central limit theorem, is determined largely by the size of the difference $d(h) = a_1(nh^p)^{-1+(p/\alpha)} - a_2h^2$ in (2.1). In principle, h can be chosen so that $d(h) = 0$, i.e.

$$h = (a_1/a_2)^{\alpha/\{\alpha(p+2)-p^2\}} n^{-(\alpha-p)/\{\alpha(p+2)-p^2\}},$$

in which case the "remainder term" in (2.1) must be investigated. However, this is a somewhat impractical suggestion. First of all, it requires careful estimation of α, a_1 and a_2, which is far from straightforward, particularly when $p \geq 2$. Secondly, it does not indicate how we might deal with circumstances where the model $f(x) \sim \text{const.}\|x\|^{-\alpha}$ is violated.

The best we can realistically hope to achieve is that h is chosen so that "variance" and "bias" contributions are of the same order of magnitude; that is, $(nh^p)^{-1+(p/\alpha)}/h^2$ is bounded away from zero and infinity as $h \to 0$ and $n \to \infty$. Subsection 2.3 will describe an empirical rule for achieving this end when $p = 1$. Achieving this balance requires taking h to be of size n^{-a}, where

$$a = (\alpha - p)/\{\alpha(p+2) - p^2\}.$$

In this case, $d(h)$ is of size n^{-2a}. If $\hat{I}$ is to be asymptotically equivalent to $\bar{I}$, up to terms of smaller order than $n^{-1/2}$, then we require $2a > 1/2$, or equivalently $\alpha(p-2) < p(p-4)$. If $p = 1$ then this condition reduces to $\alpha > 3$, but if $p \geq 2$ then the condition does not admit any solutions α which satisfy the essential constraint $\alpha > p$. Thus, the mean squared error of $\hat{I}$ is greater than that of $\bar{I}$.

Thus, we conclude that the histogram method for estimating entropy is most effective in the case of $p = 1$ dimension. In other cases binwidth choice is a critical problem, and for $p \geq 2$ it is virtually impossible to achieve root-n consistency. We shall show in Section 3 that these difficulties may be largely circumvented by employing kernel-type estimators.

2.3 *The case $p = 1$: an empirical rule*

In the case $p = 1$ it may be deduced from Theorem 4.1 of Hall (1990) that (2.1) holds with

$$a_1 = 2b^{1/\alpha}D(\alpha), \qquad a_2 = (1/24)\int (f')^2 f^{-1}, \tag{2.2}$$

where b is the constant in the regularity condition (2.3) below, and

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

with $M(x)$ denoting a Poisson-distributed random variable with mean x. The following regularity condition is sufficient for (2.1) to hold uniformly in any collection $\mathcal{H}_n$ such that for some δ, $C > 0$, $n^{-1+\delta} \le h \le n^{-\delta}$ for each $h \in \mathcal{H}_n$ and $\#\mathcal{H}_n = O(n^C)$; see Section 4 of Hall (1990):

$$\begin{aligned} &f > 0 \text{ on } (-\infty, \infty),\ f' \text{ exists and is continuous on } (-\infty, \infty), \\ &\text{and for constants } b > 0 \text{ and } \alpha > 1,\ f'(x) \sim -b\alpha x^{-\alpha-1} \\ &\text{and } f'(-x) \sim -b\alpha x^{-\alpha-1} \text{ as } x \to \infty. \end{aligned} \tag{2.3}$$

In order to determine an appropriate bandwidth for the estimator $\hat{I}$ we suggest subtracting a penalty Q from $\hat{I}$, such that for a large class of densities,

$$\breve{I} \equiv \hat{I} - Q = \bar{I} - S(h) + o_p\{(nh)^{-1+(1/\alpha)} + h^2\}, \tag{2.4}$$

where

$$S(h) = a_3(nh)^{-1+(1/\alpha)} + a_2h^2 \tag{2.5}$$

and a_2, a_3 are *positive* constants. In view of this positivity, maximizing $\breve{I}$ is asymptotically equivalent to minimizing $S(h)$, and so to minimizing the distance between $\hat{I}$ and $\bar{I}$. This operation produces a bandwidth of size $n^{-(\alpha-1)/(3\alpha-1)}$, which we showed in Subsection 2.2 to be the optimal size.

We suggest taking

$$Q = n^{-1} \text{ (number of nonempty bins)}. \tag{2.6}$$

For this penalty function it is demonstrated in Hall ((1987), Section 4) that under (2.3), formula (2.4) holds with a_2 given by (2.2) and with

$$a_3 = 2b^{1/\alpha}\{\Gamma(1 - \alpha^{-1}) - D(\alpha)\}.$$

It may be shown numerically that $a_3 > 0$ for $\alpha > \alpha_0 \simeq 2.49$, which corresponds to a density with a finite absolute moment of order greater than 1.49. (For example, finite variance is sufficient.)

A question arises as to whether the penalized version of $\hat{I}$, i.e. $\breve{I}$, should be taken as the estimator of I, or whether $\hat{I}$ itself should be used. From one point of view the question is academic, since if the bandwidth is chosen to maximize $\breve{I}$, and (2.3) holds, then $\hat{I}$ and $\breve{I}$ are both first-order equivalent to $\bar{I}$: for $J = \hat{I}$ or $\breve{I}$,

$$J = \bar{I} + O_p\big(n^{-2(\alpha-1)/(3\alpha-1)}\big) = \bar{I} + o_p\big(n^{-1/2}\big),$$

provided only that $\alpha > 3$. These formula follow from (2.1), (2.4) and (2.5) (in the context of condition (2.3), $\alpha > 3$ is equivalent to finite variance). However, it is of practical as well as theoretical interest to minimize the second-order term, of size $n^{-2(\alpha-1)/(3\alpha-1)}$. Indeed, the simulation study outlined later in this section will show that for samples of moderate size, second-order effects can be significant. We claim that if the density f has sufficiently light tails then $\hat{I}$ is preferable to $\breve{I}$.

To appreciate why, observe that with $\delta_n = n^{-2(\alpha-1)/(3\alpha-1)}$, and binwidth chosen as suggested three paragraphs above (with Q given by (2.3)), the standard deviations of $\hat{I}$ and $\check{I}$ both equal $n^{-1/2}\sigma + o(\delta_n)$, and biases equal $v + o(\delta_n)$ and $v + w + o(\delta_n)$ respectively, where

$$v = a_1(nh)^{-1+(1/\alpha)} - a_2h^2, \qquad w = -2b^{1/\alpha}\Gamma(1-\alpha^{-1})(nh)^{-1+(1/\alpha)}.$$

If we regard first-order terms as being of size $n^{-1/2}$ and second-order terms as being of size $n^{-2(\alpha-1)/(3\alpha-1)}$, then we see that standard deviations are identical to first and second orders, but that while biases agree to first order they differ to second order. The second-order bias term is less for $\hat{I}$ than it is for $\check{I}$ if and only if $|v| < |v+w|$, or equivalently if and only if $-(w+2v) > 0$, i.e.

$$b^{1/\alpha}\{\Gamma(1-\alpha^{-1}) - 2D(\alpha)\}(nh)^{-1+(1/\alpha)} + a_2h^2 > 0. \tag{2.7}$$

It may be shown numerically that $\Gamma(1-\alpha^{-1}) - 2D(\alpha) > 0$ for all $\alpha > 7.55$, which corresponds to at least 6.55 absolute moments finite. Therefore, since $a_2 > 0$, we can expect (2.7) to hold for all sufficiently light-tailed distributions.

The "penalty method" for selecting h is appropriate in a wide variety of different cases, including those where the density f has exponentially decreasing, rather than regularly varying, tails. Rigorous analysis of that case requires a little additional theory, which is developed in the next subsection. Subsection 2.5 presents numerical examples which illustrate the performance of the penalty method.

2.4 *Theory for distributions whose densities are not regularly varying*

The empirical rule developed in Subsection 2.3 is for the case where $f(x) \sim b|x|^{-\alpha}$ as $|x| \to \infty$, for constants $b > 0$ and $\alpha > 1$. A slightly more general case, where $f(x) \sim b_1x^{-\alpha_1}$ and $f(-x) \sim b_2x^{-\alpha_2}$ as $x \to \infty$, may be treated by applying results in Hall ((1987), Section 4). And similar arguments may be used to develop formulae for the case where $f(x) \sim x^{-\alpha_1}L_1(x)$ and $f(x) \sim x^{-\alpha_2}L_2(x)$ as $x \to \infty$, where L_1 and L_2 are slowly varying functions. The work in the present subsection is aimed at developing theory applicable to the case of densities whose tails decrease exponentially quickly, at a faster rate than any order of regular variation; or which decrease in a manner which is neither regularly varying nor exponentially decreasing. We deal only with the case of $p = 1$ dimension.

Our first result concerns the case of distributions whose densities decrease like const. $\exp(-\text{const.}|x|^\alpha)$, for some $\alpha > 0$. Our regularity condition, replacing (2.3), is

$$\begin{aligned} &f > 0 \text{ on } (-\infty,\infty),\ f' \text{ exists and is continuous on } (-\infty,\infty), \text{and for} \\ &\text{constants } b_{11}, b_{12}, b_{21}, b_{22}, \alpha_1, \alpha_2 > 0 \text{ we have } f'(x) \sim (d/dx)b_{11} \\ &\exp(-b_{12}x^\alpha) \text{ and } f'(-x) \sim (d/dx)b_{21}\exp(-b_{22}x^\alpha) \text{ as } x \to \infty. \end{aligned} \tag{2.8}$$

Define a_2 as at (2.2). Let $\mathcal{H}_n$ denote a collection of real numbers h satisfying $n^{-1+\delta} \le h \le n^{-\delta}$ for each $h \in \mathcal{H}_n$ and $\#\mathcal{H}_n = O(n^C)$, where $\delta, C > 0$ are fixed constants.

THEOREM 2.1. *Assume condition* (2.8). *Then*

$$\hat{I} = \bar{I} + \frac{1}{2}\sum_{i=1}^{2} b_{i2}(nh)^{-1}(\log nh)^{1/\alpha_i} - a_2h^2 + o_p\left\{\sum_{i=1}^{2}(nh)^{-1}(\log nh)^{1/\alpha_i} + h^2\right\} \tag{2.9}$$

uniformly in $h \in \mathcal{H}_n$, *as* $n \to \infty$.

Proofs of Theorems 2.1 and 2.2 are outlined in Subsection 2.6.

The penalty method of bandwidth choice, introduced in Subsection 2.3 and discussed there in the context of densities with regularly varying tails, is also appropriate for the present case of exponentially decreasing densities. To appreciate the theory appropriate for this case we must first develop analogues of formulae (2.4) and (2.5); these are,

$$\breve{I} \equiv \hat{I} - Q = \bar{I} - S(h) + o_p\left\{\sum_{i=1}^{2}(nh)^{-1}(\log nh)^{1/\alpha_i} + h^2\right\}, \tag{2.10}$$

where

$$S(h) = \frac{1}{2}\sum_{i=1}^{2} b_{i2}(nh)^{-1}(\log nh)^{1/\alpha_i} + a_2h^2. \tag{2.11}$$

(The penalty function Q is defined by (2.6), and the underlying distribution is assumed to satisfy (2.8).) Formulae (2.10) and (2.11) follow from (2.9) and the result

$$Q = \sum_{i=1}^{2} b_{i2}(nh)^{-1}(\log nh)^{1/\alpha_i} + o_p\left\{\sum_{i=1}^{2}(nh)^{-1}(\log nh)^{1/\alpha_i}\right\},$$

which may be proved by an argument similar to that employed to derive Theorem 2.1.

Application of the penalty method involves choosing $h = \breve{h}$ to maximize $\breve{I}$, and then taking $\hat{I}(\breve{h})$ as the estimator of I. Of course, $\breve{I}(\breve{h})$ is an alternative choice, but we claim that the asymptotic bias of the latter is larger in absolute value than in the case of $\hat{I}(\breve{h})$. The standard deviations of both estimators are identical, up to and including terms of the same order as the biases. To appreciate these points, put $\alpha = \min(\alpha_1, \alpha_2)$, $b = (b_1 + b_2)/2$ if $\alpha_1 = \alpha_2$, $b = b_i/2$ if $\alpha_1 \neq \alpha_2$ and $\alpha = \alpha_i$. Observe that by (2.9) and (2.11),

$$S(h) \sim b(nh)^{-1}(\log nh)^{1/\alpha} + a_2h^2,$$
$$\hat{I} - \bar{I} = b(nh)^{-1}(\log nh)^{1/\alpha} + a_2h^2 + o_p\{(nh)^{-1}(\log nh)^{1/\alpha} + h^2\}.$$

Therefore, the bandwidth which maximizes $\check{I}$, or (asymptotically equivalently) which minimizes $S(h)$, satisfies

$$\tilde{h} \sim \{(b/2a_2)(2/3)^{1/\alpha} n^{-1} (\log n)^{1/\alpha}\}^{1/3}.$$

Hence,

$$\begin{aligned}
\hat{I}(\check{h}) - \bar{I} &= (2a_2)^{1/3} b^{2/3} (3/2)^{1/(3\alpha)} \left\{1 - \frac{1}{2}(2/3)^{1/\alpha}\right\} \{n^{-1}(\log n)^{1/\alpha}\}^{2/3} \\
&\quad + o_p\big[\{n^{-1}(\log n)^{1/\alpha}\}^{2/3}\big], \\
\check{I}(\check{h}) - \bar{I} &= -(2a_2)^{1/3} b^{2/3} (3/2)^{1/(3\alpha)} \left\{1 + \frac{1}{2}(2/3)^{1/\alpha}\right\} \{n^{-1}(\log n)^{1/\alpha}\}^{2/3} \\
&\quad + o_p\big[\{n^{-1}(\log n)^{1/\alpha}\}^{2/3}\big].
\end{aligned}$$

Noting these two formulae, and the fact that $\bar{I}$ is unbiased for I, we deduce that $\hat{I}(\check{h})$ has asymptotically positive bias, whereas $\check{I}(\check{h})$ has asymptotically negative bias; and that the absolute value of the asymptotic bias is greater in the case of $\check{I}(\check{h})$ than it is for $\hat{I}(\check{h})$.

Our last result in the present subsection treats a wide variety of different contexts. It is intended to show that there exists a very large range of situations where, for suitable choice of binwidth h, the result $\hat{I} = \bar{I} + o_p(n^{-1/2})$ obtains. This formula implies that $\hat{I}$ is root-n consistent for I, and also that $n^{1/2}(\hat{I} - I)$ is asymptotically $N(0, \sigma^2)$, where $\sigma^2 = \int f(\log f)^2$.

Since our assumptions about f do not explicitly describe tail behaviour then it is not possible, in the context of the result stated below, to be as explicit about the size of second-order terms as we were in the case of distributions with regularly varying or exponentially decreasing tails.

We assume the following regularity condition on f:

$$\begin{aligned}
&f > 0 \text{ on } (-\infty, \infty),\ f'' \text{ exists on } (-\infty, \infty) \text{ and is monotone on} \\
&(-\infty, a) \text{ and on } (b, \infty) \text{ for some } -\infty < a < b < \infty,\ |f''| + |f'/f| \\
&\text{is bounded, } \int |f''| + \int (f')^2 f^{-1} < \infty, \text{ and for some } \epsilon > 0, \\
&\sup_{|y| \le \epsilon} \int |f''(x)| \{\log f(x + y)\}^2 dx < \infty.
\end{aligned} \tag{2.12}$$

Let x_{1n}, x_{2n} denote respectively the largest negative, largest positive solutions of the equation $f(x) = (nh)^{-1}$.

THEOREM 2.2. *Assume condition* (2.12) *on* f, *and that the bandwidth* h *is chosen so that*

$$(nh)^{-1}(x_{2n} - x_{1n}) + \int_{(-\infty, x_{1n}) \cup (x_{2n}, \infty)} f |\log f| + h^2 = o(n^{-1/2}). \tag{2.13}$$

Then $E|\hat{I} - \bar{I}| = o(n^{-1/2})$ *as* $n \to \infty$.

Two examples serve to illustrate that for a very wide class of distributions, h may be chosen to ensure that (2.13) holds. For the first example, assume that $f(x) \sim b|x|^{-\alpha}$ as $|x| \to \infty$, where $b > 0$ and $\alpha > 3$. Then $x_{2n} - x_{1n} \sim b'(nh)^{1/\alpha}$, where $b' > 0$, and

$$\int_{(-\infty, x_{1n}) \cup (x_{2n}, \infty)} f|\log f| = O\{(nh)^{-1+(1/\alpha)} \log nh\}.$$

It follows that if $h \sim \text{const.} n^{-(1/4)-\epsilon}$ for some $0 < \epsilon < (\alpha - 3)/\{4(\alpha - 1)\}$ then (2.13) holds.

For the second example, assume that $f(x) \sim b_1 \exp(-b_2|x|^\alpha)$ as $|x| \to \infty$, where $b_1, b_2, \alpha > 0$. Then $x_{2n} - x_{1n} \sim b_3(\log nh)^{1/\alpha}$, where $b_3 > 0$, and

$$\int_{(-\infty, x_{1n}) \cup (x_{2n}, \infty)} f|\log f| = O\{(nh)^{-1}(\log nh)^{1/\alpha}\}.$$

It follows that if $h \sim \text{const.} n^{-(1/4)-\epsilon}$ for some $0 < \epsilon < 1/4$ then (2.13) holds.

2.5 *Simulation study*

In this subsection we discuss the results of a simulation study for five univariate distributions with binwidth h chosen by the empirical rule outlined in Subsection 2.3. The five distributions chosen are a standard normal (an example of a distribution with exponentially decreasing tails), and four Student's t distributions (examples of distributions with regularly varying tails). The four t distributions have degrees of freedom ν equal to 3, 4, 6 and 10. In the notation of the previous subsections, the rate of tail decrease α for a regularly varying distribution is equal to $\nu + 1$. Thus, the t distribution with three degrees of freedom or $\alpha = 4$, has the smallest integer number of degrees of freedom for which $\hat{I}$ is asymptotically equivalent to $\bar{I}$ up to terms of smaller order than $n^{-1/2}$, as shown in Subsection 2.2.

The true value of negative entropy I is known for the standard normal distribution and is equal to $\log(\sqrt{2\pi}) + 1/2 = -1.42$. Of course, this distribution has maximum entropy among all continuous distributions with mean zero and standard deviation one. While I is not known analytically for a t distribution, it may be estimated via numerical integration.

For each of the five distributions we investigated the behaviour of $\hat{I}$ and $\breve{I}$ for four different sample sizes, $n = 50$, 100, 200, and 500. For each sample size we conducted 100 simulations and approximated expected values by taking the average over the simulations. The quantities reported are $E(\hat{I})$, $[E(\hat{I} - I)^2]^{1/2}$ (i.e. root of the mean squared error), and $E(\hat{I}) - I$ (bias). Similar quantities are calculated for $\breve{I}$. The results are shown in Table 1.

For each particular sample, we calculate $\breve{I} = \hat{I} - Q$ over a grid of binwidth values $h = 0.1, 0.2, \ldots, H$, where H is the smallest multiple of 0.1 which is large enough to contain the entire sample. That is, when $h = H$, the histogram has only

Table 1. Performance of histogram estimators, $\hat{I}$ and $\breve{I}$.

n	$E(\hat{I})$	$[E(\hat{I}-I)^2]^{1/2}$	$E(\hat{I})-I$	$E(\breve{I})$	$[E(\breve{I}-I)^2]^{1/2}$	$E(\breve{I})-I$	h^* quantiles
$N(0,1)$ analytic value $= -1.42$							
50	−1.36	0.131	0.057	−1.50	0.135	−0.079	(0.5, 0.8, 1.2, 1.7)
100	−1.38	0.089	0.035	−1.47	0.093	−0.049	(0.6, 0.7, 0.9, 1.0)
200	−1.41	0.050	0.005	−1.46	0.064	−0.042	(0.6, 0.6, 0.7, 0.9)
500	−1.42	0.030	0.002	−1.44	0.037	−0.022	(0.5, 0.5, 0.6, 0.6)
t (dof $= 10$) integrated value $= -1.52$							
50	−1.42	0.161	0.099	−1.58	0.136	−0.058	(0.4, 0.8, 1.3, 2.4)
100	−1.47	0.100	0.052	−1.57	0.095	−0.045	(0.5, 0.7, 0.9, 1.1)
200	−1.51	0.062	0.012	−1.56	0.075	−0.043	(0.6, 0.7, 0.8, 0.9)
500	−1.51	0.039	0.008	−1.54	0.044	−0.022	(0.5, 0.5, 0.6, 0.7)
t (dof $= 6$) integrated value $= -1.59$							
50	−1.47	0.185	0.125	−1.63	0.129	−0.042	(0.4, 0.7, 1.4, 2.5)
100	−1.52	0.123	0.074	−1.63	0.104	−0.036	(0.4, 0.6, 0.9, 1.2)
200	−1.56	0.078	0.032	−1.62	0.078	−0.033	(0.5, 0.6, 0.8, 0.9)
500	−1.58	0.042	0.006	−1.62	0.051	−0.028	(0.5, 0.6, 0.6, 0.7)
t (dof $= 4$) integrated value $= -1.68$							
50	−1.50	0.233	0.174	−1.70	0.156	−0.022	(0.4, 0.5, 1.2, 2.1)
100	−1.58	0.145	0.100	−1.71	0.105	−0.028	(0.4, 0.6, 0.8, 1.4)
200	−1.63	0.081	0.048	−1.71	0.080	−0.036	(0.4, 0.5, 0.7, 0.9)
500	−1.66	0.049	0.020	−1.70	0.051	−0.022	(0.5, 0.5, 0.6, 0.7)
t (dof $= 3$) integrated value $= -1.77$							
50	−1.58	0.247	0.189	−1.78	0.159	−0.019	(0.4, 0.6, 1.1, 2.3)
100	−1.65	0.159	0.114	−1.79	0.112	−0.026	(0.4, 0.5, 0.9, 1.2)
200	−1.70	0.101	0.068	−1.79	0.079	−0.026	(0.4, 0.5, 0.7, 0.9)
500	−1.74	0.058	0.030	−1.79	0.055	−0.022	(0.4, 0.5, 0.6, 0.7)

one bin. For each particular h, the bins are centred at zero. The first bin covers $(-h/2, h/2)$, the second $(-2h/2, -h/2]$, the third $[h/2, 3h/2)$, and so on. Thus for each simulation we have $\breve{I}(h)$ for a grid of h values spaced by 0.1.

There are a number of options for choosing the optimal binwidth h^*. We could take as h^* that h which maximizes $\breve{I}(h)$, as our empirical rule suggests. However, in practice $\breve{I}(h)$ is bumpy due to sampling fluctuations and to the discreteness of histogram estimators. Thus, we smooth $\breve{I}(h)$ using running medians of seven, producing $\breve{I}_{sm}(h)$ say and thus choose the largest h which maximizes $\breve{I}_{sm}(h)$ as our h^*. The lower quartile, median, upper quartile and the 90th percentile of h^* over the simulations are given in Table 1. Since our previous theory was based on the unsmoothed $\hat{I}(h)$ and $\breve{I}(h)$, we use as our estimates $\hat{I}(h^*)$ and $\breve{I}(h^*)$. In our experience, this approach works well. A cursory investigation of other strategies, such as basing the smoothing window width on sample size, using smoothed versions $\hat{I}_{sm}(h^*)$ and $\breve{I}_{sm}(h^*)$ as estimates, or using the results of a search on a coarse grid to target a search area on a finer grid, did not produce significant changes in

the results.

Several comments may be made about the results in Table 1. Further simulation investigation is warranted given the fact that only 100 simulations were conducted for each situation. However, exploration revealed that two standard deviation confidence intervals for these simulation-based approximations of $\hat{I}$ and $\check{I}$ indicate deviations of at most three points in the last significant figure for all sample sizes and distributions considered. As the number of degrees of freedom increases from three to effectively infinity for the standard normal, and thus the tail rate of decrease α increases, the mean squared error of $\hat{I}$ decreases for any specific sample size n, as predicted by (2.1). In addition, throughout the table the ratio of the square root of the mean squared errors for two sample sizes n_1 and n_2 is close to $n_1^{-1/2} : n_2^{-1/2}$. For the chosen sample sizes $n = 50$, 100, 200, and 500, these ratios are between 0.63 and 0.71. The reported error ratios, for example $0.123 : 0.185 = 0.67$ for the t distribution with 6 degrees of freedom at $n = 50$ and $n = 100$, are between 0.53 and 0.70.

For most of the distributions, $\check{I}$ tends to be more biased than $\hat{I}$, a result shown in Subsection 2.3 to be true asymptotically. The bias of $\check{I}$ is negative and larger in absolute value than the positive bias of $\hat{I}$ for the standard normal. This relationship was shown in Subsection 2.4 to be true asymptotically for distributions with exponentially decreasing tails.

The variation in the value of h^* decreases as n increases, due to reduced sampling fluctuation. In addition, it seems to decrease slightly as the tail rate α increases.

2.6 *Proofs*

Outline of proof of Theorem 2.1

For the sake of brevity we indicate only those places where the proof differs significantly from that of Theorem 4.1 of Hall (1990), the differences occurring in the derivation of approximate formulae for "$E(T_{k2})$" ($k = 1, 2$). The new approximate formulae, in the case of large s and small r, may be described as follows. Put $g_k(x) = b_{k1} \exp(-b_{k2} x^\alpha)$, and let $M(x)$ denote a Poisson-distributed random variable having mean x. Then $E(T_{k2})$ is approximated by

$$\begin{aligned} t_k &\equiv n^{-1} \int_{h^{-1} g_k^{-1}(1)}^{h^{-1} g_k^{-1}\{(nh)^{-1}\}} E\big(M\{nhg_k(hx)\} \log\big[M\{nhg_k(hx)\}/nhg_k(hx)\big]\big) dx \\ &= (nh)^{-1} \int_1^{nh} E[M(y) \log\{M(y)/y\}] d_y g^{-1}\{(nh)^{-1} y\} \\ &= (nh)^{-1} \int_1^{nh} yE(\log[\{M(y)+1\}/y]) d_y \{b_{k2}^{-1} \log(b_{k1} nh/y)\}^{1/\alpha_k}. \end{aligned}$$

Since $E(\log[\{M(y)+1\}/y]) \sim (2y)^{-1}$ as $y \to \infty$, then

$$\begin{aligned} t_k &\sim (2nh)^{-1} \int_1^{nh} d_y \{b_{k2}^{-1} \log(b_{k1} nh/y)\}^{1/\alpha_k} \\ &\sim b_{k2}^{-1/\alpha_k} (2nh)^{-1} (\log nh)^{1/\alpha_k}. \end{aligned}$$

As in the proof of Theorem 4.1 of Hall (1990), the "variance component" of $\hat{I} - \bar{I}$ is asymptotic to $t_1 + t_2$; this produces the series on the right-hand side of (2.9). The next term, the quantity $-a_2 h^2$, represents the "bias component" and is obtained in a manner identical to that in Hall (1990).

Outline of proof of Theorem 2.2

We begin with a little notation. Let

$$q_i = n^{-1} E(N_i) = \int_{B_i} f(x)dx,$$

representing the probability that a randomly chosen data value falls in bin B_i. Put $\tilde{I} = n^{-1} \sum_i N_i \log(q_i/h)$. Theorem 2.2 is a corollary of the following result.

LEMMA 2.1. *If* $h = h(n) \to 0$ *and* $nh \to \infty$ *as* $n \to \infty$, *and if* (2.12) *holds, then*

$$|E(\tilde{I}) - I| = O(h^2), \tag{2.14}$$

$$\operatorname{var}(\tilde{I} - \bar{I}) = O(n^{-1}h^2), \tag{2.15}$$

$$E|\hat{I} - \tilde{I}| = O\left\{(nh)^{-1}(x_{2n} - x_{1n}) + \int_{(-\infty, x_{1n}) \cup (x_{2n}, \infty)} f|\log f| + h^2\right\} + o(n^{-1/2}) \tag{2.16}$$

as $n \to \infty$.

For the remainder, we confine attention to proving (2.14)–(2.16). We may assume without loss of generality that the constant v, in the definition of the centre $v + ih$ of the histogram bin B_i, equals zero.

(i) PROOF OF (2.14).

Observe that $E(\tilde{I}) = \int f \log g$, where $g(x) = q_i/h$ for $x \in B_i$. For small h, and all x,

$$g(x) = \int_{-1/2}^{1/2} f(x + ih - x + hy)dy = f(x) + (ih - x)f'(x) + R_1(x),$$

where $|R_1(x)| \le C_1 h^2 \{|f''(x - h)| + |f''(x + h)|\}$ and $C_1, C_2, \ldots$ denote positive constants. Thus,

$$\log g(x) = \log f(x) + (ih - x)f'(x)f(x)^{-1} + R_2(x) \tag{2.17}$$

where

$$\begin{aligned} |R_2(x)| \le C_2 \big(& h^2 \big[\{f'(x)/f(x)\}^2 + \{|f''(x - h)| + |f''(x + h)|\} f(x)^{-1} \big] \\ & \times I\big[C_1 h^2 \{|f'(x - h)| + |f''(x + h)|\} f(x)^{-1} \le 1/2\big] \\ & + \{|\log f(x)| + |\log f(x - h)| + |\log f(x + h)|\} \\ & \times I\big[C_1 h^2 \{|f''(x - h)| + |f''(x + h)|\} f(x)^{-1} > 1/2\big] \big). \end{aligned} \tag{2.18}$$

The last inequality entails

$$\int |R_2|f \le C_2 h^2 \left[\int (f')^2 f^{-1} + 2 \int |f''| + \int \{ |\log f(x)| + |\log f(x-h)| + |\log f(x+h)| \} 2C_1 \{ |f''(x-h)| + |f''(x+h)| \} dx \right] = O(h^2).$$

Therefore, writing $i = i(x)$ for the index of the block B_i containing x,

$$E(\tilde{I}) = \int f \log g = \int f \log f + \int (ih - x) f'(x) dx + O(h^2) = I + O(h^2),$$

which establishes (2.14).

(ii) PROOF OF (2.15).
Observe that

$$\tilde{I} - \bar{I} = n^{-1} \sum_j \left\{ \sum_i I(X_j \in B_i) \log (q_i/h) - \log f(X_j) \right\},$$

whence

$$\begin{aligned} n\text{var}(\tilde{I} - \bar{I}) &= \text{var} \left[\sum_i I(X \in B_i) \{ \log (q_i/h) - \log f(X) \} \right] \\ &\le \sum_i E\left[I(X \in B_i) \{ \log(q_i/h) - \log f(X) \}^2 \right] \\ &= \int \{\log(g/f)\}^2 f, \end{aligned} \tag{2.19}$$

where g is as in (i) above. We may prove from (2.17) and (2.18) that

$$\int \{\log(g/f)\}^2 f = O(h^2),$$

which together with (2.19) implies (2.15).

(iii) PROOF OF (2.16).
Define $\Delta_i = (N_i - nq_i)/(nq_i)$, and let $\sum_i'$, $\sum_i''$ denote summation over values of i such that $nq_i > 1$, $nq_i \le 1$ respectively. Write m for the number of i's such that $nq_i > 1$. Since

$$\hat{I} - \tilde{I} = \sum N_i \log (1 + \Delta_i)$$

and $|\log(1+u)-u| \le 2\{u^2 + |\log(1+u)|I(u<1/2)\}$ then

$$n|\hat{I}-\tilde{I}| \le \left|\sum_i{}' N_i\Delta_i\right| + 2\sum_i{}' N_i\Delta_i^2 + 2\sum_i{}' N_i\left|\log\left(1+\Delta_i\right)\right|I\left(\Delta_i < -\frac{1}{2}\right) + \sum_i{}'' N_i\left|\log\left(1+\Delta_i\right)\right|. \tag{2.20}$$

Now, $N_i\Delta_i = (nq_i)^{-1}(N_i-nq_i)^2 + N_i - nq_i$, and so

$$E\left|\sum_i N_i\Delta_i\right| \le \sum_i{}'(1-q_i) + \left\{E\left(\sum_i{}' N_i - \sum_i{}' nq_i\right)^2\right\}^{1/2} \le m + n^{1/2}\left(1-\sum_i{}' q_i\right)^{1/2} = m + o(n^{1/2}),$$

since $\sum_i{}' q_i \to 1$. Likewise, noting that $N_i\Delta_i^2 = (nq_i)^{-2}(N_i-nq_i)^3 + (nq_i)^{-}(N_i-nq_i)^2$, we see that

$$E\left(\sum_i{}' N_i\Delta_i^2\right) = \sum_i{}'(nq_i)^{-1}(1-3q_i+2q_i) + \sum_i{}'(1-q_i) \le 2m.$$

If $\Delta_i < -1/2$ then $1+\Delta_i = N_i/(nq_i) < 1/2$, and so

$$N_i\left|\log\left(1+\Delta_i\right)\right| = N_i\log\left\{(nq_i)/N_i\right\} \le N_i\log(nq_i),$$

whence

$$E\left\{\sum_i{}' N_i\left|\log(1+\Delta_i)\right|I\left(\Delta_i<-\frac{1}{2}\right)\right\} \le \sum_i{}'\left\{\log(nq_i)\right\}E\left\{N_iI\left(\Delta_i<-\frac{1}{2}\right)\right\} \le 2\sum_i{}' nq_i\left|\log(nq_i)\right|P\left(\Delta_i<-\frac{1}{2}\right)^{1/2},$$

the last line by the Cauchy-Schwartz inequality. By Bernstein's inequality,

$$P\left(\Delta_i<-\frac{1}{2}\right) = P\left(N_i-nq_i<-\frac{1}{2}nq_i\right) \le \exp\left(-\frac{1}{16}nq_i\right),$$

and so, defining $C_3 = \sup_{x>0} x|\log x|\exp(-x/32)$, we have

$$E\left\{\sum_i{}' N_i\left|\log(1+\Delta_i)\right|I\left(\Delta_i<-\frac{1}{2}\right)\right\} \le 2C_3m.$$

Combining the estimates from (2.20) down we deduce that

$$nE|\hat{I} - \tilde{I}| \le (5 + 4C_3)m + o(n^{1/2}) + \sum_i{}'' E\{N_i |\log(1 + \Delta_i)|\}. \tag{2.21}$$

Observe next that $E\{N_i|\log(1+\Delta_i)|\} \le E(N_i \log N_i) + nq_i|\log(nq_i)|$. If $nq_i \le 1$ then $E(N_i \log N_i) \le E(N_i^2) \le 2nq_i$, and $|\log(nq_i)| = -\log(nq_i) \le -\log(q_i/h)$, whence

$$\sum_i{}'' E\{N_i|\log(1+\Delta_i)|\} \le 2n \sum_i{}'' q_i - n \sum_i{}'' q_i \log(q_i/h). \tag{2.22}$$

Write B for the union of B_i over indices i satisfying $nq_i \le 1$, and let g be as in (i). By that result, $\int f|\log(f/g)| = O(h^2)$, and so

$$\sum_i{}'' q_i \log(q_i/h) = \int_B f \log g = \int_b f \log f + O(h^2).$$

Put $D = (-\infty, x_{1n}) \cup (x_{2n}, \infty)$. In view of the monotonicity of the tails of f,

$$-\int_B f \log f = \int_D f|\log f| + o(n^{-1/2}),$$

implying that

$$-\sum_i{}'' q_i \log(q_i/h) = \int_D f|\log f| + O(h^2) + o(n^{-1/2}).$$

Similarly,

$$\sum_i{}'' q_i = \int_D f + O(h^2) + o(n^{-1/2}) \le \int_d f|\log f| + O(h^2) + o(n^{-1/2}),$$

whence by (2.22),

$$n^{-1} \sum_i E\{N_i|\log(1+\Delta_i)|\} \le 3 \int_D f|\log f| + O(h^2) + o(n^{-1/2}).$$

The desired result (2.16) follows from this formula and (2.21).

3. Kernel estimators

3.1 *Methodology*

Let $X_1, \ldots, X_n$ denote a random sample drawn from a p-variate distribution with density f, let K be a p-variate density function, and write h for a (scalar) bandwidth. (In practice, the data would typically be standardised for scale before using a single bandwidth; see Silverman ((1986), p. 84 ff.)) Then

$$\hat{f}_i(x) = \{(n-1)h^p\}^{-1} \sum_{j \neq i} K\{(x - X_j)/h\}$$

is a "leave-one-out" estimator of f, and

$$\hat{I}_k = n^{-1} \sum_{i=1}^{n} \log \hat{f}_i(X_i)$$

is an estimator of negative entropy, I.

In the case $p = 1$, properties of $\hat{I}_k$ have been studied in the context of estimating Kullback-Leibler loss (1987). There it was shown that, provided the kernel function has appropriately heavy tails (e.g. if K is a Student's t density, rather than a Normal density), and if the tails of f are decreasing like $|x|^{-\alpha}$ as $|x| \to \infty$, then

$$\hat{I}_k = \bar{I} - \{C_1(nh)^{-1+(1/\alpha)} + C_2 h^4\} + o_p\{(nh)^{-1+(1/\alpha)} + h^4\}, \tag{3.1}$$

where $C_1, C_2 > 0$.

More generally, suppose $p \geq 1$ and the tails of f decrease like $\|x\|^{-\alpha}$, say $f(x) = c_1(c_2 + \|x\|^2)^{-\alpha/2}$ for $c_1, c_2 > 0$ and $\alpha > p$. (The latter condition is necessary to ensure that this f is integrable.) Then, defining $\bar{I} = n^{-1} \sum \log f(X_i)$, we have

$$\hat{I}_k = \bar{I} - \{C_1(nh^p)^{-1+(p/\alpha)} + C_2 h^4\} + o_p\{(nh^p)^{-1+(p/\alpha)} + h^4\}, \tag{3.2}$$

for positive constants C_1 and C_2. The techniques of proof are very similar to those in Hall (1987), and so we shall not elaborate on the proof.

Of course, $\bar{I}$ is unbiased and root-n consistent for I, with variance $n^{-1} \cdot \{\int (\log f)^2 \, f - I^2\}$. The second order term in (3.2) is

$$J = C_1(nh^p)^{-1+(p/\alpha)} + C_2 h^4,$$

and is minimized by taking h to be of size n^{-a} where

$$a = (\alpha - p)/\{\alpha(p+4) - p^2\}. \tag{3.3}$$

Then J is of size n^{-4a}, which is of smaller order than $n^{-1/2}$ if and only if $1 \leq p \leq 3$ and $\alpha > p(8-p)/(4-p)$. When $p = 1$ this reduces to $\alpha > 7/3$, which is equivalent

to existence of a moment higher than the $1\frac{1}{3}$'rd. (For example, finite variance suffices.)

Recall from Subsection 2.2 that histogram estimators only allow root-n consistent estimation of entropy when $p = 1$. We have just seen that nonnegative kernel estimators extend this range to $1 \leq p \leq 3$, and so they do have advantages. The case $p = 2$ is of practical interest since practitioners of exploratory projection pursuit sometimes wish to project a high-dimensional distribution into two, rather than one, dimensions. As noted above, in the case of a density whose tails decrease like $\|x\|^{-\alpha}$ we need

$$\alpha > 2(8-2)/(4-2) = 6$$

if we are to get $\hat{I} = \bar{I} + o_p(n^{-1/2})$ in $p = 2$ dimensions. This corresponds to the existence of a moment higher than the fourth.

Since C_1 and C_2 in formula (3.2) are both positive then a simple practical, empirical rule for choosing bandwidth is to select h so as to maximize $\hat{I}_k = \hat{I}_k(h)$. Now, it may be proved that (3.2) is available uniformly over h's in any set $\mathcal{H}_n$ such that $\mathcal{H}_n \subseteq (n^{-1+\delta}, n^{-\delta})$ for some $0 < \delta < 1/2$ and $\#\mathcal{H}_n = O(n^C)$ for some $C > 0$. If the maximization is taken over a rich set of such h's then $\hat{I}_k = \bar{I} + O_p(n^{-4a})$, where a is given by (3.3), and so $\hat{I}_k = \bar{I} + o_p(n^{-1/2})$ if $1 \leq p \leq 3$ and $\alpha > p(8-p)/(4-p)$.

In principle, the estimator $\hat{I}_k$ may be constructed without using "leave-one-out" methods. If we define

$$\hat{f}(x) = (nh^p)^{-1} \sum_{j=1}^{n} K\{(x - X_j)/h\}$$

then an appropriate entropy estimator is given by

$$\begin{aligned} \tilde{I}_k &= n^{-1} \sum_{i=1}^{n} \log \hat{f}(X_i) \\ &= n^{-1} \sum_{i=1}^{n} \log \{(1 - n^{-1}) \hat{f}_i(X_i) + (nh)^{-1} K(0)\}. \end{aligned}$$

Here, as noted above, it is essential that the kernel have appropriately heavy tails; for example, K could be a Student's t density.

Formulae (3.1) and (3.2) continue to hold in this case, except that the constant C_1 is no longer positive. Compare formula (2.1), which is also for the case of an estimator that is not constructed by the "leave-one-out" method. Thus, the bandwidth selection argument described in the previous paragraph is not appropriate. A penalty term should be subtracted before attempting maximization, much as in the case described in Section 2.

3.2 *Simulation study*

This subsection describes a simulation study of the behaviour of our kernel estimator of negative entropy, $\hat{I}_k$. It is similar to the previous simulation study of the histogram estimator presented in Subsection 2.5 and its interpretation is subject

to the same caveat regarding the number of simulations. In all univariate cases shown, the kernel used is a Student's t with four degrees of freedom, which is heavy-tailed. In the bivariate case, the kernel was a product of two such functions. The bandwidth h chosen is that which maximizes $\hat{I}_k(h)$, this being the empirical rule discussed previously. The function $\hat{I}_k(h)$ is not smoothed first before h is chosen as it does not fluctuate as much as the inherently discrete histogram estimator. As a result, the quantiles of h vary much less for the kernel estimator examples presented in Table 2 than for the associated histogram estimator examples of Table 1.

Table 2. Performace of kernel estimator.

n	$E(\hat{I}_k)$	$[E(\hat{I}_k - I)^2]^{1/2}$	$E(\hat{I}_k) - I$	h quantiles
One dimension				
$N(0,1)$ true value $= -1.42$				
50	-1.45	0.110	-0.032	$(0.3, 0.4, 0.4, 0.5)$
100	-1.44	0.070	-0.023	$(0.3, 0.3, 0.4, 0.4)$
200	-1.44	0.046	-0.016	$(0.2, 0.3, 0.3, 0.3)$
t (dof $= 6$) integrated value $= -1.59$				
50	-1.63	0.142	-0.040	$(0.4, 0.5, 0.5, 0.6)$
100	-1.59	0.092	-0.001	$(0.3, 0.4, 0.5, 0.5)$
200	-1.61	0.067	-0.023	$(0.2, 0.3, 0.4, 0.4)$
t (dof $= 4$) integrated value $= -1.68$				
50	-1.68	0.134	-0.004	$(0.3, 0.5, 0.6, 0.7)$
100	-1.68	0.106	-0.006	$(0.2, 0.4, 0.5, 0.5)$
200	-1.67	0.078	0.007	$(0.2, 0.3, 0.4, 0.5)$
t (dof $= 3$) integrated value $= -1.77$				
50	-1.75	0.151	0.013	$(0.2, 0.5, 0.6, 0.7)$
100	-1.75	0.131	0.015	$(0.2, 0.4, 0.5, 0.6)$
200	-1.75	0.086	0.013	$(0.2, 0.3, 0.4, 0.5)$
Two dimensions				
$N(0, I)$ true value $= -2.84$				
50	-2.94	0.174	-0.098	$(0.4, 0.5, 0.5, 0.5)$
100	-2.91	0.117	-0.076	$(0.4, 0.4, 0.4, 0.5)$
200	-2.88	0.082	-0.046	$(0.3, 0.4, 0.4, 0.4)$
$N(0, V)$ correlation 0.8 true value $= -2.33$				
50	-2.48	0.207	-0.149	$(0.3, 0.3, 0.3, 0.4)$
100	-2.45	0.150	-0.119	$(0.2, 0.3, 0.3, 0.3)$
200	-2.40	0.107	-0.072	$(0.2, 0.2, 0.2, 0.3)$

Four univariate examples are presented: the normal and three Student's t with 6, 4 and 3 degrees of freedom respectively. In general, the bias is negative, as predicted by (3.1), and the error is less than that for the associated histogram

estimator examples, again as expected. The bias is positive for the most heavy-tailed distribution, the Student's t with three degrees of freedom, perhaps due to the fact that higher-order terms are having a large effect on the expansion (3.1).

Two bivariate examples are presented. Both are bivariate normals; in the first, components are independent, and in the second, the correlation coefficient is 0.8. In each case, the true negative entropy is known. The kernel estimator performs well in both examples, given the small sample size. However, the computational work required to calculate the distance between every pair of points makes the kernel estimator intractable for exploratory projection pursuit. The binning performed in the histogram estimator reduces the work required in the $p = 1$ case from $O(n^2)$ to $O(m^2)$, where m is the number of bins. Unfortunately this approach cannot be used in the $p = 2$ case, as discussed in Section 2.

Acknowledgements

The authors would like to thank the referees for their helpful comments and suggestions.

References

Friedman, J. H. (1987). Exploratory projection pursuit, *J. Amer. Statist. Assoc.*, **76**, 817–823.

Goldstein, L. and Messer, K. (1991). Optimal plug-in estimators for nonparametric functional estimation, *Ann. Statist.* (to appear).

Hall, P. (1987). On Kullback-Leibler loss and density estimation, *Ann. Statist.*, **15**, 1491–1519.

Hall, P. (1989). On polynomial-based projection indices for exploratory projection pursuit, *Ann. Statist.*, **17**, 589–605.

Hall, P. (1990). Akaike's information criterion and Kullback-Leibler loss for histogram density estimation, *Probab. Theory Related Fields*, **85**, 449–466.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution, *Ann. Statist.*, **3**, 1163–1174.

Huber, P. J. (1985). Projection pursuit (with discussion), *Ann. Statist.*, **13**, 435–525.

Joe, H. (1989). Estimation of entropy and other functionals of a multivariate density, *Ann. Inst. Statist. Math.*, **41**, 683–697.

Jones, M. C. and Sibson, R. (1987). What is projection pursuit? (with discussion), *J. Roy. Statist. Soc. Ser. B*, **150**, 1–36.

Morton, S. C. (1989). Interpretable projection pursuit, Ph.D. Dissertation, Stanford University, California.

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

Vasicek, O. (1986). A test for normality based on sample entropy, *J. Roy. Statist. Soc. Ser. B*, **38**, 54–59.

Zipf, G. K. (1965). *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology* (Facsimile of 1949 edn.), Hafner, New York.