Ann. Inst. Statist. Math.
Vol. 41, No. 4, 683-697 (1989)

ESTIMATION OF ENTROPY AND OTHER FUNCTIONALS OF A MULTIVARIATE DENSITY*

HARRY JOE

Department of Statistics, University of British Columbia, 2021 West Mall, Vancouver, B.C., Canada V6T 1W5

(Received August 25, 1987; revised March 7, 1989)

Abstract. For a multivariate density f with respect to Lebesgue measure μ, the estimation of $\int J(f)f\,d\mu$, and in particular $\int f^2 d\mu$ and $-\int f \log f\,d\mu$, is studied. These two particular functionals are important in a number of contexts. Asymptotic bias and variance terms are obtained for the estimators $\hat{I} = \int J(\hat{f})dF_n$ and $\tilde{I} = \int J(\hat{f})\hat{f}\,d\mu$, where $\hat{f}$ is a kernel density estimate of f and F_n is the empirical distribution function based on the random sample $X_1,\ldots,X_n$ from f. For the two functionals mentioned above, a first order bias term for $\tilde{I}$ can be made zero by appropriate choices of non-unimodal kernels. Suggestions for the choice of bandwidth are given; for $\hat{I} = \int \hat{f} dF_n$, a study of optimal bandwidth is possible.

Key words and phrases: Kernel density estimation, multivariate density, empirical process, entropy.

1. Introduction

This paper is concerned with the nonparametric estimation of a functional of a multivariate density of the form $I(f) = \int J(f)f\,d\mu$, where f is a p-variate density with respect to Lebesgue measure μ and J is a smooth real-valued function. Of particular interest are $I_1(f) = \int f^2 d\mu$ and $I_2(f) = -\int f \log f\,d\mu$. $I_1(f)$ is important in nonparametric inference (Bhattacharya and Roussas (1969) and Schweder (1975)) and is called a projection index by Huber (1985) in his paper on projection pursuit. The entropy function $I_2(f)$ is a measure of dispersion of the density f and negative entropy is also used as a projection index in Jones and Sibson (1987). The author's

*This research was supported by an NSERC Grant and a UBC Killam Research Fellowship.

motivation for this research is the estimation of relative entropies of a multivariate density with respect to products of marginal densities; for example, $\int f \log\left[f \Big/ \prod_{i=1}^{p} f_i\right] d\mu$, where f_i are the univariate marginals of the p-variate density f. This and other relative entropies are proposed as (probabilistic) measures of dependence and conditional dependence in Joe (1987, 1989). Sample estimates would provide measures of dependence for data. Some theory for the estimation of $I(f)$ is a starting point for estimation of relative entropies.

The estimation of $I_1(f)$ and $I_2(f)$ or the more general $I(f)$ have been discussed in the statistical literature only for the univariate case, and mainly asymptotic results have been obtained. Relevant references for estimation functionals of a density are Bhattacharya and Roussas (1969), Dmitriev and Tarasenko (1973, 1974), Schuster (1974), Schweder (1975), Ahmad (1976), Ahmad and Lin (1976), Prakasa Rao (1983), Pawlak (1986) and Silverman (1986). Schweder (1975) includes a method for choosing a bandwidth for the kernel density method. In this paper, the kernel density method is studied for finite samples; this method is easier to study analytically than other density estimation methods. Two estimators, $\hat{I}$ and $\tilde{I}$, in (2.3) and (2.4), respectively, are considered—the first one avoids numerical integration and the second one does not. Asymptotic expected values and variances up to second order terms are derived after representing the estimators in terms of the empirical process. From these, it can be seen that for $\hat{I}_1(f)$, it is possible to eliminate a bias term by subtracting a deterministic quantity or by choosing a suitable kernel, and for $\hat{I}_2(f)$, a choice of a kernel satisfying certain conditions will decrease the order of bias and mean squared error. It is indicated how to estimate the asymptotic variance to obtain a standard error. For $\hat{I}_1(f)$, a method for choosing a bandwidth based on a theoretical optimal bandwidth is given. Also, a method is given for choosing a bandwidth for $\hat{I}_2(f)$. For functionals like $I_1(f)$ and $I_2(f)$, estimates can vary a lot as the bandwidth changes. Also, the "best" bandwidth depends on the functional.

Representation in terms of the empirical process and a lemma for computation of asymptotic expected values and variances based on this representation are given in Section 2. The specific functionals $I_1(f)$ and $I_2(f)$ are studied in detail in Sections 3 and 4, respectively. All results and recommendations of estimators have been guided by some computer simulations and considerations of computational efficiency.

2. Estimation based on the kernel method

Let $X_1,\ldots, X_n$ be a random sample of size n from the p-variate density f with distribution function F. We consider estimation of

$$I = I(f) = \int_S J(f) f \, d\mu = \int_S J(f) dF \tag{2.1}$$

based on a kernel density estimate, where S may be a bounded or unbounded set. The kernel density estimate of $f(x)$ with bandwidth $h > 0$ is

$$\hat{f}(x) = (nh^p)^{-1} \sum_{i=1}^{n} k((x - X_i)/h) = n^{-1} \sum_{i=1}^{n} k_h(x - X_i), \quad x \in \mathscr{R}^p, \tag{2.2}$$

where k is a p-variate density and $k_h(u) = h^{-p} k(u/h)$.

Assumptions that are used are all summarized here for easier reference. A, B, C2, D and E are assumed throughout (unless stated otherwise) and C4, F and G are used for some results. Assumption A means that $\int_{\mathscr{R}^p} J(f) f \, d\mu$ is finite, so that it can be approximated arbitrarily closely by $\int_S J(f) f \, d\mu$ for a compact or bounded set S. This means that we can ignore the "tails" where f is small for certain expansions.

A. S is a bounded set such that f is bounded below on it by a positive constant and $\int_S J(f) f \, d\mu \approx \int_{\mathscr{R}^p} J(f) f \, d\mu$.

B. The p components of X_i have approximately the same scale (for some functionals $I(f)$ such as for the two particular functionals mentioned in Section 1, this can be assumed without loss of generality because the data can be scaled first and scaling affects $I(f)$ by a known factor).

Cm. f has continuous derivatives up to and including the m-th order.

D. J is thrice differentiable and $\int J^2(f) f \, d\mu$ exists.

E. The kernel satisfies $k(u) = k(-u)$.

F. $k(u)$ is a kernel of the form $k(u) = k(u_1, \ldots, u_p) = \prod_j k_0(u_j)$, where k_0 is a symmetric univariate density satisfying $\int v^2 k_0(v) dv = 1$.

G. $f = f(x_1, \ldots, x_p)$ has continuous first and second order derivatives and $|\partial f / \partial x_j|$, $|\partial^2 f / \partial x_j^2|$, $j = 1, \ldots, p$, are all dominated by integrable functions.

Let F_n be the empirical distribution of $X_1, \ldots, X_n$. Then $\hat{f}(x) = \int k_h(x - y) dF_n(y)$. Two estimators of $I(f)$ are

$$\hat{I} = \hat{I}(f) = \int_S J(\hat{f}) dF_n = n^{-1} \sum_{X_i \in S} J(\hat{f}(X_i)) \tag{2.3}$$

and

$$\tilde{I} = \tilde{I}(f) = \int_S J(\hat{f}) \hat{f} \, d\mu. \tag{2.4}$$

$\hat{I}$ replaces the right-hand side of (2.1) with a density estimate, and an empirical distribution and numerical integration is not necessary. $\tilde{I}$ replaces the center term of (2.1) with a density estimate. The amount of computation for $\hat{I}$ increases in the square of the sample size n and linearly in p, whereas the amount of computation for $\tilde{I}$ increases linearly in n and exponentially in p. For n in the range of 50 to a few hundred, $\tilde{I}$ may be faster to compute for $p = 1$, but $\hat{I}$ is generally much faster to compute for $p \geq 2$.

In order to study the asymptotic expected values, variances and bias terms of $\hat{I}$ and $\tilde{I}$, we define a few quantities and state a lemma. The terms in the expansions are with the bandwidth h fixed and the sample size n increasing to ∞. The h order of the terms are then obtained by letting $h \to 0$. In the expressions below, an integral without a region specified will be assumed to be an integral over $\mathscr{R}^p$.

Let $f_h(x) = E\hat{f}(x) = \int k_h(x - y)dF(y)$. Let $U_n(x) = n^{1/2}(F_n(x) - F(x))$ and let $V_n(x) = n^{1/2}(\hat{f}(x) - f_h(x))$. Then, $V_n(x) = \int k_h(x - y)dU_n(y)$. To study $\hat{I}$, an asymptotic expansion (with Assumption D) is

$$\begin{aligned}
\hat{I} &= \int_S J(f_h)dF + \int_S [J(\hat{f}) - J(f_h)]dF + n^{-1/2}\int_S J(f_h)dU_n \\
&\quad + n^{-1/2}\int_S [J(\hat{f}) - J(f_h)]dU_n \\
&= \int_S J(f_h)dF + n^{-1/2}\int_S J'(f_h)V_n dF \\
&\quad + 0.5n^{-1}\int_S J''(f_h)V_n^2 dF + \frac{1}{6}n^{-3/2}\int_S J'''(f_h)V_n^3 dF + n^{-1/2}\int_S J(f_h)dU_n \\
&\quad + n^{-1}\int_S J'(f_h)V_n dU_n + 0.5n^{-3/2}\int_S J''(f_h)V_n^2 dU_n + o(n^{-3/2}),
\end{aligned}$$

where S is bounded if necessary so that the Taylor series expansion is valid (for example, if $J(t) = -\log t$, S must be bounded for the expansion to make sense, but for $J(t) = t$, S can be unbounded). For studying $\tilde{I}$, let $L(t) = tJ(t)$. An asymptotic expansion for $\tilde{I}$ is

$$\begin{aligned}
\tilde{I} &= \int_S L(f_h)d\mu + \int_S [L(\hat{f}) - L(f_h)]d\mu \\
&= \int_S L(f_h)d\mu + n^{-1/2}\int_S L'(f_h)V_n d\mu + 0.5n^{-1}\int_S L''(f_h)V_n^2 d\mu \\
&\quad + \frac{1}{6}n^{-3/2}\int_S L'''(f_h)V_n^3 d\mu + o(n^{-3/2}).
\end{aligned}$$

Let $\chi_S(x)$ be the indicator function for the set S. After some algebra, we can write

$$\hat{I} = \int_S J(f_h)dF + n^{-1/2}\int A_1(x)dU_n(x) + n^{-1}\int\int B_1(x,y)dU_n(x)dU_n(y) + n^{-3/2}\int\int\int C_1(x,y,z)dU_n(x)dU_n(y)dU_n(z) + o(n^{-3/2}), \tag{2.5}$$

where

$$A_1(x) = \int_S J'(f_h(y))k_h(y-x)dF(y) + J(f_h(x))\chi_S(x),$$

$$B_1(x,y) = 0.5\int_S J''(f_h(z))k_h(z-x)k_h(z-y)dF(z) + J'(f_h(x))k_h(x-y)\chi_S(x)$$

and

$$C_1(x,y,z) = \frac{1}{6}\int_S J'''(f_h(w))k_h(w-x)k_h(w-y)k_h(w-z)dF(w) + 0.5J''(f_h(x))k_h(x-y)k_h(x-z)\chi_S(x).$$

Similarly,

$$\tilde{I} = \int_S L(f_h)d\mu + n^{-1/2}\int A_2(x)dU_n(x) + n^{-1}\int\int B_2(x,y)dU_n(x)dU_n(y) + n^{-3/2}\int\int\int C_2(x,y,z)dU_n(x)dU_n(y)dU_n(z) + o(n^{-3/2}), \tag{2.6}$$

where $A_2(x) = \int_S L'(f_h(y))k_h(y-x)dy$, $B_2(x,y) = 0.5\int_S L''(f_h(z))k_h(z-x)k_h(z-y)dz$ and $C_2(x,y,z) = (1/6)\int_S L'''(f_h(w))k_h(w-x)k_h(w-y)k_h(w-z)dw$.

To obtain asymptotic expected values and variances, the following lemma for working with U_n is needed. The details are straightforward but a bit tedious, so that they will not be provided here.

LEMMA 2.1. *Assume that all integrals given here are well-defined and finite.*

(i) *Let $a(x)$ be a function on $\mathscr{R}^p$. If the integral $\int a(x)dF(x)$ exists, then $E\int a(x)dU_n(x) = 0$.*

(ii) *Let $a(x,y)$ be a function on $\mathscr{R}^p \times \mathscr{R}^p$. Then*

$$E\int\int a(x,y)dU_n(x)dU_n(y) = \int a(x,x)dF(x) - \int\int a(x,y)dF(x)dF(y).$$

(iii) *Let $a(x, y, z)$ be a function on $\mathscr{R}^p \times \mathscr{R}^p \times \mathscr{R}^p$. Then*

$$E\int\int\int a(x, y, z)dU_n(x)dU_n(y)dU_n(z)$$
$$= n^{-1/2}\Big[\int a(x, x, x)dF(x) - \int\int a(x, x, z)dF(x)dF(z)$$
$$- \int\int a(x, y, x)dF(x)dF(y) - \int\int a(x, y, y)dF(x)dF(y)$$
$$+ 2\int\int\int a(x, y, z)dF(x)dF(y)dF(z)\Big].$$

(iv) *Let $a(w, x, y, z)$ be a function on $\mathscr{R}^p \times \mathscr{R}^p \times \mathscr{R}^p \times \mathscr{R}^p$. Then*

$$E\int\int\int\int a(w, x, y, z)dU_n(w)dU_n(x)dU_n(y)dU_n(z)$$
$$= \int\int [a(w, w, y, y) + a(w, y, w, y) + a(w, y, y, w)]dF(w)dF(y)$$
$$- \int\int\int [a(w, w, x, y) + a(w, x, w, y) + a(w, x, y, w)$$
$$+ a(w, x, x, y) + a(w, x, y, x)$$
$$+ a(w, x, y, y)]dF(w)dF(x)dF(y)$$
$$+ 3\int\int\int\int [a(w, x, y, z)dF(w)dF(x)dF(y)dF(z)] + O(n^{-1}).$$

By applying Lemma 2.1 to (2.5), $E\hat{I} = \theta_h + a_1(h)n^{-1} + a_2(h)n^{-2} + o(n^{-2})$, where $\theta_h = \int_S J(f_h)dF$,

$$(2.7)\quad a_1(h) = h^{-p}\Big[0.5K_2\int_S f_h^*(y)J''(f_h(y))dF(y) + k(0)\int_S J'(f_h(y))dF(y)\Big]$$
$$- \Big[0.5\int_S f_h^2(y)J''(f_h(y))dF(y) + \int_S J'(f_h(y))f_h(y)dF(y)\Big],$$

$K_2 = \int k^2(x)dx$, $l(u) = k^2(u)/K_2$, $f_h^*(y) = h^{-p}\int l((y - x)/h)dF(x)$ and $a_2(h) = O(h^{-2p})$. The $O(h^{-p})$ term of $a_1(h)$ comes from $\int B_1(x, x)dF(x)$. Also, $E(\hat{I} - \theta_h)^2 = b_1(h)n^{-1} + b_2(h)n^{-2} + o(n^{-2})$, where by expansion with $h \to 0$, $b_1(h) = O(1)$ and $b_2(h) = O(h^{-2p}) + O(h^{-p})$. The $O(h^{-p})$ term of $b_2(h)$ comes from many sources when applying Lemma 2.1, and the $O(h^{-2p})$ term of $b_2(h)$ comes from $\Big[\int B_1(x, x)dF(x)\Big]^2$. With Assumption C2, $\theta_h - I$ is $O(h^2)$, so

that by combining all of the above, in general, the mean squared error is

$$(2.8)\qquad E(\hat{I}-I)^2 = E(\hat{I}-\theta_h)^2 + 2(\theta_h - I)E(\hat{I}-\theta_h) + (\theta_h - I)^2$$
$$= O(n^{-1}) + O(n^{-2}h^{-2p}) + O(n^{-2}h^{-p}) + O(n^{-1}h^{2-p})$$
$$+ O(n^{-2}h^{2-2p}) + O(h^4)\,.$$

It will be shown in Sections 3 and 4 for $J(t) = t$ and $J(t) = -\log t$ that (2.8) and the order of bias can be improved on, because $\int B_1(x,x)dF(x)$ can be 0 in the former case and $O(h^4)$ in the latter case by choosing kernels that satisfies certain conditions. In general, improvement is not possible.

Similarly, by applying Lemma 2.1 to (2.6), $E\tilde{I} = \eta_h + c_1(h)n^{-1} + c_2(h)n^{-2} + o(n^{-2})$, where $\eta_h = \int_S L(f_h)d\mu$,

$$c_1(h) = 0.5h^{-p}K_2\int_S f_h^*(y)L''(f_h(y))dy - 0.5\int_S f_h^2(y)L''(f_h(y))dy$$

and $c_2(h) = O(h^{-2p})$. The $O(h^{-p})$ term in $c_1(h)$ comes from $\int B_2(x,x)dF(x)$. Also, $E(\tilde{I}-\eta_h)^2 = d_1(h)n^{-1} + d_2(h)n^{-2} + o(n^{-2})$, where by expansion with $h \to 0$, $d_1(h) = O(1)$ and $d_2(h) = O(h^{-2p}) + O(h^{-p})$. The $O(h^{-p})$ term of $d_2(h)$ comes from many sources when applying Lemma 2.1 and the $O(h^{-2p})$ term of d_2 comes from $\left[\int B_2(x,x)dF(x)\right]^2$. The $O(n^{-1}h^{-p})$ bias term cannot be improved by an appropriate choice of k. It can be estimated and then subtracted off from $\tilde{I}$, but it can be shown that this would not eliminate the $O(n^{-2}h^{-2p})$ term from the mean squared error, and it may increase the mean squared error. In general, then, $E(\tilde{I}-I)^2 = O(n^{-1}) + O(n^{-2}h^{-2p}) + O(n^{-1}h^{2-p}) + O(h^4)$.

3. Estimation of the integral of the square of the density

In this section, we let $I(f) = \int f^2 d\mu$. With $J(t) = t$ and $L(t) = t^2$, S can be taken to be $\mathscr{R}^p$ and the higher order terms in (2.5) and (2.6) are 0. Let $w_h = k_h * k_h$ be the convolution of k_h with itself. It is straightforward to show that $\tilde{I} = \int \hat{f}^2(x)dx = n^{-2}\sum_i\sum_j w_h(X_i - X_j)$, which is the same as $\hat{I}$ with the kernel w_h (Jones and Sibson (1987) mention that for the normal kernel, $\hat{I}$ with bandwidth h is the same as $\tilde{I}$ with bandwidth $\sqrt{2}h$. This is because if k_h is the normal density with covariance matrix h^2I_p, w_h is the normal density with covariance matrix $2h^2I_p$, where I_p is the identity matrix of order p). Therefore, we will deal only with $\hat{I}$ in this section.

From Section 2,

$$E\hat{I} = (1 - n^{-1}) \int f_h(x) dF(x) + (nh^p)^{-1} k(0) ,$$

since $C_1 \equiv 0$, so that an adjustment of $\hat{I}$ to eliminate some bias is

$$\hat{I}' = n[\hat{I} - (nh^p)^{-1} k(0)]/(n-1) \approx \hat{I} - (nh^p)^{-1} k(0) . \tag{3.1}$$

Alternatively, this bias can be removed by choosing a kernel satisfying $k(0) = 0$. By using (2.2) in (3.1), $\hat{I}' = [n(n-1)]^{-1} \sum\sum_{i \neq j} k_h(X_i - X_j)$, which is a U-statistic for $\int f_h(x) dF(x)$. $\hat{I}'$ can also be written as $n^{-1} \sum_{i=1}^{n} \hat{f}_{-i}(X_i)$, where $\hat{f}_{-i}(X_i) = (n-1)^{-1} \sum_{j \neq i} k_h(X_j - X_i)$ is a cross-validatory estimate of $f(X_i)$.

The remaining bias term depends on h. Suppose Assumption F holds in the remainder of this section; then, $f_h(x) = f(x) + 0.5h^2 \operatorname{tr} f''(x) + o(h^2)$, where $f''(x)$ is the Hessian matrix of second derivatives at x. Therefore,

$$E\hat{I}' = \int f_h dF = \int f dF + 0.5h^2 \int \operatorname{tr} f'' dF + o(h^2) . \tag{3.2}$$

Using the results of Section 2, it can be shown that subtracting off an estimate of the $O(h^2)$ bias term from $\hat{I}'$ leads to an estimate with a larger mean squared error. However, an estimate of the bias can still be used with an estimated variance to get an estimate of the mean squared error. Assuming that k_0 has support on the entire real line and that k_0 is twice differentiable (cf. Schuster (1969)), an estimate of $f''(x)$ is $\hat{f}''(x) = n^{-1} \sum_i k_h''(x - X_i)$, where $k_h''(z) = h^{-(p+2)} k''(z)$. After substitution of this into the $O(h^2)$ bias term in (3.2) and making an adjustment, an estimate of it is

$$[n(n-1)]^{-1} \sum_{i \neq i'} \operatorname{tr} k_h''(X_i - X_{i'}) . \tag{3.3}$$

Given the assumptions on k and f, the expected value of (3.3) is $\int \operatorname{tr} f'' dF + o(1)$, as $h \to 0$.

From Section 2 with substitution of $k(0) = 0$, or from the formula for asymptotic variance of a U-statistic (see Serfling (1980), p. 183), the asymptotic variance of $\hat{I}'$ is

$$4n^{-1} \left[\int f_h^2 dF - \left(\int f_h dF \right)^2 \right] + 2n^{-2} h^{-p} K_2 \int f dF + o(n^{-2} h^{-p}) , \tag{3.4}$$

where $K_2 = \int k^2(u) du$. An unbiased estimate of $\gamma = \int f_h^2 dF$ is $\hat{\gamma} = [n(n -$

$1)(n-2)]^{-1} \sum_{i,j,j' \text{ distinct}} k_h(X_i - X_j)k_h(X_i - X_{j'}) = [n(n-1)(n-2)]^{-1} \sum_i \left\{ \left(\sum_{j \neq i} k_h(X_i - X_j) \right]^2 - \sum_{j \neq i} k_h^2(X_i - X_j) \right\}$. Hence an estimated variance for $\hat{I}'$ is

$$4n^{-1}[\hat{\gamma} - (\hat{I}')^2] + 2n^{-2}h^{-p}K_2\hat{I}' . \tag{3.5}$$

Now we discuss the choice of the bandwidth h. The asymptotic mean squared error of $\hat{I}'$ is the sum of (3.4) and the square of the bias term in (3.2). Therefore, some simple algebra will show that the asymptotically optimal bandwidth is

$$h = \left(\frac{2pK_2\theta}{\beta n^2} \right)^{1/(p+4)}, \tag{3.6}$$

where $\theta = \int f^2 d\mu$ and $\beta = \left(\int \operatorname{tr} f'' dF \right)^2$. With this choice of bandwidth, the mean squared error of $\hat{I}'$ is $O(n^{-1}) + O(n^{-8/(p+4)})$ and the $O(n^{-1})$ term dominates for $p \leq 4$. It can be shown that (3.6) is scale equivariant, that is, a change of the density by a scale factor will change the optimal bandwidth by the same factor. The factor β suggests that for a unimodal density, the optimal bandwidth depends on the peakedness or concentration of mass near the peak.

We come up with a rough rule for a bandwidth based on computation of the optimal bandwidth for some densities. Let IQR denote the interquartile range. For common unimodal densities with a second derivative, $(\theta/\beta)^{0.2}/\text{IQR}$ ranges from 0.9 to 1.26 (it is 1.26 for normal, 0.94 for logistic, 0.95 for Cauchy, 0.99 to 1.21 for Gamma (3) to Gamma (12), and 1.26 to 1.15 for Weibull (3) to Weibull (20)). For a multivariate normal density $\phi(x; \Sigma)$ with zero mean vector and correlation and covariance matrix Σ, $\theta/\beta = 2^{p+2}\pi^{p/2}|\Sigma|^{1/2}/(\operatorname{tr} \Sigma^{-1})^2$; if $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of Σ, then $|\Sigma|^{1/2}/(\operatorname{tr} \Sigma^{-1})^2 = (\lambda_1 \cdots \lambda_p)^{1/2}/(\Sigma \lambda_j^{-1})^2$, so that the optimal bandwidth decreases with more dependence (eigenvalues more spread out). Provided $|\Sigma|$ is not too close to zero, a rough approximation to $(\theta/\beta)^{1/(p+4)}$ is

$$\left[\frac{2^{p+2}\pi^{p/2}}{p^2} \right]^{1/(p+4)} \frac{0.5 - \int_R \phi(x; \Sigma) dx}{0.5 - 0.5^p},$$

where $R = [-0.674, 0.674]^p$. From combining all of the above, a suggested bandwidth for a unimodal density is

$$c \left[\frac{2^{p+3}\pi^{p/2}K_2}{pn^2} \right]^{1/(p+4)} \cdot \frac{\overline{\text{IQR}}}{1.348} \cdot \frac{0.5 - \hat{\alpha}}{0.5 - 0.5^p}, \tag{3.7}$$

where c is between 0.75 and 1, $\overline{\text{IQR}}$ is the average of p interquartile ranges (assumed to be of the same order), $\hat{\alpha}$ is the proportion of data in the rectangle $\times_{j=1}^{p} [q_{1j}, q_{3j}]$ and q_{1j}, q_{3j} are the lower and upper quartiles; for the product normal kernel $k(u) = (2\pi)^{-p/2} \exp\left(-0.5 \sum_{j=1}^{p} u_j^2\right)$, $K_2 = (2\sqrt{\pi})^{-p}$. The last term of (3.7) is omitted for $p = 1$. The use of (3.7) and (3.5) worked well for simulations for the densities mentioned above, and also for densities of transformations of multivariate normal random vectors with univariate marginal distributions all equal to a distribution function G. For a non-unimodal density, measures of scale and concentration should be used relative to a mode. The suggestions here are like those in Silverman ((1986), Subsection 3.4.2).

4. Estimation of the entropy function $-\int f \log f \, d\mu$

In this section, we study estimation of the entropy function $I(f) = \int_S J(f) f \, d\mu$ with $J(t) = -\log t$. The results developed here will be useful for estimating the relative entropies mentioned in Section 1. Using the results of Section 2, it will be shown that a kernel satisfying certain conditions can lead to $\hat{I}$ having improved bias and mean squared error. This improvement is not possible for $\tilde{I}$ and since $\tilde{I}$ is computationally more difficult for $p \geq 2$, we study only $\hat{I}$ in this section.

From (2.7) and Lemma 2.1,

$$E\hat{I} = -\int_S \log f_h \, dF + a_1(h) n^{-1} + a_2(h) n^{-2} + o(n^{-2}) ,$$

where

$$a_1(h) = 0.5 - h^{-p} k(0) \int_S f/f_h \, d\mu + 0.5 h^{-p} K_2 \int_S f_h^* f/f_h^2 \, d\mu , \tag{4.1}$$

$f_h^*(x) = \int l_h(x - y) f(y) dy$, $l_h(u) = h^{-p} l(u/h)$ and l is defined following (2.7),

$$a_2(h) = h^{-2p} \left\{ 0.5 k^2(0) \int_S f/f_h^2 \, d\mu - \frac{1}{3} K_3 \int_S f_h^{**} f/f_h^3 \, d\mu \right\} + O(h^{-p}) , \tag{4.2}$$

$f_h^{**}(x) = \int m_h(x - y) f(y) dy$, $m_h(z) = h^{-p} m(z/h)$, $m(u) = k^3(u)/K_3$ and $K_3 = \int k^3(u) du$. Also, with the use of Lemma 2.1,

$$(4.3)\qquad \left[E\hat{I} + \int_S \log f_h dF\right]^2 = n^{-1}\left\{\int_S [-\log f_h(x) + d(x)]^2 dF(x) - \left(\int_S [-\log f_h(x) + d(x)] dF(x)\right)^2\right\} + O(n^{-2}h^{-p}) + O(n^{-2}h^{-2p}),$$

where $d(x) = -\int_S [f_h(y)]^{-1} k_h(y-x) dF(y)$. The $O(n^{-2}h^{-p})$ term in (4.3) is too complicated to write down. The $O(n^{-2}h^{-2p})$ term comes from $a_1^2(h)/n^2$.

Suppose Assumptions C4 and F hold in the remainder of this section. (4.1) can written as

$$a_1(h) = 0.5 + [0.5K_2 - k(0)]h^{-p}\int_S f/f_h d\mu + 0.5K_2 h^{-p}\int_S (f_h^* - f_h) f/f_h^2 d\mu,$$

and with Assumption C2,

$$\begin{aligned} f_h(x) - f_h^*(x) &= \int [k_h(y-x) - l_h(y-x)] f(y) dy \\ &= \int [k(u) - l(u)] f(x+uh) du \\ &= 0.5h^2 \operatorname{tr} f''(x)\left[1 - \int v^2 l_0(v) dv\right] + o(h^2), \end{aligned}$$

where $l_0(v) = k_0^2(v)/K_{02}$ and $K_{02} = \int k_0^2(v) dv$. Therefore, $a_1(h) = 0.5 + O(h^{2-p})$ if

$$(4.4)\qquad 2^{-1/p}K_{02} - k_0(0) = 0.$$

Furthermore, with Assumption C4, $a_1(h) = 0.5 + O(h^{4-p})$, if $\int |v|^4 k_0(v) dv$ exists, (4.4) is satisfied and

$$(4.5)\qquad \int v^2 k_0^2(v) dv / K_{02} = 1.$$

Similarly, (4.2) can be written as

$$\begin{aligned} a_2(h) = &-[K_3/3 - 0.5k^2(0)]h^{-2p}\int_S f/f_h^2 d\mu \\ &-\frac{1}{3}K_3 h^{-2p}\int_S (f_h^{**} - f_h) f/f_h^3 d\mu + O(h^{-p}), \end{aligned}$$

and $a_2(h) = O(h^{2-2p}) + O(h^{-p})$ if

$$(4.6) \qquad (1.5)^{-1/p} K_{03} - k_0^2(0) = 0 ,$$

where $K_{03} = \int k_0^3(v)dv$.

For a unimodal density k_0, $2^{-1/p} K_{02} < K_{02} \le k_0(0)$, so that a non-unimodal density is needed in order for (4.4) to be satisfied. Equations (4.4) and (4.6) together imply $K_{03} = (0.375)^{1/p} K_{02}^2$. But $K_{02}^2 = \left[\int k_0(v) \cdot k_0(v)dv\right]^2 \le \left[\int k_0^2(v) \cdot k_0(v)dv\right] \cdot \left[\int 1^2 k_0(v)dv\right] = K_{03}$ by the Cauchy-Schwarz inequality, so that (4.4) and (4.6) cannot be simultaneously satisfied. Hence we find kernels satisfying (4.4) and (4.5) to improve on the $O(n^{-1})$ bias term. Table 1 below gives kernels satisfying (4.4) and (4.5) for $p = 1$ to 4. These have the form

$$k_0(v) = \begin{cases} \alpha_1 + \alpha_2 |v|, & |v| \le \beta_1 , \\ \alpha_3 - \alpha_4 |v|, & \beta_1 \le |v| \le \beta_2 . \end{cases}$$

Table 1.

p	β_1	β_2	α_1	α_2	α_3	α_4
1	1.17747	1.86016	0.15584	0.24042	1.19597	0.64294
2	1.31849	1.85802	0.21082	0.13485	1.33829	0.72027
3	1.40397	1.84445	0.23335	0.09345	1.52649	0.82761
4	1.46008	1.83165	0.24568	0.07153	1.72603	0.94227

Suppose Assumption G now holds. The remaining bias term, in (4.7) below, depends on h. With Assumption C2, $f_h(x) = f(x) + 0.5h^2 \operatorname{tr} f''(x) + o(h^2)$, where $f''(x)$ is the Hessian matrix of second derivatives at x, and

$$(4.7) \qquad -\int_S [\log f_h - \log f]dF = 0.5h^2 \int_S \operatorname{tr} f'' d\mu + o(h^2) .$$

With Assumption C4 and if $\int |v|^4 k_0(v)dv$ exists, the two $o(h^2)$ terms can be replaced by $O(h^4)$. From Assumption G, $\int_{\mathscr{R}^p} \operatorname{tr} f'' d\mu = 0$; this can be shown by embedding $f(x)$ in the location family $f(x; v) = f(x_1 - v_1, \ldots, x_p - v_p)$, taking derivatives with respect to v_j, interchanging integration and differentiation—also needed is that if the j-th univariate marginal density f_j has a finite endpoint of upper or lower support x_j^*, then the continuity of the first derivatives of f imply $\partial f(x_1, \ldots, x_{j-1}, x_j^*, x_{j-1}, \ldots, x_p)/\partial x_j = 0$. Therefore, if S is such that $\int_{\mathscr{R}^p - S} \operatorname{tr} f'' d\mu$ is negligible, we take $-\int_S [\log f_h - \log f]dF$

to be $O(h^4)$—that is, the $O(h^2)$ is negligible relative to the other bias terms.

By combining the variance and bias terms, and using a kernel satisfying (4.4) and (4.5), the bias of $\hat{I}$ is $O(n^{-1}h^{4-p}) + O(n^{-2}h^{-2p}) + O(h^4)$ and

$$E(\hat{I} - I)^2 = O(n^{-1}) + O(n^{-2}h^{8-p}) + O(n^{-2}h^{-p}) + O(n^{-1}h^{8-p}) + O(n^{-2}h^{4-2p}) + O(h^8)\,. \tag{4.8}$$

The number of terms here, some of which cannot be simplified, makes the study of a choice of an optimal bandwidth impractical. We now restrict $p \leq 4$ (partly because the sample size n required for estimation grows quickly with increasing p) and propose to choose a bandwidth order such that the $O(n^{-1})$ term in (4.8) is dominant and such that the bias terms are $o(n^{-1/2})$. If $h = O(n^{-1/(0.5p+4)})$, then the bias is $O(n^{-4/(0.5p+4)})$ and the second dominating term in (4.8) is $O(n^{-8/(0.5p+4)})$; with this choice of the order of h, an estimate of the $O(n^{-1})$ variance term can be used for a standard error. In addition, simulations show that the bandwidth should decrease with more dependence for a multivariate density.

From these considerations, we propose the rule

$$cn^{-1/(0.5p+4)} \cdot \overline{\mathrm{IQR}} \cdot \frac{0.5 - \hat{\alpha}}{0.5 - 0.5^p}\,,$$

where c is between 0.75 and 1, $\overline{\mathrm{IQR}}$ and $\hat{\alpha}$ are defined as at the end of Section 3, and the last term is omitted for $p = 1$ (comments at the end of Section 3 concerning unimodality of the density apply here as well). This rule was found to work quite well in simulations for $p \leq 4$ with the kernels in Table 1, for various univariate densities satisfying Assumptions C4 and G, multivariate normal densities and densities of transformations of multivariate normal random vectors with univariate distributions all equal to a distribution function G. The kernels in Table 1 led to estimates $\hat{I}$ that varied much less with h than unimodal kernels. Of course, as p increases, n needs to be larger in order to get reasonable estimates. For example, for $p = 1$, n can be as small as 50, whereas for $p = 3$, at least $n = 200$ is needed for good estimates.

Finally, an estimate of (4.3) is $n^{-1}s^2$, where

$$s^2 = n^{-1} \sum_{X_i \in S} [-\log \hat{f}(X_i) + \hat{d}(X_i)]^2 - \left(\hat{I} + n^{-1} \sum_{X_i \in S} \hat{d}(X_i) \right)^2,$$

and $\hat{d}(x) = -n^{-1} \sum_{X_j \in S} k_h(x - X_j)/\hat{f}(X_j)$. This estimator, which estimates the dominate term of the asymptotic variance, was found in the simulations to be generally 10% to 25% smaller than the variance of $\hat{I}$.

5. Discussion

In this paper, we have shown how representation by an empirical process can be used to study estimation of a functional of a multivariate density based on the kernel density method. The estimation of the integral of the square of the density is easier to study than other functionals. For the entropy function $-\int f \log f d\mu$, non-unimodal kernels satisfying certain conditions can reduce the bias and mean squared error. For other functionals, or with fewer assumptions on f for entropy, subtracting estimates of the first order bias term may reduce the mean squared error. Analysis and simulations show that the sample size needed for good estimates increases rapidly with the dimension p of the multivariate density, but the methods in this paper do work well for small p. This paper just begins to study the area of estimation of a functional of a multivariate density and further research, possibly with other density estimation methods, should lead to improvements.

Acknowledgements

The author is grateful to Dr. M. C. Jones for helpful correspondence and discussion, and to the referees for their useful comments.

REFERENCES

Ahmad, I. A. (1976). On asymptotic properties of an estimate of a functional of a probability density, *Scand. Actuar. J.*, 3, 176–181.

Ahmad, I. A. and Lin, P.-E. (1976). A nonparametric estimation of the entropy for absolutely continuous distributions, *IEEE Trans. Inform. Theory*, **22**, 372–350.

Bhattacharya, G. K. and Roussas, G. G. (1969). Estimation of a certain functional of a probability density function, *Scand. Actuar. J.*, 201–206.

Dmitriev, Yu. G. and Tarasenko, F. P. (1973). On the estimation of functionals of the probability density and its derivatives, *Theory Probab. Appl.*, **18**, 628–633.

Dmitriev, Yu. G. and Tarasenko, F. P. (1974). On a class of nonparametric estimates of nonlinear functionals of density, *Theory Probab. Appl.*, **19**, 390–393.

Huber, P. J. (1985). Projection pursuit, *Ann. Statist.*, **13**, 435–474.

Joe, H. (1987). Majorization, randomness and dependence for multivariate distributions, *Ann. Probab.*, **15**, 1217–1225.

Joe, H. (1989). Relative entropy measures of multivariate dependence, *J. Amer. Statist. Assoc.*, **84**, 157–164.

Jones, M. C. and Sibson, R. (1987). What is projection pursuit?, *J. Roy. Statist. Soc. Ser. A*, **150**, 1–18.

Pawlak, M. (1986). On nonparametric estimation of a functional of a probability density, *IEEE Trans. Inform. Theory*, **32**, 79–84.

Prakasa Rao, B. L. S. (1983). *Nonparametric Functional Estimation*, Academic Press, Orlando, Florida.

Schuster, E. F. (1969). Estimation of a probability density function and its derivatives, *Ann. Math. Statist.*, **40**, 1187–1195.

Schuster, E. F. (1974). On the rate of convergence of an estimate of a functional of a probability density, *Scand. Actuar. J.*, **1**, 101–107.

Schweder, T. (1975). Window estimation of the asymptotic variance of rank estimators of location, *Scand. J. Statist.*, **2**, 113–126.

Serfling, R. J. (1980). *Approximation Theorems in Mathematical Statistics*, Wiley, New York.

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