Estimation of the Information by an Adaptive Partitioning of the Observation Space

Georges A. Darbellay and Igor Vajda, *Senior Member, IEEE*

***Abstract*—We demonstrate that it is possible to approximate the mutual information arbitrarily closely in probability by calculating relative frequencies on appropriate partitions and achieving conditional independence on the rectangles of which the partitions are made. Empirical results, including a comparison with maximum-likelihood estimators, are presented.**

***Index Terms*—Data-dependent partitions, maximum-likelihood estimation, mutual information, nonparametric estimation.**

I. Introduction

We consider a pair of random variables X, Y taking their values in the measurable spaces $\mathcal{X}, \mathcal{Y}$. We are interested in estimates $\hat{I}_n(X;Y)$ of the information

$$I(X;Y) = D(P_{X,Y} \| P_X \times P_Y) = D(f_{X,Y} \| f_X f_Y) = \int_{\mathcal{X}\times\mathcal{Y}} f_{XY} \log \frac{f_{XY}}{f_X f_Y} \quad (1)$$

where $f_{X,Y}$ and f_X, f_Y are the densities of the distributions $P_{X,Y}$ and P_X, P_Y. D denotes the Kullback–Leibler divergence, also known as the relative entropy. In this correspondence we assume for simplicity that $\mathcal{X} = \mathcal{Y} = \mathbb{R}$, but the results can be extended to $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \mathbb{R}^q$. The estimates $\hat{I}_n(X;Y)$ are assumed to be based on n independent realizations $(X_1, Y_1), \cdots, (X_n, Y_n)$ of the pair (X, Y).

If the differential entropies $h(X)[= -\int f_X \log f_X], h(Y)$ and $h(X,Y)$ exist, then the information estimation problem can be reduced to the entropy estimation problem by means of the well-known formula

$$I(X;Y) = h(X) + h(Y) - h(X,Y).$$

If $\hat{h}_n(X)$, $\hat{h}_n(Y)$, and $\hat{h}_n(X,Y)$ are the corresponding estimators of the entropies then $\hat{I}_n(X;Y) = \hat{h}_n(X) + \hat{h}_n(Y) - \hat{h}_n(X,Y)$ satisfies the relation

$$|I(X;Y) - \hat{I}_n(X;Y)| \le |h(X) - \hat{h}_n(X)| + |h(Y) - \hat{h}_n(Y)| + |h(X,Y) - \hat{h}_n(X,Y)|$$

so that the asymptotic unbiasedness, consistency, or the order of consistency of entropy estimates imply analogous properties of $\hat{I}_n(X;Y)$.

There is extensive literature dealing with entropy estimates, overviewed recently by Beirlant *et al.* [2]. Any of the entropy estimates applicable to multidimensional observations can be used for information estimation in the sense explained above. Of all these estimators, the most systematically studied seem to be the histograms associated with products $\mathcal{P}_n \times \mathcal{Q}_n$ of partitions of the marginal spaces $\mathcal{X}$ and $\mathcal{Y}$. More often than not these marginal partitions $\mathcal{P}_n$ and $\mathcal{Q}_n$ are made of intervals of the same width. Partitions $\mathcal{P}_n$ and $\mathcal{Q}_n$ made of equiprobable intervals are also encountered. The disadvantage of using product partitions $\mathcal{P}_n \times \mathcal{Q}_n$ is that different cells $A \times B \in \mathcal{P}_n \times \mathcal{Q}_n$ contribute with a very variable efficiency to the final estimate $\hat{I}_n(X;Y)$. This is obvious in extreme situations, e.g., when the support of f_{XY} is of a lesser dimension than $\mathcal{X} \times \mathcal{Y}$. For example, if $X = \sin U$ and $Y = \cos U$ where U is uniform on $(0, 2\pi)$ then the support of f_{XY} is a unit circle in $\mathbb{R}^2$. If $\mathcal{P}_n = \mathcal{Q}_n$ are uniform partitions of $[-1 - 1/n, 1 + 1/n]$ into intervals of the same width $h_n = 1/n$ then $\mathcal{R}_n = \mathcal{P}_n \times \mathcal{Q}_n$ is a partition of the observation space into $4(n+1)^2$ squares $A \times B$ of area $(1/n)^2$. All squares intersecting the support are inside the circle of radius $1 + 1/n$ and outside the circle of radius $1 - 1/n$. Therefore, the number of these squares is bounded above by

$$\frac{\pi[(1+1/n)^2 - (1-1/n)^2]}{(1/n)^2} = 4\pi n$$

which is less than the (π/n)th part of all squares. Thus for large n, a vast majority of squares from $\mathcal{R}_n$ are not effectively used for the estimation of $I(X;Y)$. These inefficient squares can in fact be replaced by fewer squares or rectangles with the same statistical effect, but not belonging to $\mathcal{R}_n$. Similar situations are typical also for $\mathcal{X} \times \mathcal{Y}$ of higher dimensions where most of the probability mass is concentrated in tails or various hardly predictable "corners" of the observation space (see, e.g., [15, Sec. 1.5]).

A step toward balancing the influence of all partition cells was undertaken by Barron *et al.* [1] who proposed nonuniform and possibly nonproduct partitions into cells of a constant dominating probability rather than of a constant Lebesgue volume. An even more inspiring method (applicable, however, only to one-dimensional spaces) is the finite partitioning by sample quantiles, see [3]; the sample $1/n$-quantiles are equivalent to the spacings discussed in [2], see also [16]. The random grouping of data into m_n cells specified by the j/m_n-quantiles of empirical distribution, $1 \le j \le m_n - 1$, is an example of adaptive partitioning, leading moreover to the uniform distribution of observations into cells (the number of observations in a cell differs by at most one from n/m_n), provided that there are no repeating values, which is a fair assumption for continuous random variables).

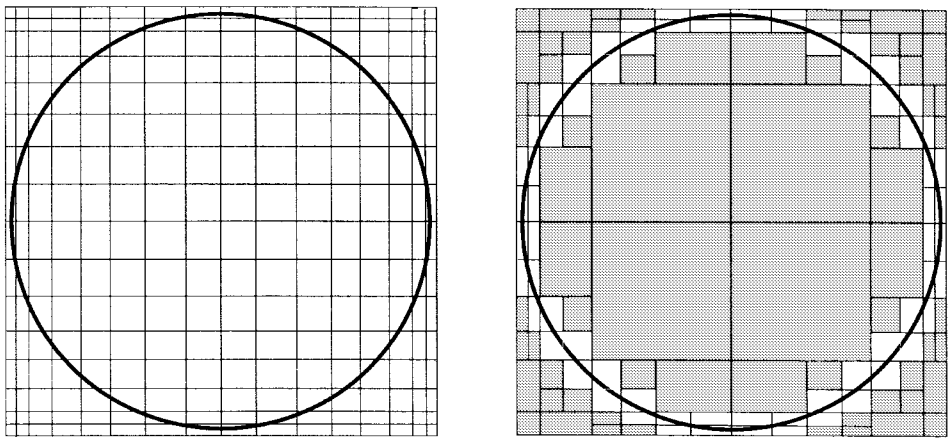
In this correspondence we consider estimates based on the relative frequencies calculated on the cells of adaptive partitions $\mathcal{R}_n$ of $\mathcal{X} \times \mathcal{Y}$, an example of which is shown in Fig. 1(b) for the case of a circle in $\mathbb{R}^2$. The partitions $\mathcal{R}_n$ consist of rectangles $A \times B$ specified by marginal empirical quantiles and they are not of the product type $\mathcal{P}_n \times \mathcal{Q}_n$, an example of which is shown in Fig. 1(a). The partitions $\mathcal{R}_n$ are not uniform in the sense that, in contrast to product partitions, they are not made up of a grid of vertical and horizontal lines, irrespective of whether these lines are equally spaced or not. Note that both the product partition $\mathcal{P}_n \times \mathcal{Q}_n$ of Fig. 1(a) and the partition $\mathcal{R}_n$ of Fig. 1(b) are adaptive, because they both use marginal empirical quantiles. The superior adaptivity of $\mathcal{R}_n$ in Fig. 1(b) comes from the multistep procedure upon which it is built. The number of rectangles in $\mathcal{R}_n$ is kept under control so that the estimator remains relatively simple. In Section II we present a theoretical motivation and description of this estimator. In Section III experiments with data generated from known parametric families are reported. Our conclusions are summarized in Section IV.

Manuscript received April 24, 1998; revised November 11, 1998. This work was supported by the Fonds National Suisse de la Recherche Scientifique and by the EU Grant "Copernicus 579."

The authors are with the Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, 182 08 Prague, Czech Republic.

Communicated by P. Moulin, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier S 0018-9448(99)03551-8.



(a) (b)

Fig. 1. Example where the support of f_{XY} is a circle in $\mathbb{R}^2$. (a) Product partition. It is constructed in a single step. (b) Partition with product cells. It is constructed by a multistep procedure, as described in Section II. Should the partitioning proceed further, the shaded rectangles would remain unaffected.

II. The Estimator

Let us consider a measurable partition $\mathcal{R}$ of $\mathcal{X} \times \mathcal{Y}$ into rectangles of the type $A \times B$ and the densities of the conditional distributions

$$P_{X,Y|A\times B} = P_{X,Y|X\in A, Y\in B}$$

and

$$P_{X|A} = P_{X|X\in A} \qquad P_{Y|B} = P_{Y|Y\in B}$$

$$f_{X,Y|A\times B} = \frac{1_{A\times B} f_{X,Y}}{\int 1_{A\times B} f_{X,Y}} = \frac{1_{A\times B} f_{X,Y}}{P_{X,Y}(A\times B)}$$

and, similarly,

$$f_{X|A} = \frac{1_A f_X}{P_X(A)} \qquad f_{Y|B} = \frac{1_B f_Y}{P_Y(B)}$$

where 1_E denotes the indicator function of the set E. By $P^{\mathcal{R}}_{X,Y}$ and $(P_X \times P_Y)^{\mathcal{R}}$ we denote the restrictions of the corresponding distributions on the σ-algebra generated by $\mathcal{R}$ and we define the *restricted divergence*

$$\begin{aligned} D^{\mathcal{R}}(X;Y) &= D(P^{\mathcal{R}}_{X,Y} \| (P_X \times P_Y)^{\mathcal{R}}) \\ &= \sum_{A\times B\in\mathcal{R}} P_{X,Y}(A\times B) \log \frac{P_{X,Y}(A\times B)}{P_X(A) P_Y(B)}. \end{aligned} \tag{2}$$

In general, $(P_X \times P_Y)^{\mathcal{R}}$ differs from the product of marginals of $P^{\mathcal{R}}_{X,Y}$. This means that $D^{\mathcal{R}}(X;Y)$ cannot be understood as a mutual information. In the special case, where $\mathcal{R}$ is a product partition $\mathcal{R} = \mathcal{P} \times \mathcal{Q}$

$$(P_X \times P_Y)^{\mathcal{R}} = P^{\mathcal{P}}_X \times P^{\mathcal{Q}}_Y$$

holds. In this case one can define the quantized versions $X^{\mathcal{P}}, Y^{\mathcal{Q}}$ of the random variables X, Y by $P_{X^{\mathcal{P}},Y^{\mathcal{Q}}} = P^{\mathcal{R}}_{X,Y}$ and now the marginals of $P^{\mathcal{R}}_{X,Y}$ are $P^{\mathcal{P}}_X$ and $P^{\mathcal{Q}}_Y$. Consequently,

$$D^{\mathcal{R}}(X;Y) = D(P_{X^{\mathcal{P}},Y^{\mathcal{Q}}} \| P_{X^{\mathcal{P}}} \times P_{Y^{\mathcal{Q}}}) = I(X^{\mathcal{P}}; Y^{\mathcal{Q}}).$$

However, for a general partition $\mathcal{R}$ the interpretation of the restricted divergence as an information is impossible.

A similar conclusion applies also to the *residual divergence*

$$\begin{aligned} D_{\mathcal{R}}(X;Y) &= \sum_{A\times B\in\mathcal{R}} P_{X,Y}(A\times B) \\ &\quad \cdot D(P_{X,Y|A\times B} \| (P_X \times P_Y)_{|A\times B}) \\ &= \sum_{A\times B\in\mathcal{R}} P_{X,Y}(A\times B) \int f_{X,Y|A\times B} \log \frac{f_{X,Y|A\times B}}{f_{X|A} f_{Y|B}} \end{aligned} \tag{3}$$

where $(P_X \times P_Y)_{|A\times B}$ denotes the conditional $(P_X \times P_Y)$-distribution on $\mathcal{X} \times \mathcal{Y}$ with the density $f_{X|A} f_{Y|B}$.

We see that the residual divergence is the expected conditional divergence of $P_{X,Y}$ and $P_X \times P_Y$ which remains after specification of the rectangular events yielding the restricted divergence. Consequently, it is nonnegative. The following result is a kind of chain rule for the information divergence (cf., [5, Theorem 2.5.3]).

Proposition 1: For every partition $\mathcal{R}$

$$I(X;Y) = D^{\mathcal{R}}(X;Y) + D_{\mathcal{R}}(X;Y).$$

Proof: By (1) and the additivity of logarithm

$$\begin{aligned} I(X;Y) &= \sum_{A\times B\in\mathcal{R}} \int_{A\times B} f_{X,Y} \log \frac{f_{X,Y}}{f_X f_Y} \\ &= \sum_{A\times B\in\mathcal{R}} \int 1_{A\times B} f_{X,Y} \log \frac{1_{A\times B} f_{X,Y}}{1_A f_X 1_B f_Y} \\ &= \sum_{A\times B\in\mathcal{R}} P_{X,Y}(A\times B) \\ &\quad \cdot \int f_{X,Y|A\times B} \log \frac{f_{X,Y|A\times B} P_{X,Y}(A\times B)}{f_{X|A} f_{Y|B} P_X(A) P_Y(B)} \\ &= \sum_{A\times B\in\mathcal{R}} P_{X,Y}(A\times B) \log \frac{P_{X,Y}(A\times B)}{P_X(A) P_Y(B)} \\ &\quad + \sum_{A\times B\in\mathcal{R}} P_{X,Y}(A\times B) \\ &\quad \cdot \int f_{X,Y|A\times B} \log \frac{f_{X,Y|A\times B}}{f_{X|A} f_{Y|B}} \end{aligned}$$

so that the desired result follows from (2) and (3). □

We say that a sequence of partitions $\{\mathcal{R}^{(k)}, k \in \mathbb{N}\}$ is *nested* if every cell $C \in \mathcal{R}^{(k)}$ is a disjoint union

$$C = \sum_{\ell=1}^{L} C_\ell \tag{4}$$

of cells $C_\ell \in \mathcal{R}^{(k+1)}$, where L varies with C. $\mathcal{R}^{(k+1)}$ is called a refinement or a subpartition of $\mathcal{R}^{(k)}$. If the cells C and C_ℓ are rectangles, we can write

$$C = A \times B = \sum_{\ell=1}^{L} A_\ell \times B_\ell. \tag{5}$$

Obviously, for nested partitions the generated σ-algebras $\mathcal{S}(\mathcal{R}^{(k)})$ are increasing.

A nested sequence $\{\mathcal{R}^{(k)}, k \in \mathbb{N}\}$ is said to be *asymptotically sufficient* for X, Y if for every $\varepsilon > 0$ there exists k_ε such that for every measurable $C \subset \mathcal{X} \times \mathcal{Y}$ one can find $C_0 \in \mathcal{S}(\mathcal{R}^{(k_\varepsilon)})$ satisfying the condition

$$P_{X,Y}(C \triangle C_0) < \varepsilon$$

where $\triangle$ denotes the symmetric difference. If the latter condition holds, then it also holds with $P_{X,Y}$ replaced by $P_X \times P_Y$. If the union of nested partitions $\mathcal{R}^{(k)}$ generates the σ-algebra of all measurable subsets of $\mathcal{X} \times \mathcal{Y}$, then the sequence $\{\mathcal{R}^{(k)}, k \in \mathbb{N}\}$ is asymptotically sufficient for every pair X, Y.

Proposition 2: If a nested sequence of partitions $\mathcal{R}^{(k)}$ is asymptotically sufficient for X, Y, then

$$\lim_{k \to \infty} D^{\mathcal{R}^{(k)}}(X;Y) = I(X;Y) \tag{6}$$

and

$$\lim_{k \to \infty} D_{\mathcal{R}^{(k)}}(X;Y) = 0 \tag{7}$$

where both convergences are monotone.

Proof: By Proposition 1, (6) and (7) are equivalent. The monotone convergence (6) follows from [14, Theorems 1.24 and 1.30]. □

Equation (6) is not a new result—a similar statement is proved, e.g., in Dobrushin [12].

Now, let $\{\mathcal{R}^{(k)}, k \in \mathbb{N}\}$ be a sequence of finite partitions satisfying the assumption of Proposition 2. Then, for $\varepsilon > 0$, there exists k_ε such that

$$D_{\mathcal{R}^{(k_\varepsilon)}}(X;Y) \le \varepsilon. \tag{8}$$

By Proposition 1, this implies

$$D^{\mathcal{R}^{(k_\varepsilon)}}(X;Y) \le I(X;Y) \le D^{\mathcal{R}^{(k_\varepsilon)}}(X;Y) + \varepsilon. \tag{9}$$

We are now ready to turn to the problem of estimating the mutual information from a sample of n independent observations $(X_1, Y_1), \cdots, (X_n, Y_n)$ of the pair (X, Y). Let

$$P_n(C) = \frac{1}{n} \sum_{i=1}^{n} 1_C(X_i, Y_i), \qquad C \subset \mathcal{X} \times \mathcal{Y}$$

be the empirical probability distribution on $\mathcal{X} \times \mathcal{Y}$. We recall that 1_C denotes the indicator function of the set C. We introduce the estimate

$$\hat{D}_{n,k}(X;Y) = \sum_{A \times B \in \mathcal{R}^{(k)}} P_n(A \times B) \cdot \log \frac{P_n(A \times B)}{P_n(A \times \mathbb{R}) P_n(\mathbb{R} \times B)}. \tag{10}$$

It may be rewritten as

$$\hat{D}_{n,k}(X;Y) = \frac{1}{n} \sum_{A \times B \in \mathcal{R}^{(k)}} \sum_{i=1}^{n} 1_{A \times B}(X_i, Y_i) \cdot \log \frac{P_n(A \times B)}{P_n(A \times \mathbb{R}) P_n(\mathbb{R} \times B)} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{P_n(A_i \times B_i)}{P_n(A_i \times \mathbb{R}) P_n(\mathbb{R} \times B_i)}$$

where $A_i \times B_i$ is the cell where (X_i, Y_i) falls. By the law of large numbers and (2)

$$\lim_{n \to \infty} \hat{D}_{n,k} = D^{\mathcal{R}^{(k)}}(X;Y) \quad \text{in probability.} \tag{11}$$

Therefore, it follows from (9) that

$$\lim_{n \to \infty} P(|\hat{D}_{n,k_\varepsilon} - I(X;Y)| < \varepsilon) = 1. \tag{12}$$

A smaller ε will require a larger k_ε, i.e., a finer partition. The last equation says that $\hat{D}_{n,k_\varepsilon}$ is an estimate of the mutual information (though the estimated divergence $\hat{D}$ is not a mutual information as was made clear before). In the sense of (12) we shall write $\hat{I} = \hat{D}$.

Thus we have proved the possibility of approximating the information arbitrarily closely in probability by calculating relative frequencies on appropriate rectangles. As mentioned before in connection with histograms, the problem is to select the appropriate partition. This is usually done *a priori*. Our practical experience has shown that it is far better to specify the partitions *a posteriori*, i.e., by means of a *data-dependent* procedure.

An *adaptive partitioning* $\{\mathcal{R}_n^{(k)}, k \in \mathbb{N}\}$ of $\mathcal{X} \times \mathcal{Y}$ into rectangles and based on the sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ is defined recursively by means of the three steps described below. Its parameters are r, $s \in \{2, 3, \cdots\}$, $s \ge r$, and $\delta > 0$. Once the values of the parameters have been chosen, they remain fixed during the whole partitioning process. The parameter r is used to construct subpartitions and takes a single value. In practice, it often makes sense to choose $r = 2$. The parameter s is used for testing whether a subpartition should be made. We may wish to test on several subpartitions. For this reason, the parameter s may take several values. This applies to δ as well, because it depends on s.

Step 0 Let $\mathcal{R}_n^{(1)} = \mathcal{P}_n \times \mathcal{Q}_n$, where $\mathcal{P}_n$ and $\mathcal{Q}_n$ are partitions of $\mathbb{R}$ into r intervals specified by the marginal sample quantiles

$$a_j = F_n^{-1}(j/r), \qquad 1 \le j \le r-1$$

and

$$b_j = G_n^{-1}(j/r), \qquad 1 \le j \le r-1.$$

$F_n(x) = P_n((-\infty, x) \times \mathbb{R}), G_n(x) = P_n(\mathbb{R} \times (-\infty, x))$ are the marginal distribution functions of P_n. In other words, $\mathcal{R}_n^{(1)}$ is the product of statistically equivalent marginal blocks.

Step 1 $\mathcal{R}_n^{(k+1)}$ contains all rectangles from $\mathcal{R}_n^{(k)}$ not intersecting with the sample. For the remaining rectangles $A \times B \in \mathcal{R}_n^{(k)}$ there are two possibilities. Define at first the conditional marginal distribution functions

$$F_{A,n}(x) = \frac{P_n(((-\infty, x) \cap A) \times \mathbb{R})}{P_n(A \times \mathbb{R})}, \qquad x \in \mathbb{R}$$

and

$$G_{B,n}(x) = \frac{P_n(\mathbb{R} \times ((-\infty, x) \cap B))}{P_n(\mathbb{R} \times B)}, \qquad x \in \mathbb{R}$$

and the partition $\mathcal{R}_{A \times B} = \mathcal{P}_A \times \mathcal{P}_B$ of $A \times B$, where $\mathcal{P}_A$ and $\mathcal{P}_B$ are partitions of A and B specified by the corresponding conditional marginal sample quantiles defined similarly as in Step 1, but with r replaced by $s \ge r$. The decomposition (5) holds for $A_\ell \times B_\ell \in \mathcal{R}_{A \times B}$ and $L = s^2$. If

$$D_{A \times B}(X;Y) \equiv \sum_{\ell=1}^{s^2} \frac{P_n(A_\ell \times B_\ell)}{P_n(A \times B)} \cdot \log \left(\frac{P_n(A_\ell \times B_\ell)}{P_n(A \times B)} \Big/ \frac{P_n(A_\ell \times \mathbb{R}) P_n(\mathbb{R} \times B_\ell)}{P_n(A \times \mathbb{R}) P_n(\mathbb{R} \times B)} \right)$$

is larger than $\delta = \delta_s$, then $\mathcal{R}_n^{(k+1)}$ contains all rectangles from the partition $\mathcal{R}_{A \times B}$ constructed with r (not s)

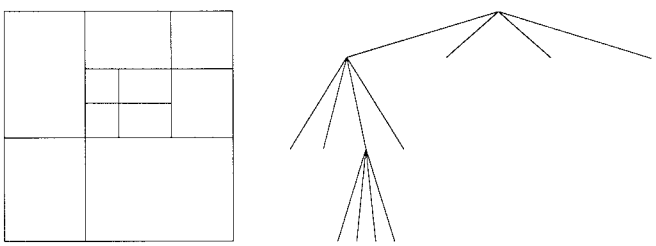


Fig. 2. Example of a partition in $\mathbb{R}^2$, with $r = 2$, and corresponding tree. We used the convention that on each rectangle the subrectangles are labeled from one to four counterclockwise, starting from the top right. At each level, the branches of the tree are labeled from one to four from left to right.

marginal quantiles. Otherwise, $\mathcal{R}_n^{(k+1)}$ contains $A \times B$ itself. In this manner one obtains $\mathcal{R}_n^{(k+1)}$, which is a refinement of $\mathcal{R}_n^{(k)}$.

Step 2 If $\mathcal{R}_n^{(k+1)} \neq \mathcal{R}_n^{(k)}$, then Step 1 is repeated. Otherwise, the process is terminated and $\mathcal{R}_n$ is defined to be $\mathcal{R}_n^{(k)}$.

The *adaptive estimator* $\hat{I}_n(X;Y) = \hat{D}_n(X;Y)$ of $I(X;Y)$ is defined by (10) with $\mathcal{R}_n^{(k)}$ replaced by the adaptive partition $\mathcal{R}_n$. The adaptive partitioning algorithm described above is well suited to being implemented on a computer, as it corresponds to a tree structure—which is not true for every partitioning scheme. The root of the tree is the unpartitioned space. Each node corresponds to a cell and it has r^2 (r^d in $\mathbb{R}^d$) branches. A schematic representation is shown in Fig. 2 for a very simple partition. A leaf is a cell for which the refinement process has been stopped. It is also worth noting that the use of marginal sample quantiles in Steps 0 and 1 above makes the estimator $\hat{I}_n$ invariant with respect to one-to-one transformations of its marginal variables. This means $\hat{I}_n(u(X);v(Y)) = \hat{I}_n(X;Y)$, where u and v are bijective functions.

The statistic $D_{A\times B}(X;Y)$ in Step 1 is used as an estimator of

$$\int f_{X,Y|A\times B} \log \frac{f_{X,Y|A\times B}}{f_{X|A} f_{Y|B}}, \qquad A \times B \in \mathcal{R}_n \tag{13}$$

figuring in (3). The condition $D_{A\times B}(X;Y) \leq \delta$ means that $D_{\mathcal{R}_n}(X;Y)$ will satisfy a condition similar to (8). Since (8) implies (9), $\hat{I}_n(X;Y)$, which estimates in some sense $D^{\mathcal{R}_n}(X;Y)$, will not be too far from $I(X;Y)$. Of course, $D_{A\times B}(X;Y)$ may be replaced by another measure of independence. For example, we could use

$$\chi^2_{A\times B} \equiv \sum_{\ell=1}^{s^2} \frac{P_n(A\times\mathbb{R})P_n(\mathbb{R}\times B)}{P_n(A_\ell\times\mathbb{R})P_n(\mathbb{R}\times B_\ell)} \cdot \left(\frac{P_n(A_\ell\times B_\ell)}{P_n(A\times B)} - \frac{P_n(A_\ell\times\mathbb{R})P_n(\mathbb{R}\times B_\ell)}{P_n(A\times\mathbb{R})P_n(\mathbb{R}\times B)}\right)^2.$$

Note that $D_{A\times B}(X;Y) \leq \log(e)\chi^2_{A\times B}$. This follows easily from the inequality $\log x \leq (\log e)(x-1)$.

The integral (13) vanishes if and only if $f_{X,Y|A\times B} = f_{X|A} f_{Y|B}$, i.e., if and only if the random variables are independent on $A \times B$. This means that the partitioning is stopped when *conditional independence* has been achieved on every cell $A \times B \in \mathcal{R}_n$.

The accuracy of the approximation of $I(X;Y)$ by $\hat{I}_n(X;Y)$ depends on

i) how accurately $\hat{I}_n(X;Y)$ estimates the discrete divergence $D^{\mathcal{R}_n}(X;Y)$, i.e., on the accuracy of the empirical probability distributions in (13). This accuracy is influenced by the choice of the parameter r;

ii) how accurately $D_{A\times B}(X;Y), A \times B \in \mathcal{R}_n$, estimate the integrals (13), which itself depends essentially on how large s is, i.e., on how fine the partitions $\mathcal{R}_{A\times B}$ are when testing for conditional independence;

iii) how small is the parameter δ.

The proof of the consistency of $\hat{I}_n(X;Y)$ for slowly increasing $s = s_n$ and slowly decreasing $\delta = \delta_{s,n}$ seems to be a difficult mathematical problem. Nevertheless, we know of only one type of situation for which the estimator is not consistent. It could happen that on a cell $A \times B \in \mathcal{R}_n$ the data, though being dependent, are distributed symmetrically with respect to the product partition $\mathcal{R}_{A\times B} = \mathcal{P}_A \times \mathcal{P}_B$ specified by the marginal sample quantiles. In this case, the partitioning of the cell $A \times B$ will be stopped, regardless of the number of points n. For such distributions, i.e., those having a symmetry with respect to the marginal sample quantiles, the estimated mutual information will underestimate the true mutual information. The failure of detecting such symmetries is of course more likely for small values of s. One way of reducing this risk of incorrectly stopping the partitioning of a cell $A \times B$ is to apply the independence test to several subpartitions of $A \times B$. This means letting the parameter $s \geq r$ take several values.

In order to guarantee the consistency of nonparametric estimators, smoothness and tail conditions are usually imposed: the density function must be reasonably smooth and its tails not too fat. Our estimator encounters no difficulty at all with slowly decreasing tails, because it uses marginal quantiles. If the density is not smooth, our estimator approaches more slowly the true value than for a smooth density, but it still does.

Our practical experience with the nonparametric estimator described above is excellent. This experience, which has in part already been reported in [6]–[8], shows that the choice of the parameters is not difficult.

The parameter r must be small for the estimator to be adaptive (if r is large the risk of ending up with a product partition is high). It is also clear from i) above that r should be small. We simply choose $r = 2$. For the parameter s we will also choose $s = 2$. But in order to detect possible symmetries with respect to 2×2 partitions, we test a second time by dividing the marginal quantiles once more. In effect, we test with $s = 2$ and with $s = 2^2 = 4$.

The parameters δ_s depend on the choice of the independence test. In Section III, we will use the χ^2 statistic. This means that for stopping the partitioning of a cell $A\times B$ we require $\chi^2_{A\times B} < \delta_{s=2}$ and $\chi^2_{A\times B} < \delta_{s=4}$. For convenience, rather than give the actual values of δ_s which depend on s and the sample size, we will determine them by means of the significance level of the χ^2 test. Except for the partitioning of the root cell, the significance level has no meaning *per se*. It is simply a quantity which conveniently summarizes the values of the δ_s. The significance level for the χ^2 test is determined by matching the estimated mutual informations with their true values, which is done by using distributions for which the mutual information is known analytically. A too high significance level, i.e., too low values for δ_s, will result in $\hat{I}$ overestimating I. Conversely, a too low significance level, i.e., too high values for δ_s, will result in $\hat{I}$ underestimating I. It turns out that a significance level between 1 and 3% works well. For small samples ($n < 500$) one may possibly go up to 5% but not higher. For large samples ($n \approx 10^6$), the significance level should be chosen around 1%. All simulation results reported in the next section were obtained with a significance level of 3%, whatever the value of n.

TABLE I
AVERAGE ESTIMATES OF THE MUTUAL INFORMATION $\mathsf{avg}(\hat{I})$ FOR GAUSSIAN DISTRIBUTIONS. ML REFERS TO THE MAXIMUM-LIKELIHOOD ESTIMATOR, CI TO THE ESTIMATOR BASED ON THE CONDITIONAL INDEPENDENCE CRITERION. THE AVERAGES WERE CALCULATED OVER 1000 ESTIMATES. THE SAMPLE SIZE IS GIVEN AT THE TOP OF THE COLUMN AND RUNS FROM 250 TO 10000 DATA POINTS. THE LAST COLUMN CONTAINS THE THEORETICAL VALUE OF THE MUTUAL INFORMATION. THESE FOUR VALUES CORRESPOND TO THE VALUES $r = 0,\ 0.3,\ 0.6,$ AND 0.9 OF THE COEFFICIENT OF LINEAR CORRELATION r USED TO GENERATE THE SAMPLES

		250	500	1 000	2 000	10 000	I
$r = 0$	$\mathsf{avg}(\hat{I}_{\mathrm{ML}})$	0.0021	0.0010	0.0005	0.0003	0.0001	0
	$\mathsf{avg}(\hat{I}_{\mathrm{CI}})$	0.0001	0	0	0	0	
$r = 0.3$	$\mathsf{avg}(\hat{I}_{\mathrm{ML}})$	0.0497	0.0483	0.0477	0.0477	0.0474	0.0472
	$\mathsf{avg}(\hat{I}_{\mathrm{CI}})$	0.0231	0.0321	0.0402	0.0458	0.0473	
$r = 0.6$	$\mathsf{avg}(\hat{I}_{\mathrm{ML}})$	0.2260	0.2244	0.2238	0.2239	0.2236	0.2231
	$\mathsf{avg}(\hat{I}_{\mathrm{CI}})$	0.1856	0.2047	0.2116	0.2199	0.2234	
$r = 0.9$	$\mathsf{avg}(\hat{I}_{\mathrm{ML}})$	0.8328	0.8315	0.8311	0.8309	0.8309	0.8304
	$\mathsf{avg}(\hat{I}_{\mathrm{CI}})$	0.7329	0.7774	0.7997	0.8144	0.8302	

III. EMPIRICAL RESULTS

In order to assess the quality of our estimator we conducted a series of empirical studies in which we compared our nonparametric estimator to maximum likelihood (ML) estimators. The statistical optimality of ML estimators is well established and we use them as a benchmark. Being a parametric technique, ML estimation is applicable only if (i) the distribution which governs the data is known, (ii) the mutual information of that distribution is known analytically, and (iii) the maximum likelihood equations can be solved for that particular distribution. It is clear that ML estimators have an 'unfair advantage' over any nonparametric estimator which would be applied to data coming from a distribution satisfying the conditions (i), (ii), and (iii). The main objective of this section is to investigate how far our nonparametric estimator is from ML estimators.

We chose four bivariate distributions for which the mutual information is known analytically [10]. These are the Gaussian distribution, the Pareto distribution, the gamma-exponential distribution, and the ordered Weinman exponential distribution. From these four distributions we generated samples of n independent observations (X_i, Y_i). We applied to them both the nonparametric estimator, to which we will refer to as the conditional independence (CI) estimator since it is based on a conditional independence criterion, and the appropriate maximum-likelihood (ML) estimator. The estimates will be denoted by, respectively, $\hat{I}_{\mathrm{CI}}$ and $\hat{I}_{\mathrm{ML}}$. For the numerical calculations, we used the natural logarithm, i.e., $\log = \ln$.

In all the simulations, the averages $\mathsf{avg}(\hat{I})$ and the standard deviations $\mathsf{stdv}(\hat{I})$ were calculated over 1000 estimates. This was done for five different sample sizes namely $n = 250, 500, 1000, 2000, 10000$, and various values of the distribution parameters. Due to space limitations we report results only for the Gaussian distribution. The results for the other three distributions confirm the soundness of our estimator [11].

The bivariate normal probability density function is

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} \cdot \exp\left\{-\frac{1}{2(1-r^2)}\left[\left(\frac{x-\mu_x}{\sigma_x}\right)^2 + \left(\frac{y-\mu_y}{\sigma_y}\right)^2 - 2r\frac{x-\mu_x}{\sigma_x}\frac{y-\mu_y}{\sigma_y}\right]\right\} \quad (14)$$

where the parameters μ_x and μ_y are the expectations of X and Y, σ_x^2 and σ_y^2 the variances of X and Y, and r the coefficient of correlation between X and Y. It is well known that the mutual information is

$$I_{\mathrm{NORMAL}}(X,Y) = -\frac{1}{2}\log(1-r^2). \quad (15)$$

Therefore, to estimate I it suffices to estimate r. The ML estimator of r is

$$\hat{r} = \frac{\sum_{i=1}^{n}(x_i-\hat{\mu}_x)(y_i-\hat{\mu}_y)}{\sqrt{\sum_{i=1}^{n}(x_i-\hat{\mu}_x)^2\sum_{i=1}^{n}(y_i-\hat{\mu}_y)^2}} \quad (16)$$

where

$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n}x_i \qquad \hat{\mu}_y = \frac{1}{n}\sum_{i=1}^{n}y_i. \quad (17)$$

Inserting (16) into (15) yields the sample mutual information $\hat{I}_{\mathrm{ML}}$.

For Gaussian distributions the results are shown in Table I and in Fig. 3. The results displayed correspond to the values $r = 0, 0.3, 0.6,$ and 0.9 of the coefficient of linear correlation appearing in (14). The other parameters were chosen as $\sigma_x = \sigma_y = 1$ and $\mu_x = \mu_y = 0$, but this is irrelevant since $\hat{I}_{\mathrm{ML}}$ and $\hat{I}_{\mathrm{CI}}$ are invariant with respect to shifts and linear scaling. The values of the mutual information I, as obtained from (15), are listed in the last column of Table I. The other columns contain the average mutual information for the sample size $n = 250, 500, 1000, 2000, 10000$. Each average $\mathrm{avg}(\hat{I})$ is calculated over 1000 estimates $\hat{I}$. For the CI estimator it may be seen that $\mathrm{avg}(\hat{I}_{\mathrm{CI}})$ does approach I in the last column as n increases. Except for $r = 0$, the ML estimator is usually better, as expected. For small samples, the underestimating bias of the CI estimator is larger (in absolute value) than the overestimating bias of the ML estimator. Yet, the values of the CI estimator are quite good, and the CI estimator does catch up as n increase: in the penultimate column, the CI estimator performs as well as the ML estimator. The corresponding standard deviations $\mathsf{stdv}(\hat{I}) = (\mathsf{var}(\hat{I}))^{1/2}$ are shown in Fig. 3. Those of the CI estimator are not so much larger than those of the ML estimator. For small samples, $\mathsf{stdv}(\hat{I}_{\mathrm{CI}})$ is typically 20–30% higher than $\mathsf{stdv}(\hat{I}_{\mathrm{ML}})$. For larger samples, it is typically 10–20% higher. For $r = 0$, $\mathsf{stdv}(\hat{I}_{\mathrm{CI}})$ is in fact lower than $\mathsf{stdv}(\hat{I}_{\mathrm{ML}})$.

We also investigated the performance of our nonparametric estimator with respect to two other nonparametric estimators. The first one is based on product partitions (PP), an example of which appeared in Fig. 1(a). It is a grid constructed, in a single step, with marginal empirical quantiles [9]. The second one is a classical histogram (CH)

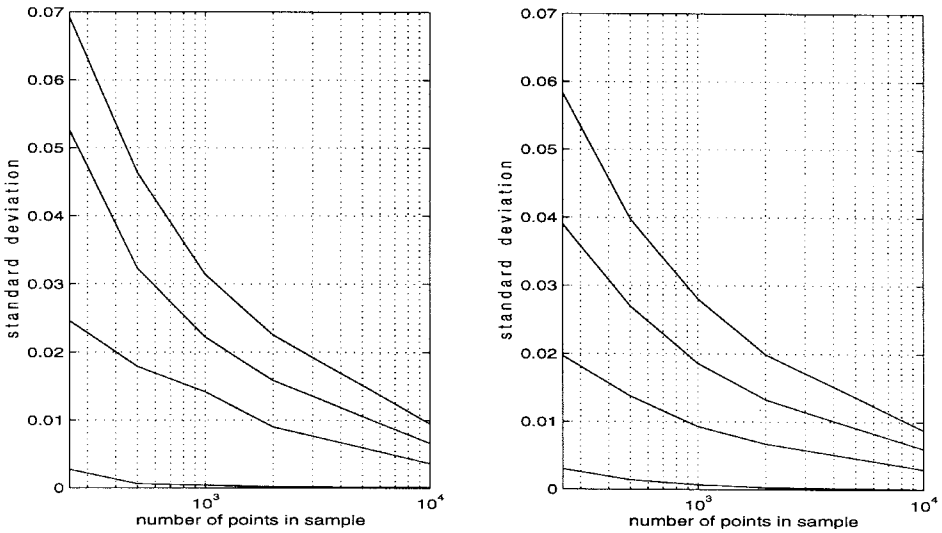


Fig. 3. Standard deviations of $\hat{I}_{\mathrm{CI}}$ and $\hat{I}_{\mathrm{ML}}$ for Gaussian distributions, plotted against the number of points in the sample. A logarithmic scale was used for the number of points. On each graph, the highest curve corresponds to $r = 0.9$, the second highest to $r = 0.6$, the third highest to $r = 0.3$, and the lowest to $r = 0$.

TABLE II
AVERAGE ESTIMATES OF THE MUTUAL INFORMATION avg $(\hat{I})$ FOR GAUSSIAN SAMPLES OF $n = 10000$ DATA POINTS. PP REFERS TO THE PRODUCT PARTITION ESTIMATOR, CH TO A CLASSICAL HISTOGRAM ESTIMATOR. THE RESULTS ARE TO BE COMPARED WITH THE PENULTIMATE COLUMN OF TABLE I

	avg$(\hat{I}_{\mathrm{PP}})$	avg$(\hat{I}_{\mathrm{CH}})$
$r = 0$	0.0181	0.0301
$r = 0.3$	0.0623	0.0698
$r = 0.6$	0.2308	0.2338
$r = 0.9$	0.8010	0.7605

estimator [13]. Both estimators are universally consistent. We found that the variance of the PP estimator is slightly lower, and the variance of the CH estimator slightly higher, than the CI estimator. The bias of these two estimators is far worse than that of our CI estimator. For Gaussian samples of $n = 10000$ data points, with $r = 0, 0.3, 0.6, 0.9$ as above, the results are shown in Table II. A comparison with the penultimate column of Table I speaks for itself. The PP and CH estimators overestimate the mutual information when it is small (weakly dependent X and Y), and they underestimate it when it is large (strongly dependent X and Y). It is impossible to choose the parameters governing these two estimators in such a way that they perform better over the whole spectrum of the dependence (i.e., the range of r for Gaussian distributions) [4], [9]. If the parameters are adjusted so as to decrease the overestimation for weakly dependent X and Y, then the underestimation for strongly dependent X and Y will worsen, and *vice versa*. We also noticed that for non-Gaussian distributions the CH estimator often produces results which are even poorer.

IV. CONCLUSIONS

We have presented a nonparametric estimator of the mutual information based on data-dependent partitions. Unlike its parametric counterparts it is in principle applicable to any distribution. Yet, it is intuitively easy to understand: the partition must be refined until conditional independence has been achieved on its cells. Since the partitioning procedure may be associated to a tree, it lends itself to a computer implementation which can be optimized so as to be very fast—typically the calculation for a sample of 10 000 data pairs takes a fraction of a second on an average PC.

From our empirical study the nonparametric estimator appears to be asymptotically unbiased and efficient. Its variance decreases with the number n of points, and it is of the same order of magnitude as that of the corresponding maximum-likelihood estimator. Except for very low values of the mutual information, the variance is inversely proportional to n. This indicates that the estimator is $\sqrt{n}$-consistent for a reasonably large class of distributions.

To guarantee the consistency of a nonparametric estimator, conditions are usually imposed upon the tails and the smoothness of the density. No such conditions are necessary for our estimator. However, some typical distributions for which our estimator would grossly underestimate the mutual information were identified. To be fair, such distributions are in fact rather exotic. Severe overestimation seems excluded. We stress that the limitations on the consistency are not due to the partitioning procedure itself but to the independence test. A way of extending the consistency would be to use more than one statistical technique when testing for conditional independence.

We also compared our nonparametric estimator to two other nonparametric estimators. We found that the variance of these two estimators is of the same order of magnitude as that of our estimator. However, with regard to the bias, the performance of these two estimators is appalling. In our view, the superiority of the estimator described in this correspondence comes from its adaptivity and its "naturalness": the mutual information is a supremum over partitions.

ACKNOWLEDGMENT

It is a pleasure to thank J. Franěk for his assistance with the programming.

REFERENCES

[1] A. R. Barron, L. Györfi, and E. C. van der Meulen, "Distribution estimation consistent in total variation and in two types of information divergence," *IEEE Trans. Inform. Theory*, vol. 35, pp. 1437–1454, Sept. 1992.

[2] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. van der Meulen, "Nonparametric entropy estimation: An overview," *Int. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, 1997.

[3] E. Bofinger, "Goodness-of-fit using sample quantiles," *J. Roy. Stat. Soc., Ser. B*, vol. 35, pp. 277–284, 1973.

[4] I. Cižova, "Test of a histogram estimator for the differential entropy," Master's thesis, Czech Tech. Univ. (ČVUT), Prague, 1997 (in Czech).
[5] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[6] G. A. Darbellay, "An adaptive histogram estimator for the mutual information," Res. Rep. 1889, UTIA, Academy of Sciences, Prague, Czech Republic, 1996. Also, *Computat. Statist. Data Anal.*, to be published.
[7] ——, "The mutual information as a measure of statistical dependence," in *Proc. Int. Symp. Information Theory* (Ulm, Germany, June 29–July 4, 1997). Piscataway, NJ: IEEE Press, 1997, p. 405.
[8] ——, "Predictability: An information-theoretic perspective," in *Signal Analysis and Prediction*, A. Procházka, J. Uhlíř, P. J. W. Rayner, and N. G. Kingsbury, Eds. Boston, MA: Birkhäuser-Verlag, 1998, pp. 249–262.
[9] ——, "Statistical dependences in $\mathbb{R}^d$: An information-theoretic approach," in *Proc. 3rd European IEEE Workshop Computationally Intensive Methods in Control and Data Processing* (Prague, Czech Republic, Sept. 7–9, 1998). Available via e-mail at library@utia.cas.cz.
[10] G. A. Darbellay and I. Vajda, "Entropy expressions for multivariate continuous distributions," Res. Rep. 1920, UTIA, Academy of Sciences, Prague, Czech Republic, 1998. Available via e-mail at library@utia.cas.cz.
[11] ——, "Estimation of the mutual information with data-dependent partitions," Res. Rep. 1921, UTIA, Academy of Sciences, Prague, Czech Republic, 1998. Available via e-mail at library@utia.cas.cz.
[12] R. L. Dobrushin, "General formulation of Shannon's main theorem in information theory," *Usp. Mat. Nauk*, vol. 14, pp. 3–104, 1959 (in Russian). Translated in *Amer. Math. Soc. Trans.*, vol. 33, pp. 323–438.
[13] L. Györfi and E. C. van der Meulen, "Density-free convergence properties of various estimators of entropy," *Comput. Statist. Data Anal.*, vol. 5, pp. 425–436, 1987.
[14] F. Liese and I. Vajda, *Convex Statistical Distances*. Leipzig, Germany: Teubner, 1987.
[15] D. W. Scott, *Multivariate Density Estimation*. New York: Wiley, 1992.
[16] Y. Shao and M. G. Hahn, "Limit theorems for the logarithm of sample spacings," *Statist. Probab. Lett.*, vol. 24, pp. 121–132, 1995.

Best Asymptotic Normality of the Kernel Density Entropy Estimator for Smooth Densities

Paul P. B. Eggermont and Vincent N. LaRiccia

***Abstract*— In the random sampling setting we estimate the entropy of a probability density distribution by the entropy of a kernel density estimator using the double exponential kernel. Under mild smoothness and moment conditions we show that the entropy of the kernel density estimator equals a sum of independent and identically distributed (i.i.d.) random variables plus a perturbation which is asymptotically negligible compared to the parametric rate $n^{-1/2}$. An essential part in the proof is obtained by exhibiting almost sure bounds for the Kullback–Leibler divergence between the kernel density estimator and its expected value. The basic technical tools are Doob's submartingale inequality and convexity (Jensen's inequality).**

***Index Terms*—Convexity, entropy estimation, kernel density estimators, Kullback–Leibler divergence, submartingales.**

I. Introduction

Let $X_1, X_2, \cdots, X_n$ be independent and identically distributed (i.i.d.) univariate random variables, with common probability density function $g(x)$. Let g^{nh} be the kernel density estimator

$$g^{nh}(x) = \frac{1}{n}\sum_{i=1}^{n} s_h(x - X_i), \qquad x \in \mathbb{R} \tag{1.1}$$

with the *double exponential kernel* $s_h(x) = (2h)^{-1}\exp(-h^{-1}|x|)$. We are interested for practical and theoretical reasons in the estimation of the negative entropy of g

$$H(g) = \int_{\mathbb{R}} g(x)\,\log\, g(x)\,dx \tag{1.2}$$

by the *natural* estimator $H(g^{nh})$ with $h \asymp n^{-\beta}$, for some β with $\frac{1}{4} < \beta < \frac{1}{2}$, depending on the smoothness and decay of g. For some practical applications, see Györfi and van der Meulen [12], Joe [16], and references therein. Our interest in the entropy estimation problem ties in with our attempt at understanding likelihood discrepancy principles for the automatic selection of the window parameter in nonparametric deconvolution problems, see Eggermont and LaRiccia [8].

Under suitable assumptions we prove that

$$H(g^{nh}) = \frac{1}{n}\sum_{i=1}^{n} \log\, g(X_i) + \varepsilon_{nh} \tag{1.3}$$

with $\varepsilon_{nh} = o(n^{-1/2})$ almost surely.

The conditions on g involve smoothness and that g has a finite moment of order >2. If $\mathbb{E}[\{\log g\}^2] < \infty$ then the asymptotic normality of $H(g^{nh})$ is assured by the central limit theorem,

$$\sqrt{n}\,\{H(g^{nh}) - H(g)\} \to_{\rm d}\ Y \sim N(0, \mathrm{Var}[\log g]) \tag{1.4}$$

Manuscript received March 13, 1996; revised November 24, 1998. The material in this correspondence was presented in part at the Conference on Nonparametric Function Estimation, Montreal, Que., Canada, October 13–24, 1997.

The authors are with the Department of Mathematical Sciences, University of Delaware, Newark, DE 19716 USA.

Communicated by K. Zeger, Associate Editor at Large.
Publisher Item Identifier S 0018-9448(99)03764-5.