PROCEEDINGS OF THE ROYAL SOCIETY A

Proc. R. Soc. A (2008) **464**, 1203–1215
doi:10.1098/rspa.2007.0196
Published online 12 February 2008

A computationally efficient estimator for mutual information

By Dafydd Evans*

School of Computer Science, University of Cardiff, 5 The Parade, Cardiff CF24 3AA, UK

Mutual information quantifies the determinism that exists in a relationship between random variables, and thus plays an important role in exploratory data analysis. We investigate a class of non-parametric estimators for mutual information, based on the nearest neighbour structure of observations in both the joint and marginal spaces. Unless both marginal spaces are one-dimensional, we demonstrate that a well-known estimator of this type can be computationally expensive under certain conditions, and propose a computationally efficient alternative that has a time complexity of order $\mathcal{O}(N \log N)$ as the number of observations $N \to \infty$.

Keywords: mutual information; nearest neighbour analysis; non-parametric estimation

1. Introduction

Mutual information quantifies the dependence between two or more random vectors, and has many attractive properties (Cover & Thomas 1991). In particular, it provides a natural criterion for variable selection in regression problems, and has been used to identify appropriate time lags in nonlinear time-series analysis (Kantz & Schreiber 1997). The mutual information $I(X, Y)$ between two random vectors $X \in \mathbb{R}^m$ and $Y \in \mathbb{R}^n$ is defined by

$$I(X, Y) = \mathcal{E}\left(-\log \frac{\phi_z}{\phi_x \phi_y}\right) = -\iint \log\left(\frac{\phi_z(x, y)}{\phi_x(x)\phi_y(y)}\right)\phi_z(x, y)\,\mathrm{d}x\,\mathrm{d}y, \tag{1.1}$$

where ϕ_z is the joint density of the pair $Z=(X, Y)$, and ϕ_x and ϕ_y are the marginal densities of X and Y, respectively. Let $Z_1, \ldots, Z_N$ be a sample of independent and identically distributed observations $Z_i=(X_i, Y_i)$ of the joint variable $Z=(X, Y)$. In this paper, we investigate a class of estimators for mutual information based on the nearest neighbour structure of the observations in both the joint and marginal spaces. Unless both the marginal spaces are one-dimensional $(m,n=1)$, we demonstrate that, under certain conditions, an existing estimator of this type due to Kraskov *et al.* (2004) has computational complexity of order $\mathcal{O}(N^{1+\alpha})$ for some $\alpha > 0$ as the number of observations $N \to \infty$. We also propose an alternative estimator that, at the expense of increased estimation error, has computational complexity of order $\mathcal{O}(N \log N)$ as $N \to \infty$, regardless of the dimension of the marginal spaces.

*d.evans@cs.cardiff.ac.uk

Received 31 August 2007
Accepted 14 January 2008

To model the behaviour of some response vector Y, the first step is to choose an appropriate explanatory vector X, consisting of other observable parameters that, as far as possible, combine to determine the behaviour of Y. In the literature, this is known as feature selection or model identification (Liu & Motoda 1998), and is the central problem in data mining (Witten & Frank 2005). For time-series analysis, a related problem is to determine optimal embeddings for time-delay reconstruction (Kantz & Schreiber 1997).

Suppose we have identified a set of d variables that we think might influence a given response vector. To construct an accurate model, we must choose a subset of these candidate variables whose elements combine to determine the behaviour of the response vector as far as possible. A candidate subset of explanatory variables can be evaluated by estimating the mutual information between it and the response vector. There are 2^d-1 non-trivial candidate subsets, so if d is large it is not feasible to evaluate them all. Instead, an optimization routine must be used to search through the set of candidate subsets, seeking those whose mutual information with respect to the observed response is a maximum. To simplify the model construction process, after choosing a suitable subset of explanatory variables, we should also remove any redundant variables in the subset, i.e. those that depend almost entirely on one or more of the other chosen variables. Mutual information can again be used to quantify this dependence, and this elimination step may require further optimization if the number of candidate variables is large. For successful optimization, it is essential that the evaluation metric used to assess candidate solutions is rapidly computable, in order that a comprehensive search of the solution space can be carried out. Thus, we are motivated to investigate computationally efficient estimators of mutual information.

As we shall see, if both the marginal spaces are one-dimensional, the estimator of Kraskov *et al.* (2004) has the same computational complexity as our proposed alternative, namely of order $\mathcal{O}(N\log N)$ as $N\to\infty$. Kraskov *et al.* (2004) also demonstrate that their estimator performs particularly well if the explanatory and response vectors are independent. Prior to the optimization, the estimator of Kraskov *et al.* (2004) could therefore be used to eliminate those candidate variables that appear to be (almost) independent of the response vector, and thus reduce the dimension of the search space.

2. Estimation

The estimators that we consider are based on the metric properties of nearest neighbour balls in both the joint and marginal spaces. First, we impose a condition on the probability distributions to ensure that the probability measure of a nearest neighbour ball can be bounded in terms of its radius.

(a) *The positive density condition*

Let $F:\mathbb{R}^m\to[0,1]$ be a distribution function whose density function $\phi(x)=F'(x)$ is smooth (i.e. has bounded partial derivatives) every point $x\in\mathbb{R}^m$. For $x\in\mathbb{R}^m$, let $B_x(r)$ denote the ball of radius r centred at x, and let $\omega_x(r)$ and $v_x(r)$ denote its probability measure and volume (Lebesgue measure), respectively,

$$\omega_x(r)=\int_{B_x(r)}\phi(\xi)\,\mathrm{d}\xi \quad \text{and} \quad v_x(r)=\int_{B_x(r)}\mathrm{d}\xi. \tag{2.1}$$

We restrict our attention to distributions that satisfy a *positive density* condition (Gruber 2004), which ensures that the probability measure of an arbitrary ball can be bounded in terms of its radius.

Definition 2.1. A probability distribution satisfies a positive density condition if constants $\beta > 1$ and $\delta > 0$ exist, such that

$$\frac{r^m}{\beta} \le \omega_x(r) \le \beta r^m \quad \text{for all} \quad 0 \le r \le \delta. \tag{2.2}$$

For smooth densities, Evans *et al.* (2002) show that the positive density condition is satisfied if the support $S = \{x \in \mathbb{R}^m : \phi(x) > 0\}$ of the density is a compact convex body in $\mathbb{R}^m$. The positive density condition implies that the samples are drawn from a bounded region in $\mathbb{R}^m$, and thus excludes the possibility that a sample point can take arbitrarily large values. This is acceptable in practical non-parametric data analysis (where the analysis must be performed using only a finite number of empirical observations) because there is nothing to suggest that the variable in question can assume values significantly beyond, for example, the convex hull of the observed samples.

(*b*) *Nearest neighbours*

Let $X_1, \ldots, X_N$ be a sample of independent and identically distributed observations in $\mathbb{R}^m$. For $k \in \mathbb{N}$, let $X_{i(k)}$ denote the kth nearest neighbour of X_i among the sample points $X_1, \ldots, X_N$, where proximity relations are defined relative to the ℓ^∞ norm

$$\|x\| = \max_{1 \le j \le m} |x_j| \quad \text{for} \quad x = (x_1, \ldots, x_m) \in \mathbb{R}^m.$$

Let $B_x(i, k)$ be the kth nearest neighbour ball of X_i, defined to be the ball (hypercube) centred at X_i and having the kth nearest neighbour of X_i on its boundary. Let $\omega_x(i, k)$ and $v_x(i, k)$ denote the probability measure and volume (Lebesgue measure) of $B_x(i, k)$, respectively,

$$\omega_x(i,k) = \int_{B_x(i,k)} \phi_x(\xi)\, \mathrm{d}\xi \quad \text{and} \quad v_x(i,k) = \int_{B_x(i,k)} \mathrm{d}\xi. \tag{2.3}$$

For a sampling distribution having smooth density, it is well known that the probability measure $\omega_x(i, k)$ of the kth nearest neighbour ball of any sample point has a beta distribution with parameters k and $N-k$. In particular,

$$\mathcal{E}(\omega_x(i,k)) = k/N \quad \text{and} \quad \mathcal{E}(\log \omega_x(i,k)) = \psi(k) - \psi(N), \tag{2.4}$$

where ψ is the digamma function $\psi(k) = \Gamma'(k)/\Gamma(k)$.

(*c*) *Entropy estimation*

The (differential) entropy $H(X)$ of a random variable X is defined by

$$H(X) = \mathcal{E}(-\log \phi_x) = -\int \log \phi_x(x) \phi_x(x)\, \mathrm{d}x, \tag{2.5}$$

where ϕ_x is the density of X. Non-parametric methods for entropy estimation are surveyed by Beirlant *et al.* (1997). One approach is to partition the space into a finite set of ‘boxes’, and approximate the integral of (2.5) by the sum

$$\hat{H}_{\text{part}} = -\sum_{j=1}^{M} \widehat{p_j \log p_j}, \tag{2.6}$$

where p_j is the probability measure of the jth box and the index runs over all boxes. The simplest method estimates p_j by the fraction of sample points falling in the jth box, which yields a negatively biased estimator. Grassberger (1988, submitted) presents a number of improved estimators for $p_j \log p_j$, and, in each case, provides an explicit analytic formula for the bias.

Another class of methods estimate $\log \phi(x_i)$ at every sample point x_i, and proceed to approximate the integral of (2.5) by the sample mean

$$\hat{H}_{\text{samp}} = -\frac{1}{N}\sum_{i=1}^{N} \widehat{\log \phi(x_i)}. \tag{2.7}$$

The simplest approach is to estimate $\phi(x_i)$ by the ratio $\omega_{x_i}(r)/v_{x_i}(r)$ for some fixed $r>0$, where $\omega_{x_i}(r)$ and $v_{x_i}(r)$ are defined as in (2.1) and where $\omega_{x_i}(r)$ is estimated by the proportion of sample points contained in the ball $B_{x_i}(r)$. If the radius r is small, the ball is likely to contain relatively few sample points and the estimate is therefore prone to significant sampling error. As r increases the sampling error will decrease, but because the density now becomes increasingly non-uniform over the ball, the estimate becomes increasingly affected by systematic error (bias).

Alternatively, we can estimate $\phi(x_i)$ by the ratio $\omega_x(i,k)/v_x(i,k)$, where $\omega_x(i,k)$ and $v_x(i,k)$ are, respectively, the probability measure and volume (Lebesgue measure) of the kth nearest neighbour ball of x_i, which, by (2.4), yields the estimator

$$\widehat{\log \phi(x_i)} = -\psi(k) + \psi(N) + \log v_x(i,k).$$

For $k=1,2,\ldots$, we thus obtain the entropy estimator of Kozachenko & Leonenko (1987),

$$\hat{H}^{(k)} = -\psi(k) + \psi(N) + \frac{1}{N}\sum_{i=1}^{N} \log v_x(i,k). \tag{2.8}$$

For this estimator, which is also discussed by Wolsztynski *et al.* (2005), the trade-off between sampling error and bias is determined by the size of the kth nearest neighbour ball, and k must therefore be chosen appropriately to ensure that the total error is as small as possible. To determine the bias of $\hat{H}^{(k)}$, we need an analytic expression for the expectation $\mathcal{E}(\log d_x(i,k))$, where $d_x(i,k)$ is the distance from an arbitrary sample point to its kth nearest neighbour in the sample, and the expectation is taken over all realizations of the sample. For smooth sampling densities satisfying (2.2), Evans *et al.* (2002) obtain an asymptotic expression (with an explicit first-order term) for the expectation $\mathcal{E}(d_x(i,k))$ as the number of points $N\to\infty$. It is a matter for conjecture whether a similar approach can yield a similar expression for $\mathcal{E}(\log d_x(i,k))$. Furthermore, the way in which finite sample corrections, similar to those presented by Grassberger (submitted) for partition-based estimators, might be developed for sample mean estimators, such as (2.8), is also an open question.

(*d*) *Mutual information estimation*

Mutual information can be expressed as $I(X, Y)=H(X)+H(Y)-H(X, Y)$, where $H(X, Y)$ is the entropy of the joint variable $Z=(X, Y)$. Thus, for $k_x, k_y, k_z \in \mathbb{N}$, it follows by (1.1) and (2.8) that:

$$I(X, Y) \approx \mathcal{E}\left(\log\left(\frac{\omega_x(i, k_x)\omega_y(i, k_y)}{\omega_z(i, k_z)}\,\frac{v_z(i, k_z)}{v_x(i, k_x)v_y(i, k_y)}\right)\right). \tag{2.9}$$

Let $k \in \mathbb{N}$ be fixed and $Z_{i(k)}=(X_{i(k)}, Y_{i(k)})$ be the kth nearest neighbour of Z_i in the joint sample $(Z_i, \ldots, Z_N)$. Following Kraskov *et al.* (2004), we take $k_x=k_x(i, k)$ to be the number such that X_i is closer to exactly k_x-1 points of the marginal sample $(X_1, \ldots, X_N)$ than it is to $X_{i(k)}$, and $k_y=k_y(i, k)$ to be the number such that Y_i is closer to exactly k_y-1 points of the marginal sample $(Y_1, \ldots, Y_N)$ than it is to $Y_{i(k)}$. Then $\omega_x(i, k_x)$ is the probability measure of the k_xth nearest neighbour ball of X_i in $\mathbb{R}^m$, and $\omega_y(i, k_y)$ is the probability measure of the k_yth nearest neighbour ball of Y_i in $\mathbb{R}^n$. Because the nearest neighbours are computed with respect to the ℓ^∞ norm, we have that

$$v_z(i, k) = v_x(i, k_x)v_y(i, k_y)$$

and hence

$$I(X, Y) \approx \mathcal{E}\left(\log\frac{\omega_x(i, k_x)\omega_y(i, k_y)}{\omega_z(i, k)}\right). \tag{2.10}$$

Thus, by (2.4), for each $k=1, 2, \ldots$ we obtain the following sample mean estimator for mutual information due to Kraskov *et al.* (2004), based on the ratios of probability measures

$$\hat{I}_{\mathrm{P}}^{(k)}(X, Y) = \psi(k)+\psi(N)-\frac{1}{N}\sum_{i=1}^{N}(\psi(k_x(i, k))+\psi(k_y(i, k))). \tag{2.11}$$

As we shall see, under certain conditions the numbers $k_x(i, k)$ and $k_y(i, k)$ will rapidly increase as the number of data points increases, which can have a significant impact on the expected computation time required for this estimator.

Instead of using ratios of probability measures, we can estimate $I(X, Y)$ using volume ratios. In this case, rather than using balls of the same radius but different probability measure, we use balls of the same probability measure but different radius. Thus, we return to (2.9), setting $k_x=k_y=k_z=k$. By (2.4), each of $\log\omega_x(i, k)$, $\log\omega_y(i, k)$ and $\log\omega_z(i, k)$ has an expected value $\psi(k)-\psi(N)$, so (2.9) becomes

$$I(X, Y) \approx \psi(k)-\psi(N)+\mathcal{E}\left(\log\frac{v_z(i, k)}{v_x(i, k)v_y(i, k)}\right).$$

Thus, for each $k=1, 2, \ldots$, we propose the following sample mean estimator for mutual information, based on volume ratios:

$$\hat{I}_{\mathrm{V}}^{(k)}(X, Y) = -\psi(k)+\psi(N)-\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{v_z(i, k)}{v_x(i, k)v_y(i, k)}\right). \tag{2.12}$$

Estimator $\hat{I}_{\mathrm{V}}^{(k)}$ ensures that the number of nearest neighbours to be computed remains fixed, and therefore avoids the possible computational overheads associated with $\hat{I}_{\mathrm{P}}^{(k)}$.

3. Estimating probability measures can be expensive

For a set of N points, the naive approach to finding the k nearest neighbours of every point in the set is to first compute the distance between every pair of points in the set, which has a computation time of asymptotic order $\mathcal{O}(N^2)$ as $N\to\infty$. However, for one-dimensional data we can simply sort the points, following which only $N-k+1$ comparisons are needed to find the k nearest neighbours of every point. The computation time of this method is thus dominated by the sort algorithm, the most efficient of which (e.g. *quicksort*) scales as $\mathcal{O}(N\log N)$ as $N\to\infty$.

For multi-dimensional data, more sophisticated methods are required. For example, methods based on quadtrees (Bentley 1975) first recursively partition the space until each cell contains a small number of sample points, a procedure that has time complexity of order $\mathcal{O}(N\log N)$ as $N\to\infty$. The cell containing an arbitrary point can then be found in time of order $\mathcal{O}(\log N)$, and, having done this, the first k nearest neighbours of the point can be located in time of order $\mathcal{O}(k)$. To find the k nearest neighbours of every point therefore requires a computation time of order $\mathcal{O}(Nk+N\log N)$ as $N\to\infty$. If k is constant with respect to N, this is simply of order $\mathcal{O}(N\log N)$ as $N\to\infty$, but if k increases with N, say $k=N^{\alpha}$ for some $\alpha>0$, the computational cost will be of order $\mathcal{O}(N^{1+\alpha})$ as $N\to\infty$.

In theorem 3.1, we show that if X and Y are independent and satisfy positive density conditions on their distributions, the expected value of k_x, and hence the expected number of near neighbours that must be found for each point in the marginal sample $X_1, \dots, X_N$, grows at least as quickly as N^{α} for some $\alpha>0$ as $N\to\infty$. Hence if either $m>1$ or $n>1$, the time required to compute estimator $\hat{I}_{\mathrm{P}}^{(k)}$ will be of asymptotic order $\mathcal{O}(N^{1+\alpha})$ as $N\to\infty$. In the following, we use the asymptotic notation $f(N)=\Omega(g(N))$ as $N\to\infty$ to denote that $g(N)$ is an asymptotic lower bound for $f(N)$, in the sense that a constant $C>0$ and a number N_0 exist, such that $|f(N)|\geq C|g(N)|$ for all $N>N_0$.

Theorem 3.1. *If $X\in\mathbb{R}^m$ and $Y\in R^n$ are independent random vectors whose distributions satisfy positive density conditions, then*

$$\mathcal{E}(k_x)=\Omega(N^{n/(m+n)}) \quad \text{as} \quad N\to\infty. \tag{3.1}$$

Proof. Let $z=(x,y)\in\mathbb{R}^{m+n}$ be a fixed point, and consider all samples $(Z_1,\dots,Z_N)$ for which $Z_i=z$. Let $i(k)$ be the index of the kth nearest neighbour of Z_i among the sample points $Z_1,\dots,Z_N$, $r_z=\|z-Z_{i(k)}\|$ be the distance from $Z_i=z$ to its kth nearest neighbour in $\mathbb{R}^{m+n}$ and $\omega_z=\omega_z(i,k)$ be the probability measure of the kth nearest neighbour ball $B_z(r_z)$ in $\mathbb{R}^{m+n}$. Similarly, let $r_x=\|x-X_{i(k)}\|$ be the distance from $X_i=x$ to the point $X_{i(k)}$ in $\mathbb{R}^m$, k_x be the number of points $Z_j=(X_j,Y_j)$ for which $\|x-X_j\|\leq r_x$ and $\omega_x=\omega_x(i,k_x)$ be the probability measure of the ball $B_x(r_x)$ in $\mathbb{R}^m$. Then, $X_{i(k)}$ is the k_xth nearest neighbour of $X_i=x$ among the points $X_1,\dots,X_N$ and $B_x(r_x)$ is the k_xth nearest neighbour ball of X_i.

The expected value $\mathcal{E}(k_x)$ depends on the probability measure ω_x of the ball $B_x(r_x)\subset\mathbb{R}^m$, computed with respect to the marginal density ϕ_x. In addition to the fixed point $Z_i=z$, the ball $B_z(r_z)\subset\mathbb{R}^{m+n}$ contains exactly k points (the k nearest neighbours of Z_i), while the remaining $N-k-1$ points are independently and identically distributed in its complement $\mathbb{R}^{m+n}\setminus B_z(r_z)$. Thus, because the points Z_j are independent and identically distributed, the expected number of points k_x lying in the region $B_x(r_x)\times\mathbb{R}^n$ satisfies

$$\mathcal{E}(k_x) = k + (N-k-1)\mathcal{E}(\omega_x-\omega_z).$$

By (2.4) we have $\mathcal{E}(\omega_z)=k/N$, which means that $(N-k-1)\mathcal{E}(\omega_z)\le k$ and hence

$$\mathcal{E}(k_x)\ge (N-k-1)\mathcal{E}(\omega_x). \tag{3.2}$$

Let $F(r)=P(r_z\le r)$ be the distribution of the kth nearest neighbour distance r_z, so that

$$\mathcal{E}(\omega_x) = \int_0^\infty \mathcal{E}(\omega_x|r_z=r)F'(r)\,\mathrm{d}r, \tag{3.3}$$

where $\mathcal{E}(\omega_x|r_z=r)$ is the expected value of ω_x taken over all samples for which $Z_i=z$ and $\|z-Z_{i(k)}\|=r$. To compute a lower bound on $\mathcal{E}(\omega_x|r_z=r)$, let

$$\begin{aligned} G(s) &= P(r_x\le s|r_z\le r)\\ &= P(x'\in B_x(s)|z'\in B_z(r)) \end{aligned}$$

and consider

$$\mathcal{E}(\omega_x|r_z\le r) = \int_0^r \omega_x(s)\,G'(s)\,\mathrm{d}s. \tag{3.4}$$

Given that a point $z'=(x',y')$ lies in the ball $B_z(r)$, the conditional probability $G(s)$ is equal to the conditional probability that z' lies in the region $B_x(s)\times B_y(r)$, given that $z'\in B_z(r)$. Furthermore, because X and Y are independent, and the proximity relations are determined with respect to the ℓ^∞ norm, the probability measure of $B_x(s)\times B_y(r)$ is equal to $\omega_x(s)\omega_y(r)$, and the probability measure of $B_z(r)$ is equal to $\omega_x(r)\omega_y(r)$. Hence for $0\le s\le r$, we have

$$G(s) = \frac{\omega_x(s)\omega_y(r)}{\omega_z(r)} = \frac{\omega_x(s)}{\omega_x(r)}, \tag{3.5}$$

so, by (3.4),

$$\mathcal{E}(\omega_x|r_z\le r) = \int_0^r \omega_x(s)\,G'(s)\,\mathrm{d}s = \frac{1}{2}\omega_x(r). \tag{3.6}$$

Because $\mathcal{E}(\omega_x|r_z\le r)$ is an increasing function of r, this means that

$$\mathcal{E}(\omega_x|r_z=r)\ge\frac{1}{2}\omega_x(r), \tag{3.7}$$

so, by (3.3),

$$\mathcal{E}(\omega_x)\ge\frac{1}{2}\int_0^\infty \omega_x(r)F'(r)\,\mathrm{d}r. \tag{3.8}$$

The distribution function $F(r)=P(r_z\le r)$ is the probability that at least k points Z_j $(j\ne i)$ lie inside the ball $B_z(r)$. Using elementary probabilistic and combinatorial arguments (Evans *et al.* 2002), it can be shown that

$$F'(r) = 2c_1(N,k)\omega_z(r)^{k-1}(1-\omega_z(r))^{N-k-1}\omega_z'(r), \tag{3.9}$$

where

$$c_1(N,k) = \frac{\Gamma(N)}{2\Gamma(N-k)\Gamma(k)}$$

and Γ denotes the Euler gamma function. Hence, by (3.8),

$$\mathcal{E}(\omega_x)\ge c_1(N,k)\int_0^\infty \omega_x(r)\omega_z(r)^{k-1}(1-\omega_z(r))^{N-k-1}\,\mathrm{d}\omega_z(r). \tag{3.10}$$

To evaluate the integral, we note that for any fixed $\delta>0$, the error incurred by neglecting all kth nearest neighbour balls $B_z(r)$ of radius $r>\delta$ (or equivalently, of probability measure $\omega_z(r)>\omega_z(\delta)$) becomes exponentially small as $N\to\infty$. Loosely speaking, this is because the probability that a region of fixed positive measure contains at most some fixed number of points must rapidly approach zero as the number of points increases without bound. To make this precise, let

$$e_N(k) = c_1(N,k)\int_\delta^\infty \omega_x(r)\omega_z(r)^{k-1}(1-\omega_z(r))^{N-k-1}\,\mathrm{d}\omega_z(r). \tag{3.11}$$

Because $\omega_z(r)$ is increasing with respect to r, we have $1-\omega_z(r)\le 1-\omega_z(\delta)$ for all $0\le r\le\delta$. Hence, because $\omega_x(r)\le 1$ and $\omega_z(r)\le 1$, and since $c_1(N,k)\le N^k$, it follows that:

$$|e_N(t)|\le N^k(1-\omega_z(\delta))^{N-k-1}.$$

Let $u=(1-\omega_z(\delta))^{N-k-1}$ so that $\log u=(N-k-1)\log(1-\omega_z(\delta))$. Because $\omega(\delta)$ is fixed, some constant $c>0$ exists, such that $\log(1-\omega_z(\delta))\le -c$ and hence $\log u\le -c(N-k-1)$. This means that $(1-\omega_z(\delta))^{N-k-1}\le \mathrm{e}^{-c(N-k-1)}$ and therefore

$$|e_N(t)|\le N^k\mathrm{e}^{-c(N-k-1)} = \mathcal{O}(1) \quad \text{as} \quad N\to\infty.$$

Thus, it follows that:

$$\mathcal{E}(\omega_x)\ge c_1(N,k)\int_0^\delta \omega_x(r)\omega_z(r)^{k-1}(1-\omega_z(r))^{N-k-1}\,\mathrm{d}\omega_z(r) + \Omega(1) \quad \text{as} \quad N\to\infty. \tag{3.12}$$

The distributions of X and Y are assumed to satisfy positive density conditions, so by (2.2) constants β_x, $\beta_y>1$ and δ_x, $\delta_y>0$ exist, such that

$$\frac{r^m}{\beta_x}\le\omega_x(r)\le\beta_x r^m \quad \text{for all} \quad r\le\delta_x,$$

$$\frac{r^n}{\beta_y}\le\omega_y(r)\le\beta_y r^n \quad \text{for all} \quad r\le\delta_y$$

and hence

$$\frac{r^{m+n}}{\beta_x\beta_y} \le \omega_x(r)\omega_y(r) \le \beta_x\beta_y r^{m+n} \quad \text{for all} \quad r \le \delta = \min\{\delta_x, \delta_y\}. \tag{3.13}$$

From these relations, and using the fact that $\omega_z(r) \le \omega_x(r)\omega_y(r)$, we obtain

$$\omega_x(r) \ge \frac{1}{\beta_x(\beta_x\beta_y)^{m/(m+n)}}\,\omega_z(r)^{m/(m+n)} \quad \text{for all} \quad r \le \delta.$$

Hence, by (3.12),

$$\mathcal{E}(\omega_x) \ge c_2(N,k)\int_0^\delta \omega_z(r)^{k+(m/(m+n))-1}(1-\omega_z(r))^{N-k-1}\,\mathrm{d}\omega_z(r) + \Omega(1) \ \text{ as } \ N \to \infty, \tag{3.14}$$

where

$$c_2(N,k) = \frac{c_1(N,k)}{2\beta_x(\beta_x\beta_y)^{m/(m+n)}}. \tag{3.15}$$

Changing the variable of integration and using a similar argument to that leading up to (3.12), we obtain

$$\mathcal{E}(\omega_x) \ge c_2(N,k)\int_0^1 \omega^{k+(m/(m+n))-1}(1-\omega)^{N-k-1}\,\mathrm{d}\omega + \Omega(1) \ \text{ as } \ N \to \infty.$$

We recognize this as the beta integral

$$\int_0^1 t^{a-1}(1-t)^{b-1}\,\mathrm{d}t = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

with parameters $a = k + (m/(m+n))$ and $b = N-k$. Thus, by (3.10), we obtain

$$\mathcal{E}(\omega_x) \ge c_2(N,k)\frac{\Gamma\left(k+\frac{m}{m+n}\right)}{\Gamma(k)}\frac{\Gamma(N)}{\Gamma\left(N+\frac{m}{m+n}\right)} + \Omega(1) \ \text{ as } \ N \to \infty.$$

Using the asymptotic expansion (Artin 1964),

$$\frac{\Gamma(N)}{\Gamma(N+\sigma)} = N^{-\sigma}\left(1 + \mathcal{O}\left(\frac{1}{N}\right)\right) \ \text{ as } \ N \to \infty,$$

it follows that:

$$\mathcal{E}(\omega_x) \ge c_3(k)N^{-(m/(m+n))} + \Omega(1) \ \text{ as } \ N \to \infty,$$

where $c_3(k)$ is the finite constant (independent of N)

$$c_3(k) = \frac{\Gamma\left(k+\frac{m}{m+n}\right)}{2\Gamma(k)\beta_x(\beta_x\beta_y)^{m/(m+n)}}.$$

Finally, because this bound is independent of the fixed point $Z_i = z$, it follows by (3.2) that:

$$\mathcal{E}(k_x) = \Omega(N^{n/(m+n)}) \ \text{ as } \ N \to \infty,$$

which completes the proof. ■

Corollary 3.2.

$$\mathcal{E}(k_y) = \Omega(N^{m/(m+n)}) \quad \textit{as} \quad N \to \infty. \tag{3.16}$$

4. Experimental results

For multi-variate Gaussian distributions, exact expressions for entropy and mutual information are known, which allow us to evaluate the estimators $\hat{I}_{\mathrm{P}}$ and $\hat{I}_{\mathrm{V}}$ described above. If Z has an $(m+1)$-dimensional Gaussian distribution with covariance matrix Σ_z, it is known (Cover & Thomas 1991) that

$$I(Z_1, \ldots, Z_{m+1}) = (-1/2)\log(\det \Sigma_z). \tag{4.1}$$

Let $X=(Z_1, \ldots, Z_m) \in \mathbb{R}^m$ and $Y=Z_{m+1} \in \mathbb{R}$. Then their mutual information satisfies

$$\begin{aligned} I(X, Y) &= I(Z_1, \ldots, Z_{m+1}) - I(Z_1, \ldots, Z_m) \\ &= (-1/2)\log(\det \Sigma_z / \det \Sigma_x), \end{aligned} \tag{4.2}$$

where Σ_x is the covariance matrix of $X=(Z_1, \ldots, Z_m) \in \mathbb{R}^m$.

We consider three cases: *independent* (A); *weakly dependent* (B); and *strongly dependent* (C) random vectors, each distributed according to the $(m+1)$-dimensional Gaussian distribution with mean zero, unit variance and respective covariance matrices Σ_{A}, Σ_{B} and Σ_{C} defined by

$$\left.\begin{aligned} \Sigma_{\mathrm{A}}(j,k) &= 0 && \text{for } j \neq k, \\ \Sigma_{\mathrm{B}}(j,k) &= 0.5 && \text{for } j \neq k \\ \text{and} \quad \Sigma_{\mathrm{C}}(j,k) &= 0.9 && \text{for } j \neq k. \end{aligned}\right\} \tag{4.3}$$

(*a*) *The estimators* $I_P^{(k)}$ *and* $I_V^{(k)}$

For $m=1$, 2, 3, 4 and for each of the covariance matrices Σ_{A}, Σ_{B} and Σ_{C}, we independently select $N=1000$ points from $\mathbb{R}^{m+1}$, distributed according to the $(m+1)$-dimensional Gaussian distribution having zero mean, unit variance and covariance matrix Σ. In each case, we record the values computed by the estimators $\hat{I}_{\mathrm{P}}^{(k)}$ and $\hat{I}_{\mathrm{V}}^{(k)}$ for every $k=1, \ldots, 20$, repeating the process over 25 independent realizations of the sample. The results are shown in figure 1*a*–*l*.

In each plot, we show the mean estimate sequence computed by $\hat{I}_{\mathrm{P}}^{(k)}$ and $\hat{I}_{\mathrm{V}}^{(k)}$ for $1 \leq k \leq 20$, taken over all 25 sample realizations. The error bars indicate the standard error of the mean taken over these repetitions. In addition, the solid line represents the true value as determined by (4.2), and the dotted line denotes the mean empirical value of $I(X, Y)$ that is also computed according to (4.2), but using the empirical covariance matrix of each sample realization.

As expected, the figures show that both the bias and the variance (and hence the mean squared error) of estimator $\hat{I}_{\mathrm{V}}$ are greater than those of estimator $\hat{I}_{\mathrm{P}}$.

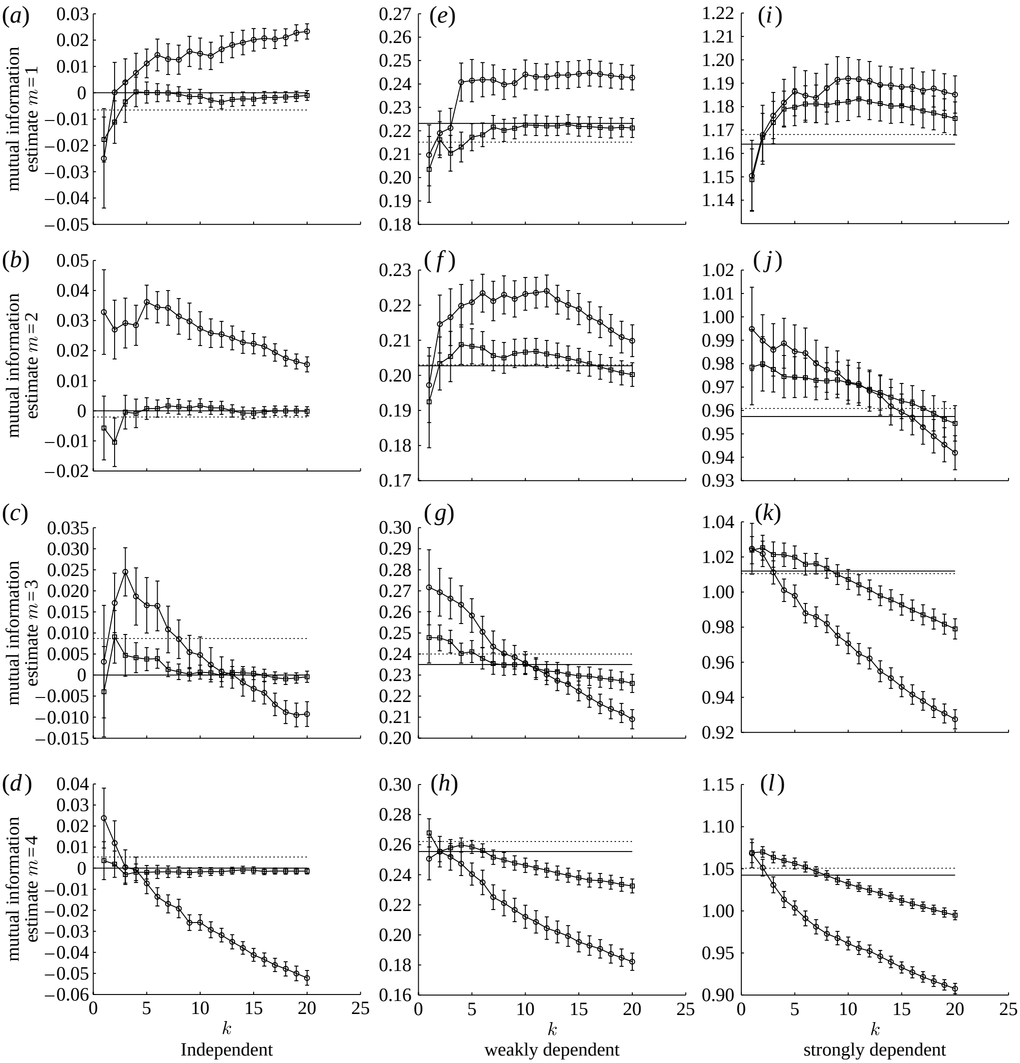

Figure 1. $\hat{I}_{\mathrm{P}}^{(k)}$ and $\hat{I}_{\mathrm{V}}^{(k)}$. Independent (*a*) $m=1$, (*b*) $m=2$, (*c*) $m=3$, (*d*) $m=4$. Weakly dependent (*e*) $m=1$, (*f*) $m=2$, (*g*) $m=3$, (*h*) $m=4$. Strongly dependent (*i*) $m=1$, (*j*) $m=2$, (*k*) $m=3$, (*l*) $m=4$. Squares, $\hat{I}_{\mathrm{prob}}^{(k)}$; circles, $\hat{I}_{\mathrm{vol}}^{(k)}$; solid line, true mutual information; dotted line, empirical mutual information.

(*b*) *Scaling exponents*

Table 1 shows the average values $\langle k_x \rangle$ and $\langle k_y \rangle$ for $k_z=10$ and 20, computed over 25 instances of the joint sample. As one might expect, $\langle k_x \rangle$ and $\langle k_y \rangle$ decrease as the random variables X and Y become increasingly dependent. It is also apparent that $\langle k_x \rangle$ decreases as m increases, but that this is balanced by a corresponding increase in $\langle k_y \rangle$.

Theorem 3.1 states that $k_x \geq N^{\alpha}$ for some $\alpha>0$. To investigate the scaling exponent, we record the average value $\langle k_x \rangle$ for $N=100$, 200, … up to 2500, then use least-squares regression to estimate the gradient of the ($\log N$, $\log k_x$) curve.

Table 1. The empirical values $\langle k_x \rangle$ and $\langle k_y \rangle$ for $N=1000$ points with $k_z=10$ and 20.

		$k_z=10$		$k_z=20$	
m	Σ	$\langle k_x \rangle$	$\langle k_y \rangle$	$\langle k_x \rangle$	$\langle k_y \rangle$
1	Σ_A	106	101	153	147
	Σ_B	92	95	133	137
	Σ_C	57	57	81	82
2	Σ_A	52	218	83	281
	Σ_B	49	192	78	249
	Σ_C	35	117	57	153
3	Σ_A	35	317	59	389
	Σ_B	33	279	55	343
	Σ_C	25	159	42	198
4	Σ_A	28	398	49	469
	Σ_B	25	340	43	405
	Σ_C	21	195	37	238

Table 2. Empirical scaling exponents α_x and α_y for $N=1000$ points with $k_z=10$.

m	Σ	$n/(m+n)$	α_x	$n/(m+n)$	α_y
1	Σ_A	0.50	0.516	0.50	0.519
	Σ_B		0.526		0.490
	Σ_C		0.497		0.496
2	Σ_A	0.33	0.378	0.67	0.672
	Σ_B		0.370		0.670
	Σ_C		0.332		0.676
3	Σ_A	0.25	0.254	0.75	0.794
	Σ_B		0.266		0.764
	Σ_C		0.256		0.742
4	Σ_A	0.20	0.233	0.80	0.818
	Σ_B		0.224		0.806
	Σ_C		0.200		0.782

The empirical scaling exponents, shown in table 2, are close to the theoretical values $\alpha_x=n/(m+n)$ for k_x and $\alpha_y=m/(m+n)$ for k_y derived in theorem 3.1 and corollary 3.2, respectively.

5. Conclusion

We have investigated two non-parametric estimators for mutual information, demonstrated that the method of Kraskov *et al.* (2004) can be computationally expensive under certain circumstances and proposed a more computationally efficient alternative. Experimental results confirm that our estimator is more computationally efficient than the estimator of Kraskov *et al.* (2004), albeit at the expense of increased estimation error. Other estimators for mutual information, for example (Darbellay & Vajda 1999), have appeared in the literature, and there is considerable scope for an extensive numerical investigation into the estimation accuracy and computational cost associated with these techniques. There is also scope to extend the theoretical analysis of such estimators, for example, along the lines developed by Grassberger (submitted).

As a closing remark, in Evans (2007) we prove a weak law of large numbers for functions of a point and its nearest neighbours, where the number of nearest neighbours remains fixed relative to the sample size. This result extends previous research to include unbounded functions of a point and its nearest neighbours, and thus asserts that our sample mean estimator $\hat{I}_{\mathrm{V}}^{(k)}$ is weakly consistent as the number of observations $N \to \infty$.

The author would like to thank the Royal Society for supporting his research through its University Research Fellowship scheme.

References

Artin, E. 1964 *The gamma function*. New York, NY: Holt, Rinehart and Winston.

Beirlant, J., Dudewicz, E. J., Györfi, L. & van der Meulen, E. C. 1997 Non-parametric entropy estimation: an overview. *Int. J. Math. Stat. Sci.* **6**, 17–39.

Bentley, J. L. 1975 Multidimensional binary search trees used for associative search. *Commun. ACM* **18**, 509–517. (doi:10.1145/361002.361007)

Cover, T. M. & Thomas, J. A. 1991 *Elements of information theory*. New York, NY: Wiley.

Darbellay, G. A. & Vajda, I. 1999 Estimation of the information by an adaptive partitioning of the observation space. *IEEE Trans. Inf. Theory* **45**, 1315–1321. (doi:10.1109/18.761290)

Evans, D. 2007 A law of large numbers for nearest neighbour statistics. Preprint, Cardiff University.

Evans, D., Jones, A. J. & Schmidt, W. M. 2002 Asymptotic moments of near-neighbour distance distributions. *Proc. R. Soc. A* **458**, 2839–2849. (doi:10.1098/rspa.2002.1011)

Grassberger, P. 1988 Finite sample corrections to entropy and dimension estimates. *Phys. Lett. A* **128**, 369–373. (doi:10.1016/0375-9601(88)90193-4)

Grassberger, P. Submitted. Entropy estimates from insufficient samplings. (http://arXiv.org/abs/physics/0307138)

Gruber, P. M. 2004 Optimum quantization and its applications. *Adv. Math.* **186**, 456–497. (doi:10.1016/j.aim.2003.07.017)

Kantz, H. & Schreiber, T. 1997 *Nonlinear time series analysis*. Cambridge, UK: Cambridge University Press.

Kozachenko, L. F. & Leonenko, N. N. 1987 Sample estimate of entropy of a random vector. *Probl. Inf. Transm.* **23**, 95–101.

Kraskov, A., Stögbauer, H. & Grassberger, P. 2004 Estimating mutual information. *Phys. Rev. E* **69**, 066 138. (doi:10.1103/PhysRevE.69.066138)

Liu, H. & Motoda, H. 1998 *Feature selection for knowledge discovery and data mining*. Berlin, Germany: Springer.

Witten, I. H. & Frank, E. 2005 *Data mining: practical machine learning tools and techniques*. San Francisco, CA: Morgan Kaufmann.

Wolsztynski, E., Thierry, E. & Pronzato, L. 2005 Minimum-entropy estimation in semi-parametric models. *Signal Process.* **85**, 937–949. (doi:10.1016/j.sigpro.2004.11.028)