PHYSICAL REVIEW E **69**, 066138 (2004)

Estimating mutual information

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger
John-von-Neumann Institute for Computing, Forschungszentrum Jülich, D-52425 Jülich, Germany
(Received 28 May 2003; published 23 June 2004)

We present two classes of improved estimators for mutual information $M(X,Y)$, from samples of random points distributed according to some joint probability density $\mu(x,y)$. In contrast to conventional estimators based on binnings, they are based on entropy estimates from k-nearest neighbor distances. This means that they are data efficient (with $k=1$ we resolve structures down to the smallest possible scales), adaptive (the resolution is higher where data are more numerous), and have minimal bias. Indeed, the bias of the underlying entropy estimates is mainly due to nonuniformity of the density at the smallest resolved scale, giving typically systematic errors which scale as functions of k/N for N points. Numerically, we find that both families become *exact* for independent distributions, i.e. the estimator $\hat{M}(X,Y)$ vanishes (up to statistical fluctuations) if $\mu(x,y)=\mu(x)\mu(y)$. This holds for all tested marginal distributions and for all dimensions of x and y. In addition, we give estimators for redundancies between more than two random variables. We compare our algorithms in detail with existing algorithms. Finally, we demonstrate the usefulness of our estimators for assessing the actual independence of components obtained from independent component analysis (ICA), for improving ICA, and for estimating the reliability of blind source separation.

DOI: 10.1103/PhysRevE.69.066138 PACS number(s): 05.90.+m, 02.50.−r, 87.10.+e

I. INTRODUCTION

Among the measures of independence between random variables, mutual information (MI) is singled out by its information theoretic background [1]. In contrast to the linear correlation coefficient, it is sensitive also to dependences which do not manifest themselves in the covariance. Indeed, MI is zero if and only if the two random variables are strictly independent. The latter is also true for quantities based on Renyi entropies [2], and these are often easier to estimate (in particular if their order is 2 or some other integer >2). Nevertheless, MI is unique in its close ties to Shannon entropy and the theoretical advantages derived from this. Some well-known properties of MI and some simple consequences thereof are collected in the Appendix.

But it is also true that estimating MI is not always easy. Typically, one has a set of N bivariate measurements, $z_i=(x_i,y_i), i=1,\ldots,N$, which are assumed to be iid (independent identically distributed) realizations of a random variable $Z=(X,Y)$ with density $\mu(x,y)$. Here, x and y can be either scalars or can be elements of some higher-dimensional space. In the following, we shall assume that the density is a proper smooth function, although we could also allow more singular densities. All we need is that the integrals written below exist in some sense. In particular, we will always assume that $0\log(0)=0$, i.e., we do not have to assume that densities are strictly positive. The marginal densities of X and Y are $\mu_x(x)=\int dy\,\mu(x,y)$ and $\mu_y(y)=\int dx\,\mu(x,y)$. The MI is defined as

$$I(X,Y)=\int\int dxdy\,\mu(x,y)\log\frac{\mu(x,y)}{\mu_x(x)\mu_y(y)}. \qquad (1)$$

The base of the logarithm determines the units in which information is measured. In particular, taking base 2 leads to information measured in bits. In the following, we always will use natural logarithms. The aim is to estimate $I(X,Y)$ from the set $\{z_i\}$ alone, without knowing the densities μ, μ_x, and μ_y.

One of the main fields where MI plays an important role, at least conceptually, is independent component analysis (ICA) [3,4]. In the ICA literature, very crude approximations to MI based on cumulant expansions are popular because of their ease of use. But they are valid only for distributions close to Gaussians and can mainly be used for ranking different distributions by interdependence, and much less for estimating the actual dependences. Expressions obtained by entropy maximalization using averages of some functions of the sample data as constraints [4] are more robust, but are still very crude approximations. Finally, estimates based on explicit parametrizations of the densities might be useful but are not very efficient. More promising are methods based on kernel density estimators [5,6]. We will not pursue these here either, but we will comment on them in Sec. IV A.

The most straightforward and widespread approach for estimating MI more precisely consists in partitioning the supports of X and Y into bins of finite size, and approximating Eq. (1) by the finite sum

$$I(X,Y)\approx I_{\text{binned}}(X,Y)\equiv\sum_{ij}p(i,j)\log\frac{p(i,j)}{p_x(i)p_y(j)}, \qquad (2)$$

where $p_x(i)=\int_i dx\;\mu_x(x)$, $p_y(j)=\int_j dy\;\mu_y(y)$, and $p(i,j)=\int_i\int_j dxdy\;\mu(x,y)$, and $\int_i$ means the integral over bin i. An estimator of $I_{\text{binned}}(X,Y)$ is obtained by counting the numbers of points falling into the various bins. If $n_x(i)$ [$n_y(j)$] is the number of points falling into the ith bin of X [jth bin of Y], and $n(i,j)$ is the number of points in their intersection, then we approximate $p_x(i)\approx n_x(i)/N$, $p_y(j)\approx n_y(j)/N$, and $p(i,j)\approx n(i,j)/N$. It is easily seen that the right-hand side of Eq. (2) indeed converges to $I(X,Y)$ if we first let $N\to\infty$ and then

let all bin sizes tend to zero, if all densities exist as proper (not necessarily smooth) functions. If not, i.e., if the distributions are, e.g., (multi)fractal, this convergence might no longer be true. In that case, Eq. (2) would define resolution-dependent mutual entropies which diverge in the limit of infinite resolution. Although the methods developed below could be adapted to apply also to that case, we shall not do this in the present paper.

The bin sizes used in Eq. (2) do not need to be the same for all bins. Optimized estimators [7,8] use indeed adaptive bin sizes which are essentially geared to having equal numbers $n(i,j)$ for all pairs (i,j) with nonzero measure. While such estimators are much better than estimators using fixed bin sizes, they still have systematic errors which result on the one hand from approximating $I(X,Y)$ by $I_{\mathrm{binned}}(X,Y)$, and on the other hand by approximating (logarithms of) probabilities by (logarithms of) frequency ratios. The latter could be presumably minimized by using corrections for finite $n_x(i)$ and $n(i,j)$, respectively [9]. These corrections are in the form of asymptotic series which diverge for finite N, but whose first two terms improve the estimates in typical cases. The first correction term—which often is not sufficient—was taken into account in [6,10].

In the present paper we will not follow these lines, but rather estimate MI from k-nearest neighbor statistics. There exists an extensive literature on such estimators for the simple Shannon entropy

$$H(X) = -\int dx \mu(x) \log\, \mu(x), \tag{3}$$

dating back at least to [11,12]. But it seems that these methods have hardly ever been used for estimating MI (for an exception see [13], where they were used to estimate transfer entropies). In [12,14–19] it is assumed that x is one-dimensional, so that the x_i can be ordered by magnitude and $x_{i+1}-x_i \to 0$ for $N \to \infty$. In the simplest case, the estimator based only on these distances is

$$H(X) \approx \frac{1}{N-1} \sum_{i=1}^{N-1} \log(x_{i+1} - x_i) + \psi(1) - \psi(N). \tag{4}$$

Here, $\psi(x)$ is the digamma function, $\psi(x) = \Gamma(x)^{-1} d\Gamma(x)/dx$. It satisfies the recursion $\psi(x+1) = \psi(x) + 1/x$ and $\psi(1) = -C$, where $C = 0.577\,215\,6\ldots$ is the Euler-Mascheroni constant. For large x, $\psi(x) \approx \log x - 1/2x$. Similar formulas exist which use $x_{i+k} - x_i$ instead of $x_{i+1} - x_i$, for any integer $k < N$.

Although Eq. (4) and its generalizations to $k > 1$ seem to give the best estimators of $H(X)$, they cannot be used for MI because it is not obvious how to generalize them to higher dimensions. Here we have to use a slightly different approach, due to [20] [see also [21,22]; the latter authors were only interested in fractal measures and estimating their information dimensions, but the basic concepts are the same as in estimating $H(X)$ for smooth densities].

Assume some metrics to be given on the spaces spanned by X, Y and $Z = (X,Y)$. We can then rank, for each point $z_i = (x_i, y_i)$, its neighbors by distance $d_{i,j} = \|z_i - z_j\|$: $d_{i,j_1} \leqslant d_{i,j_2} \leqslant d_{i,j_3} \leqslant \cdots$. Similar rankings can be done in the subspaces X and Y. The basic idea of [20–22] is to estimate $H(X)$ from the average distance to the k-nearest neighbor, averaged over all x_i. Details will be given in Sec. II. Mutual information could be obtained by estimating in this way $H(X)$, $H(Y)$, and $H(X,Y)$ separately and using [1]

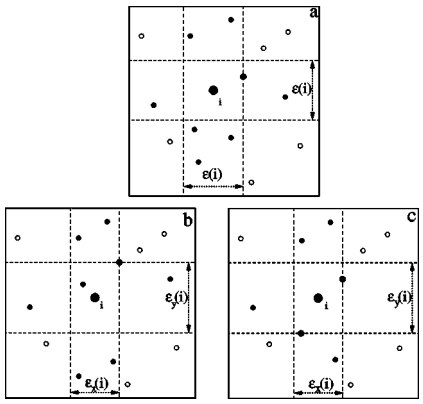

FIG. 1. Panel (a): Determination of $\epsilon(i)$, $n_x(i)$, and $n_y(i)$ in the first algorithm, for $k=1$ and some fixed i. In this example, $n_x(i)=5$ and $n_y(i)=3$. Panels (b),(c): Determination of $\epsilon_x(i)$, $\epsilon_y(i)$, $n_x(i)$, and $n_y(i)$ in the second algorithm for $k=2$. Panel (b) shows a case in which $\epsilon_x(i)$ and $\epsilon_y(i)$ are determined by the same point, while panel (c) shows a case in which they are determined by different points.

$$I(X,Y) = H(X) + H(Y) - H(X,Y). \tag{5}$$

But this would mean that the errors made in the individual estimates would presumably not cancel, and therefore we proceed differently.

Indeed we will present two slightly different algorithms, both based on the above idea. Both use for the space $Z = (X,Y)$ the maximum norm,

$$\|z - z'\| = \max\{\|x - x'\|, \|y - y'\|\}, \tag{6}$$

while any norms can be used for $\|x-x'\|$ and $\|y-y'\|$ (they need not be the same, as these spaces could be completely different). Let us denote by $\epsilon(i)/2$ the distance from z_i to its kth neighbor, and by $\epsilon_x(i)/2$ and $\epsilon_y(i)/2$ the distances between the same points projected into the X and Y subspaces. Obviously, $\epsilon(i) = \max\{\epsilon_x(i), \epsilon_y(i)\}$.

In the first algorithm, we count the number $n_x(i)$ of points x_j whose distance from x_i is strictly less than $\epsilon(i)/2$, and similarly for y instead of x. This is illustrated in Fig. 1(a). Notice that $\epsilon(i)$ is a random (fluctuating) variable, and therefore also $n_x(i)$ and $n_y(i)$ fluctuate. We denote by $\langle \cdots \rangle$ averages both over all $i \in [1, \ldots, N]$ and over all realizations of the random samples,

$$\langle \cdots \rangle = N^{-1} \sum_{i=1}^{N} \mathsf{E}[\cdots(i)]. \tag{7}$$

The estimate for MI is then

$$I^{(1)}(X,Y) = \psi(k) - \langle \psi(n_x + 1) + \psi(n_y + 1) \rangle + \psi(N). \tag{8}$$

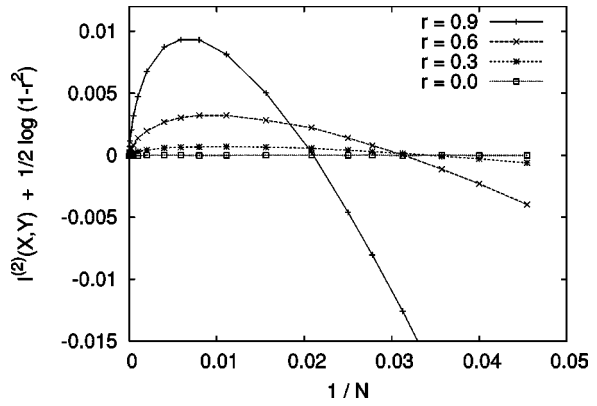

FIG. 2. Estimates of $I^{(2)}(X,Y)-I_{\text{exact}}(X,Y)$ for Gaussians with unit variance and covariances $r=0.9, 0.6, 0.3$, and 0.0 (from top to bottom), plotted against $1/N$. In all cases $k=1$. The number of trials is $>2\times10^6$ for $N\leqslant1000$ and decreases to $\approx10^5$ for $N=40\,000$. Error bars are smaller than the sizes of the symbols.

Alternatively, in the second algorithm, we replace $n_x(i)$ and $n_y(i)$ by the number of points with $\|x_i-x_j\|\leqslant\epsilon_x(i)/2$ and $\|y_i-y_j\|\leqslant\epsilon_y(i)/2$ [see Figs. 1(b) and 1(c)]. The estimate for MI is then

$$I^{(2)}(X,Y) = \psi(k) - 1/k - \langle\psi(n_x) + \psi(n_y)\rangle + \psi(N). \tag{9}$$

The derivations of Eqs. (8) and (9) will be given in Sec. II. There we will also give formulas for generalized redundancies in higher dimensions,

$$I(X_1,X_2,\ldots,X_m) = H(X_1) + H(X_2) + \cdots + H(X_m) - H(X_1,X_2,\cdots,X_m). \tag{10}$$

In general, both formulas give very similar results. For the same k, Eq. (8) gives slightly smaller statistical errors [because $n_x(i)$ and $n_y(i)$ tend to be larger and have smaller relative fluctuations], but have larger systematic errors. The latter is only severe if we are interested in very high dimensions where $\epsilon(i)$ tends typically to be much larger than the marginal $\epsilon_{x_j}(i)$. In that case the second algorithm seems preferable. Otherwise, both can be used equally well.

A systematic study of the performance of Eqs. (8) and (9) and comparison with previous algorithms will be given in Sec. III. Here we will just show results of $I^{(2)}(X,Y)$ for Gaussian distributions. Let X and Y be Gaussians with zero mean and unit variance, and with covariance r. In this case $I(X,Y)$ is known exactly [8],

$$I_{\text{Gauss}}(X,Y) = -\frac{1}{2}\log(1-r^2). \tag{11}$$

In Fig. 2, we show the errors $I^{(2)}(X,Y)-I_{\text{Gauss}}(X,Y)$ for various values of r, obtained from a large number (typically 10^5-10^7) of realizations of N-tuples of vectors (x_i,y_i). We show only results for $k=1$, plotted against $1/N$. Results for $k>1$ are similar. To a first approximation $I^{(1)}(X,Y)$ and $I^{(2)}\times(X,Y)$ depend only on the ratio k/N.

The most conspicuous feature seen in Fig. 2, apart from the fact that indeed $I^{(2)}(X,Y)-I_{\text{Gauss}}(X,Y)\to0$ for $N\to\infty$, is that the systematic error is compatible with zero for $r=0$, i.e., when the two Gaussians are uncorrelated. We checked this with high statistics runs for many different values of k and N (*a priori* one should expect that systematic errors become large for very small N), and for many more distributions (exponential, uniform, etc.). In all cases we found that both $I^{(1)}(X,Y)$ and $I^{(2)}(X,Y)$ become exact for independent variables. Moreover, the same seems to be true for higher-order redundancies. We thus have the following conjecture.

Conjecture. Equations (8) and (9) are exact for independent X and Y, i.e., $I^{(1)}(X,Y)=I^{(2)}(X,Y)=0$ if and only if $I(X,Y)=0$.

We have no proof for this very surprising result. We have numerical indications that moreover

$$\frac{|I^{(1,2)}(X,Y)-I(X,Y)|}{I(X,Y)} \leqslant \text{const} \tag{12}$$

as X and Y become more and more independent, but this is much less clean and therefore much less sure.

In Sec. II we shall give formal arguments for our estimators, and for generalizations to higher dimensions. Detailed numerical results for cases where the exact MI is known will be given in Sec. III. In Sec. IV A we give two preliminary applications to gene expression data and to ICA. Conclusions are drawn in the final section, Sec. V. Finally, some general aspects of MI are recalled in an Appendix.

II. FORMAL DEVELOPMENTS

A. Kozachenko-Leonenko estimate for Shannon entropies

We first review the derivation of the Shannon entropy estimate [20–23], since the estimators for MI are obtained by very similar arguments.

Let X be a continuous random variable with values in some metric space, i.e., there is a distance function $\|x-x'\|$ between any two realizations of X, and let the density $\mu(x)$ exist as a proper function. Shannon entropy is defined as

$$H(X) = -\int dx\mu(x)\log\,\mu(x), \tag{13}$$

where "log" will always mean natural logarithm so that information is measured in natural units. Our aim is to estimate $H(X)$ from a random sample $(x_1\cdots x_N)$ of N realizations of X.

The first step is to realize that Eq. (13) can be understood (up to the minus sign) as an average of $\log\,\mu(x)$. If we had unbiased estimators $\widehat{\log\,\mu(x)}$ of the latter, we would have an unbiased estimator

$$\hat{H}(X) = -N^{-1}\sum_{i=1}^{N}\widehat{\log\,\mu(x_i)}. \tag{14}$$

In order to obtain the estimate $\widehat{\log\,\mu(x_i)}$, we consider the probability distribution $P_k(\epsilon)$ for the distance between x_i and its kth nearest neighbor. The probability $P_k(\epsilon)d\epsilon$ is equal to the chance that there is one point within distance r

$\in[\epsilon/2,\epsilon/2+d\epsilon/2]$ from x_i, that there are $k-1$ other points at smaller distances, and that $N-k-1$ points have larger distances from x_k. Let us denote by p_i the mass of the ϵ ball centered at x_i, $p_i(\epsilon)=\int_{\|\xi-x_i\|<\epsilon/2} d\xi\mu(\xi)$. Using the trinomial formula we obtain

$$P_k(\epsilon)d\epsilon = \frac{(N-1)!}{1!\,(k-1)!\,(N-k-1)!}\frac{dp_i(\epsilon)}{d\epsilon}d\epsilon \times p_i^{k-1} \times (1-p_i)^{N-k-1} \tag{15}$$

or

$$P_k(\epsilon) = k\binom{N-1}{k}\frac{dp_i(\epsilon)}{d\epsilon}p_i^{k-1}(1-p_i)^{N-k-1}. \tag{16}$$

One easily checks that this is correctly normalized, $\int d\epsilon P_k(\epsilon)=1$. Using Eq. (16), one can also compute the expectation value of $\log p_i(\epsilon)$,

$$\begin{aligned}\mathsf{E}(\log p_i) &= \int_0^\infty d\epsilon P_k(\epsilon)\log p_i(\epsilon)\\ &= k\binom{N-1}{k}\int_0^1 dp p^{k-1}(1-p)^{N-k-1}\log p\\ &= \psi(k)-\psi(N),\end{aligned} \tag{17}$$

where $\psi(x)$ is the digamma function. The expectation is taken here over the positions of all other $N-1$ points, with x_i kept fixed. An estimator for $\log \mu(x)$ is then obtained by assuming that $\mu(x)$ is constant in the entire ϵ ball. The latter gives

$$p_i(\epsilon) \approx c_d \epsilon^d \mu(x_i), \tag{18}$$

where d is the dimension of x and c_d is the volume of the d-dimensional unit ball. For the maximum norm one has simply $c_d=1$, while $c_d=\pi^{d/2}/\Gamma(1+d/2)/2^d$ for the Euclidean norm.

Using Eqs. (17) and (18), one obtains

$$\log \mu(x_i) \approx \psi(k)-\psi(N)-d\mathsf{E}(\log\epsilon)-\log c_d, \tag{19}$$

which finally leads to

$$\hat{H}(X) = -\psi(k)+\psi(N)+\log c_d + \frac{d}{N}\sum_{i=1}^N \log\epsilon(i), \tag{20}$$

where $\epsilon(i)$ is twice the distance from x_i to its kth neighbor.

From the derivation it is obvious that Eq. (20) would be unbiased, if the density $\mu(x)$ were strictly constant. The only approximation is in Eq. (18). For points on a torus (e.g., when x is a phase) with a strictly positive density one can easily estimate the leading corrections to Eq. (18) for large N. One finds that they are $O(1/N^2)$ and that they scale, for large k and N, as $\sim(k/N)^2$. In most other cases (including, e.g., Gaussians and uniform densities in bounded domains with a sharp cutoff) it seems numerically that the error is $\sim k/N$ or $\sim k/N\log(N/k)$.

B. Mutual information: Estimator $I^{(1)}(X,Y)$

Let us now consider the joint random variable $Z=(X,Y)$ with maximum norm. Again we take one of the N points z_i and consider the distance $\epsilon/2$ to its kth neighbor. Again this is a random variable with distribution given by Eq. (16). Also Eq. (17) holds without changes. The first difference from the previous subsection is in Eq. (18), where we have to replace d by $d_Z=d_X+d_Y$, c_d by $c_{d_X}c_{d_Y}$, and of course x_i by $z_i=(x_i,y_i)$. With these modifications we obtain therefore

$$\hat{H}(X,Y) = -\psi(k)+\psi(N)+\log(c_{d_X}c_{d_Y}) + \frac{d_X+d_Y}{N}\sum_{i=1}^N \log\epsilon(i). \tag{21}$$

In order to obtain $I(X,Y)$, we have to subtract this from estimates for $H(X)$ and $H(Y)$. For the latter, we could use Eq. (20) directly with the same k. But this would mean that we would effectively use different distance scales in the joint and marginal spaces. For any fixed k, the distance to the kth neighbor in the joint space will be larger than the distances to the neighbors in the marginal spaces. The bias in Eq. (20) results from the nonuniformity of the density. Since the effect of the latter depends of course on the kth neighbor distances, the biases in $\hat{H}(X)$, $\hat{H}(Y)$, and in $\hat{H}(X,Y)$ would be very different and would thus not cancel.

To avoid this, we notice that Eq. (20) holds for *any* value of k, and that we do not have to choose a fixed k when estimating the marginal entropies. Assume, as in Fig. 1(a), that the kth neighbor of x_i is on one of the vertical sides of the square of size $\epsilon(i)$. In this case, if there are altogether $n_x(i)$ points within the vertical lines $x=x_i\pm\epsilon(i)/2$, then $\epsilon(i)/2$ is the distance to the $[n_x(i)+1]$st neighbor of x_i, and

$$\hat{H}(X) = \frac{-1}{N}\sum_{i=1}^N \psi[n_x(i)+1] + \psi(N) + \log c_{d_X} + \frac{d_X}{N}\sum_{i=1}^N \log\epsilon(i). \tag{22}$$

For the other direction [the y direction in Fig. 1(a)] this is not exactly true, i.e., $\epsilon(i)$ is not exactly equal to twice the distance to the $[n_y(i)+1]$st neighbor, if $n_y(i)$ is analogously defined as the number of points with $\|y_j-y_i\|<\epsilon(i)/2$. Nevertheless, we can consider Eq. (22) also as a good approximation for $H(Y)$, if we replace everywhere X by Y in its right-hand side [this approximation becomes exact when $n_y(i)\to\infty$, and thus also when $N\to\infty$]. If we do this, subtracting $\hat{H}(X,Y)$ from $\hat{H}(X)+\hat{H}(Y)$ leads directly to Eq. (8).

We should stress that the errors in $\hat{H}(X)$, $\hat{H}(Y)$, and in $\hat{H}(X,Y)$ will not cancel eactly in general. But the chances that they will do so approximately are bigger with the above procedure than if we had used different length scales in the three estimates. The real proof that our proposed estimator is better than that obtained when using the same k in $\hat{H}(X)$, $\hat{H}(Y)$, and $\hat{H}(X,Y)$ comes of course from detailed numerical tests.

These arguments can be easily extended to m random variables and lead to

$$I^{(1)}(X_1,X_2,\ldots,X_m) = \psi(k) + (m-1)\psi(N) - \langle \psi(n_{x_1}) + \psi(n_{x_2}) + \cdots + \psi(n_{x_m})\rangle. \quad (23)$$

C. Mutual information: Estimator $I^{(2)}(X,Y)$

The main drawback of the above derivation is that the Kozachenko-Leonenko estimator is used correctly in only one marginal direction. This seems unavoidable if one wants to stick to "balls," i.e., to (hyper-) cubes in the joint space. In order to avoid it we have to switch to (hyper) rectangles.

Let us first discuss the case of two marginal variables X and Y, and generalize later to m variables $X_1,\ldots,X_m$. As illustrated in Figs. 1(b) and 1(c), there are two cases to be distinguished [all other cases, where more points fall onto the boundaries $x_i \pm \epsilon_x(i)/2$ and $y_i \pm \epsilon_y(i)/2$, have zero probability; see, however, the third paragraph of Sec. III]: Either the two sides $\epsilon_x(i)$ and $\epsilon_y(i)$ are determined by the same point [Fig. 1(b)], or by different points [Fig. 1(c)]. In either case we have to replace $P_k(\epsilon)$ by a two-dimensional density,

$$P_k(\epsilon_x,\epsilon_y) = P_k^{(b)}(\epsilon_x,\epsilon_y) + P_k^{(c)}(\epsilon_x,\epsilon_y) \quad (24)$$

with

$$P_k^{(b)}(\epsilon_x,\epsilon_y) = \binom{N-1}{k} \frac{d^2[q_i^k]}{d\epsilon_x d\epsilon_y}(1-p_i)^{N-k-1} \quad (25)$$

and

$$P_k^{(c)}(\epsilon_x,\epsilon_y) = (k-1)\binom{N-1}{k} \frac{d^2[q_i^k]}{d\epsilon_x d\epsilon_y}(1-p_i)^{N-k-1}. \quad (26)$$

Here, $q_i \equiv q_i(\epsilon_x,\epsilon_y)$ is the mass of the rectangle of size $\epsilon_x \times \epsilon_y$ centered at (x_i,y_i), and p_i is, as before, the mass of the square of size $\epsilon = \max\{\epsilon_x,\epsilon_y\}$. The latter is needed since by using the maximum norm we guarantee that there are no points in this square which are not inside the rectangle.

Again we verify straightforwardly that P_k is normalized, while we have now instead of Eq. (17)

$$\mathrm{E}(\log q_i) = \int \int_0^\infty d\epsilon_x d\epsilon_y P_k(\epsilon_x,\epsilon_y) \log\, q_i(\epsilon_x,\epsilon_y) = \psi(k) - 1/k - \psi(N). \quad (27)$$

Denoting now by $n_x(i)$ and $n_y(i)$ the number of points with distance less than *or equal* to $\epsilon_x(i)/2$ and $\epsilon_y(i)/2$, respectively, we arrive at Eq. (9).

For the generalization to m variables we have to consider m-dimensional densities $P_k(\epsilon_{x_1},\ldots,\epsilon_{x_m})$. The number of distinct cases [analogous to the two cases shown in Figs. 1(b) and 1(c)] proliferates as m grows, but fortunately we do not have to consider all these cases explicitly. One sees easily that each of them contributes to P_k a term

$$\propto \frac{d^m[q_i^k]}{d\epsilon_{x_1}\cdots d\epsilon_{x_m}}(1-p_i)^{N-k-1}. \quad (28)$$

The direct calculation of the proportionality factors would be extremely tedious (we did it for $m=3$), but it can be avoided by simply demanding that the sum is correctly normalized. This gives

$$P_k(\epsilon_{x_1},\ldots,\epsilon_{x_m}) = k^{m-1}\binom{N-1}{k} \frac{d^m[q_i^k]}{d\epsilon_{x_1}\cdots d\epsilon_{x_m}} \times (1-p_i)^{N-k-1}. \quad (29)$$

Calculating again $\mathrm{E}(\log q_i) = \psi(k) - (m-1)/k - \psi(N)$ analytically and approximating the density by a constant inside the hyper-rectangle, we obtain finally

$$I^{(2)}(X_1,X_2,\ldots,X_m) = \psi(k) - (m-1)/k + (m-1)\psi(N) - \langle \psi(n_{x_1}) + \psi(n_{x_2}) + \cdots + \psi(n_{x_m})\rangle. \quad (30)$$

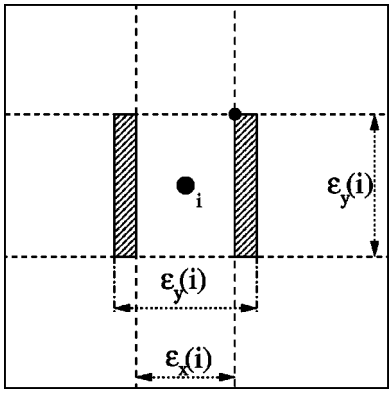

FIG. 3. There cannot be any points inside the shaded rectangles. For method 2, this means that the estimates of the marginal entropy $H(X)[H(Y)]$ should be modified, since part of the area outside [inside] the stripe of with ϵ_x [ϵ_y] is forbidden. This is neglected in Eq. (9).

Before leaving this section, we should mention that we cheated slightly in deriving $I^{(2)}(X,Y)$ (and its generalization to $m>2$). Assume that in a particular realization we have $\epsilon_x(i) < \epsilon_y(i)$, as in Figs. 1(b) and 1(c). In that case we know that there cannot be any point in the two rectangles $[x_i-\epsilon_y(i)/2, x_i-\epsilon_x(i)/2]\times[y_i-\epsilon_y(i)/2, y_i+\epsilon_y(i)/2]$ and $[x_i+\epsilon_x(i)/2, x_i+\epsilon_y(i)/2]\times[y_i-\epsilon_y(i)/2, y_i+\epsilon_y(i)/2]$ (see Fig. 3). While we have taken this correctly into account when estimating $H(X,Y)$ (where it was crucial), we have neglected it in $H(X)$ and $H(Y)$. There, the corrections are $O(1/n_x)$ and $O(1/n_y)$, and should vanish for $N\to\infty$. It could be that their net effect vanishes, because they contribute with opposite signs to $H(X)$ and $H(Y)$. But we have no proof for it. Anyhow, due to the approximation of constant density within each rectangle, we cannot expect our estimates to be exact for finite N, and any justification ultimately relies on numerics.

III. IMPLEMENTATION AND RESULTS

A. Some implementation details

Mutual information is invariant under reparametrization of the marginal variables. If $X'=F(X)$ and $Y'=G(Y)$ are ho-

meomorphisms, then $I(X,Y)=I(X',Y')$ (see the Appendix). This is in contrast to $H(X)$, which changes in general under a homeomorphism. This can be used to rescale both variables first to unit variance. In addition, if the distributions are very skewed and/or rough, it might be a good idea to transform them such as to become more uniform (or at least single-humped and more or less symmetric). Although this is not required, strictly speaking it will reduce errors in general. One example is the Γ-exponential distribution in two variables, $\mu(x,y)=x^{\theta}\exp(-x-xy)/\Gamma(\theta)$ for $x,y>0$ [24], when $\theta<1$. For $\theta\to 0$, the marginal distributions develop $1/x$ and $1/y$ singularities (for $x\to 0$ and for $y\to\infty$, respectively), and the joint distribution is nonzero only in a very narrow region near the two axes. In this case our algorithm failed when applied directly, but it gave excellent results after transforming the variables to $x'=\log x$ and $y'=\log y$.

When implemented straightforwardly, the algorithm spends most of the CPU time searching for neighbors. In the most naive version, we need two nested loops through all points which gives a CPU time $O(N^2)$. While this is acceptable for very small data sets (say $N\leqslant 300$), fast neighbor search algorithms are needed when dealing with larger sets. Let us assume that X and Y are scalars. An algorithm with complexity $O(N\sqrt{k\ N})$ is then obtained by first ranking the x_i by magnitude (this can be done by any sorting algorithm such as QUICKSORT), and coranking the y_i with them [25]. Nearest neighbors of (x_i,y_i) can then be obtained by searching x neighbors on both sides of x_i and verifying that their distance in the y direction is not too large. Neighbors in the marginal subspaces are found even easier by ranking both x_i and y_i. Most results in this paper were obtained by this method, which is suitable for N up to a few thousand. The fastest (but also most complex) algorithm is obtained by using grids ("boxes") [26,27]. Indeed, we use three grids: A two-dimensional one with box size $O(\sqrt{k/N})$ and two one-dimensional ones with box sizes $O(1/N)$. First the k neighbors in 2D space are searched using the 2D grid, then the boxes at distances $\pm\epsilon$ from the central point are searched in the 1D grids to find n_x and n_y. If the distributions are smooth, this leads to complexity $O(\sqrt{kN})$. The last algorithm is comparable in speed to the algorithm of [8]. For all three versions of our algorithm it costs only little additional CPU time if one also evaluates, together with $I(X,Y)$ for some $k>1$, the estimators for smaller k.

Empirical data usually are obtained with few (e.g., 12 or 16) binary digits, which means that many points in a large set may have identical coordinates. In that case, the numbers $n_x(i)$ and $n_y(i)$ need no longer be unique (the assumption of continuously distributed points is violated). If no precautions are taken, any code based on nearest-neighbor counting is then bound to give wrong results. The simplest way out of this dilemma is to add very low-amplitude noise to the data ($\approx 10^{-10}$, say, when working with double precision) which breaks this degeneracy. We found this to give satisfactory results in all cases.

Often, MI is estimated after *rank ordering* the data, i.e., after replacing the coordinate x_i by the rank of the ith point when sorted by magnitude. This is equivalent to applying a monotonic transformation $x\to x',y\to y'$ to each coordinate, which leads to a strictly uniform empirical density, $\mu_x'(x')=\mu_y'(x')=(1/N)\Sigma_{i=1}^{N}\delta(x'-i)$. For $N\to\infty$ and $k\gg 1$ this clearly leaves the MI estimate invariant. But it is not obvious that it leaves invariant also the estimates for finite k, since the transformation is not smooth at the smallest length scale. We found numerically that rank ordering gives correct estimates also for small k, if the distance degeneracies implied by it are broken by adding low-amplitude noise as discussed above. In particular, both estimators still gave zero MI for independent pairs. Although rank ordering can reduce statistical errors, we did not apply it in the following tests, and we did not study in detail the properties of the resulting estimators.

B. Results: Two-dimensional distributions

We shall first discuss applications of our estimators to correlated Gaussians, mainly because we can in this way most easily compare with analytic results and with previous numerical analyses. In all cases we shall deal with Gaussians of unit variance and zero mean. For m such Gaussians with covariance matrix $\sigma_{ik}i,k=1\cdots m$, one has

$$I(X_1,\ldots,X_m)=-\frac{1}{2}\log[\det(\sigma)]. \tag{31}$$

For $m=2$ and using the notation $r=\sigma_{XY}$, this gives Eq. (11).

First results for $I^{(2)}(X,Y)$ with $k=1$ were already shown in Fig. 2. Results obtained with $I^{(1)}(X,Y)$ are very similar and would indeed be hard to distinguish in this figure. In Fig. 4 we compare values of $I^{(1)}(X,Y)$ (left panel) with those for $I^{(2)}(X,Y)$ (right panel) for different values of N and for $r=0.9$. The horizontal axes show k/N (left) and $(k-1/2)/N$ (right). Except for very small values of k and N, we observe scaling of the form

$$I^{(1)}(X,Y)\approx\Phi\left(\frac{k}{N}\right),\quad I^{(2)}(X,Y)\approx\Phi\left(\frac{k-1/2}{N}\right). \tag{32}$$

This is a general result and is found also for other distributions. The scaling with k/N of $I^{(1)}(X,Y)$ results simply from the fact that the number of neighbors within a fixed distance would scale $\propto N$, if there were no statistical fluctuations. For large k these fluctuations should become irrelevant, and thus the MI estimate should depend only on the ratio k/N. For $I^{(2)}(X,Y)$ this argument has to be slightly modified, since the smaller one of ϵ_x and ϵ_y is determined [for large k, where the situation illustrated in Fig. 1(c) dominates over that in Fig. 1(b)] by $k-1$ instead of k neighbors.

The fact that $I^{(2)}(X,Y)$ for a given value of k is between $I^{(1)}(X,Y)$ for $k-1$ and $I^{(1)}(X,Y)$ for k is also seen from the variances of the estimates. In Fig. 5 we show the standard deviations, again for covariance $r=0.9$. These statistical errors depend only weakly on r. For $r=0$ they are roughly 10% smaller. As seen from Fig. 5, the errors of $I^{(2)}(X,Y;k)$ are roughly halfway between those of $I^{(1)}(X,Y;k-1)$ and $I^{(1)}(X,Y;k)$. They scale roughly as $\sim\sqrt{N}$, except for very large k/N. Their dependence on k does not follow a simple scaling law. The fact that statistical errors increase when k decreases is intuitively obvious, since then the width of the distribution of ϵ increases too. Qualitatively the same depen-

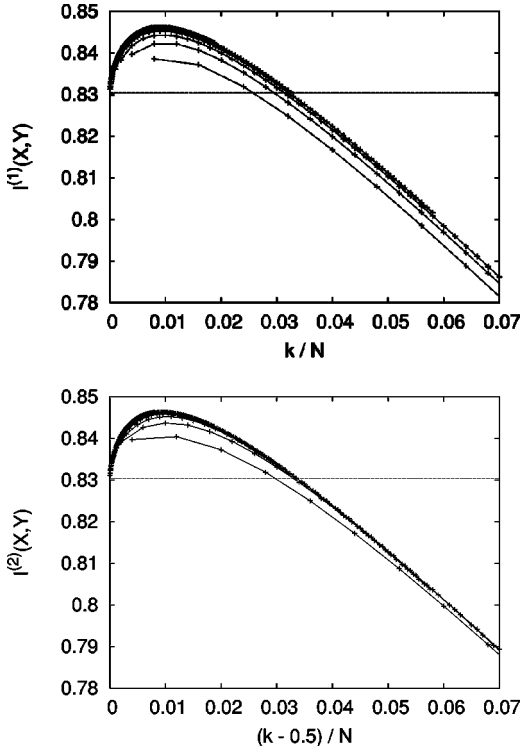

FIG. 4. Mutual information estimates $I^{(1)}(X,Y)$ (left panel) and $I^{(2)}(X,Y)$ (right panel) for Gaussian deviates with unit variance and covariance $r=0.9$, plotted against k/N (left panel) and $(k-1/2)/N$ (right panel), respectively. Each curve corresponds to a fixed value of N, with $N=125,250,500,1000,2000,4000,10\,000$, and $20\,000$, from bottom to top. Error bars are smaller than the size of the symbols. The dashed line indicates the exact value $I(X,Y)=0.830\,366$.

dence of the errors was observed also for different distributions. For practical applications, it means that one should use $k>1$ in order to reduce statistical errors, but too large values of k should be avoided since then the increase of systematic

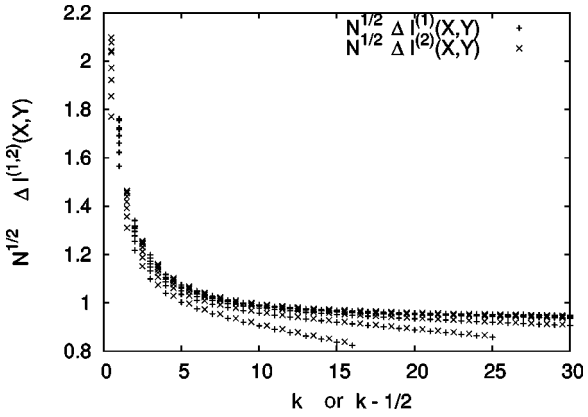

FIG. 5. Standard deviations of the estimates $I^{(1)}(X,Y)(+)$ and $I^{(2)}(X,Y)(\times)$ for Gaussian deviates with unit variance and covariance $r=0.9$, multiplied by $\sqrt{N}$ and plotted against $k[I^{(1)}(X,Y)]$ or $k-1/2[I^{(2)}(X,Y)]$. Each curve corresponds to a fixed value of N, with $N=125,250,500,1000,2000,4000,10\,000$, and $20\,000$, from bottom to top.

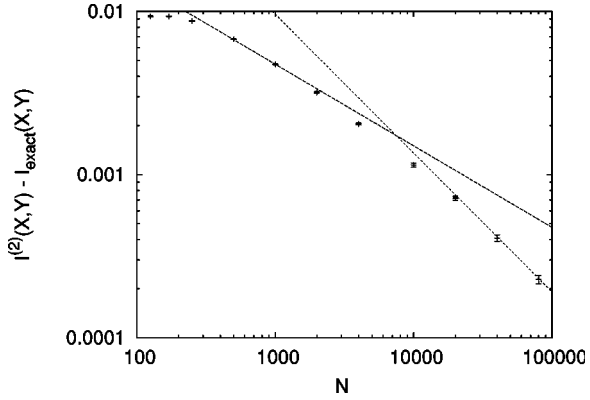

FIG. 6. Systematic error $I^{(2)}(X,Y)-I_{\text{exact}}(X,Y)$ for $k=3$ plotted against N on a log-log scale, for $r=0.9$. The dashed lines are $\propto N^{-0.5}$ and $\propto N^{-0.85}$.

errors outweighs the decrease of statistical ones. We propose to use typically $k=2-4$, except when testing for independence. In the latter case we do not have to worry about systematic errors, and statistical errors are minimized by taking k to be very large (up to $k\approx N/2$, say).

The above shows that $I^{(1)}(X,Y)$ and $I^{(2)}(X,Y)$ behave very similarly. Also CPU times needed to estimate them are nearly the same. In the following, we shall only show data for one of them, understanding that everything holds also for the other, unless the opposite is said explicitly.

For $N\to\infty$, the systematic errors tend to zero, as they should. From Figs. 2 and 4 one might conjecture that $I^{(1,2)}\times(X,Y)-I_{\text{exact}}(X,Y)\sim N^{-1/2}$, but this is not true. Plotting this difference on a double logarithmic scale (Fig. 6), we see a scaling $\sim N^{-1/2}$ for $N\approx 10^3$, but faster convergence for larger N. It can be fitted by a scaling $\sim 1/N^{0.85}$ for the largest values of N reached by our simulations, but the true asymptotic behavior is presumably just $\sim 1/N$.

As said in the Introduction, the most surprising feature of our estimators is that they seem to be exact for independent random variables X and Y. In Fig. 7 we show how the *relative* systematic errors behave for Gaussians when $r\to 0$. More precisely, we show $I^{(1,2)}(X,Y)/I^{(1,2)}_{\text{exact}}(X,Y)$ for $k=1$, plotted against N for four different values of r. Obviously these data converge, when $r\to 0$, to a finite function of N. We

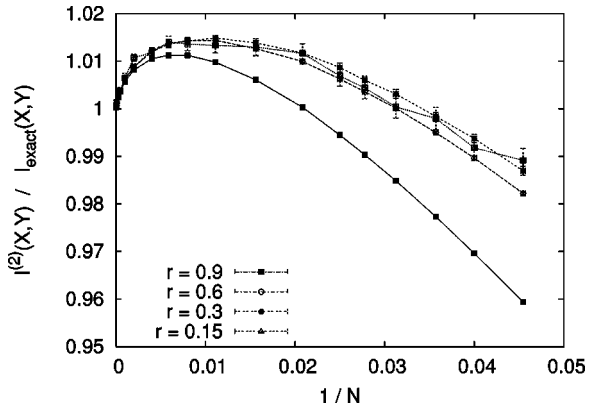

FIG. 7. Ratios $I^{(2)}(X,Y)/I_{\text{exact}}(X,Y)$ for $k=1$ plotted against $1/N$, for four different values of r.

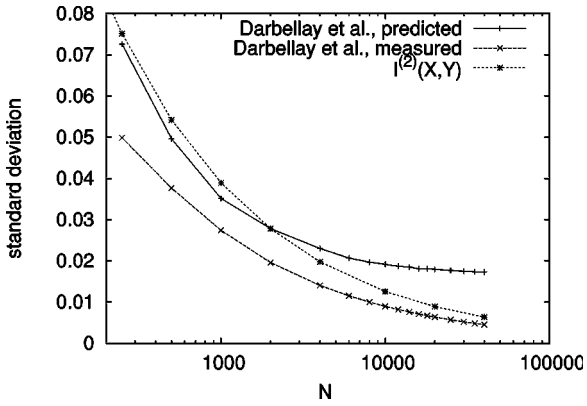

FIG. 8. Statistical errors (one standard deviation) for Gaussian deviates with $r=0.9$, plotted against N. Results from $I^{(2)}(X,Y)$ for $k=1$ (full line) are compared to theoretically predicted (dashed line) and actually measured (dotted line) errors from [8].

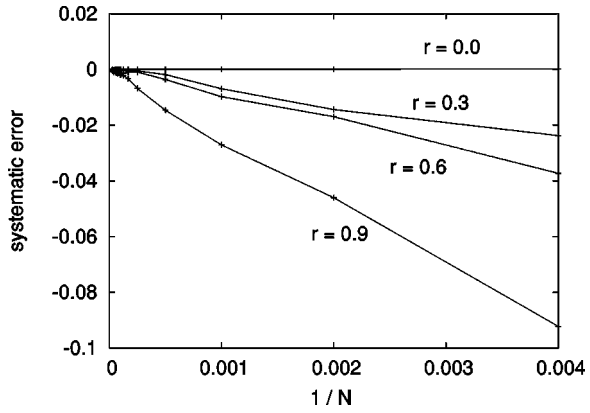

FIG. 9. Systematic errors for Gaussian deviates with $r=0.0, 0.3, 0.6$, and 0.9, plotted against $1/N$, obtained with the algorithm of [8]. These should be compared to the systematic errors obtained with the present algorithm shown in Fig. 2.

have observed the same also for other distributions, which leads to a conjecture stronger than the conjecture made in the Introduction: Assume that we have a one-parameter family of 2D distributions with densities $\mu(x,y;r)$, with r being a real-valued parameter. Assume also that μ factorizes for $r=r_0$, and that it depends smoothly on r in the vicinity of r_0, with $\partial\mu(x,y;r)/\partial r$ finite. Then we propose that for many distributions (although not for all)

$$I^{(1,2)}(X,Y)/I_{\text{exact}}(X,Y) \to F(k,N) \tag{33}$$

for $r \to r_0$, with some function $F(k,N)$ which is close to 1 for all k and all $N \gg 1$, and which converges to 1 for $N \to \infty$. We have not found a general criterion for which families of distributions we should expect Eq. (33).

The most precise and efficient previous algorithm for estimating MI is that of Darbellay and Vajda [8], and we will compare here only with their algorithm (some less systematic comparisons with a KDE method will be discussed in Sec. IV A). As far as speed is concerned, it seems to be faster than the present one, which might, however, be due to a more efficient implementation. In any case, also with the present algorithm we were able to obtain extremely high statistics on work stations within reasonable CPU times. To compare our statistical and systematic errors with those of [8], we have used the code basic.exe from Ref. [42]. We used the parameter settings recommended in its description.

This code provides an estimate of the statistical error, even if only one data set is provided. When running it with many (typically $\approx 10^4$) data sets, we found that these error bars are always underestimated, sometimes by rather large margins. This seems to be due to occasional outliers which point presumably to some numerical instability. Unfortunately, having no source code we could not pin down the troubles. In Fig. 8 we compare the predictions of the statistical errors provided by the code of [8], the actual errors obtained from the variance of the estimators provided by this code, and the error obtained from $I^{(2)}(X,Y)$ with $k=3$. We see that the latter is larger than the theoretical error from [8], but smaller than the actual error. For Gaussians with smaller correlation coefficients, the statistical errors of [8] decrease strongly with r, because the partitionings are followed to less and less depth. But, as we shall see, this comes with a risk for systematic errors.

Systematic errors of [8] for Gaussians with various values of r are shown in Fig. 9. Comparing with Fig. 2 we see that they are, for $r \neq 0$, about an order of magnitude larger than ours, except for very large N, where they seem to decrease as $1/N$. Systematic errors of [8] are also very small when $r=0$, but this seems to result from fine tuning the parameter δ_s which governs the pruning of the partitioning tree in [8]. Bad choices of δ_s lead to wrong MI estimates, and optimal choices should depend on the problem to be analyzed. No such fine tuning is needed with our method.

As examples of non-Gaussian distributions we studied (i) the Γ-exponential distribution [29], (ii) the ordered Weinman exponential distribution [29], and (iii) the "circle distribution" of Ref. [28]. For all these, both exact formulas for the MI and detailed simulations using Darbellay-Vajda algorithm exist. In addition, we tested that $I^{(1)}$ and $I^{(2)}$ vanish, within statistical errors, for independent uniform distributions, for exponential distributions, and when X was Gaussian and Y was either uniform or exponentially distributed. Notice that "uniform" means uniform within a finite interval and zero outside, so that the Kozachenko-Leonenko estimate is not exact for this case either.

In all cases with independent X and Y we found that $I^{(1,2)}(X,Y)=0$ within the statistical errors (which typically were $\approx 10^{-3}-10^{-4}$). We do not show these data.

The Γ-exponential distribution depends on a parameter θ (after a suitable rescaling of x and y) and is defined [29] as

$$\mu(x,y;\theta) = \frac{1}{\Gamma(\theta)} x^{\theta} e^{-x-xy} \tag{34}$$

for $x>0$ and $y>0$, and $\mu(x,y;\theta)=0$ otherwise. The MI is [29] $I(X,Y)_{\text{exact}} = \psi(\theta+1) - \log\theta$. For $\theta > 1$ the distribution becomes strongly peaked at $x=0$ and $y=0$. Therefore, as we already said, our algorithms perform poorly for $\theta \gg 1$, if we use x_i and y_i themselves. But using $x_i' = \log x_i$ and $y_i' = \log y_i$ we obtain excellent results, as seen from Fig. 10.

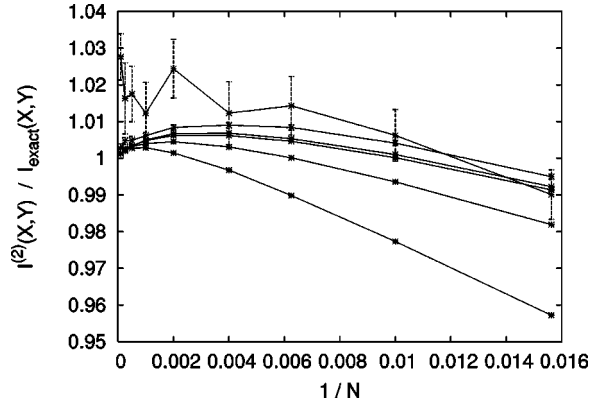

FIG. 10. Ratios $I(X,Y)_{\text{estim}}/I_{\text{exact}}(X,Y)$ for the Γ-exponential distribution, plotted against $1/N$. These data were obtained with $I^{(2)}$ using $k=1$, after transforming x_i and y_i to their logarithms. The five curves correspond to $\theta=0.1, 0.3, 1.0, 2.0, 10.0$, and 100.0 (from bottom to top).

There we plot again $I^{(2)}(X',Y')/I(X,Y)_{\text{exact}}$ for $k=1$ against $1/N$ for five values of θ. These data obviously support our conjecture that $I^{(2)}(X',Y')/I(X,Y)_{\text{exact}}$ tends towards a finite function as independence is approached. To compare with [29], we show in Fig. 11 our data together with those of [29] for the same four values of θ also studied there, namely $\theta=0.1, 0.3, 2.0$, and 100.0. We see that MI was grossly underestimated in [29], in particular for large θ where $I(X,Y)$ is very small [for $\theta \gg 1$, one has $I(X,Y) \approx 1/2\theta$].

The ordered Weinman exponential distribution depends on two continuous parameters. Following [29] we consider here only the case where one of these parameters (called θ_0 in [29]) is set equal to 1, in which case the density is

$$\mu(x,y;\theta) = \frac{2}{\theta} e^{-2x-(y-x)/\theta} \tag{35}$$

for $x>0$ and $y>0$, and $\mu(x,y;\theta)=0$ otherwise. The MI is [29]

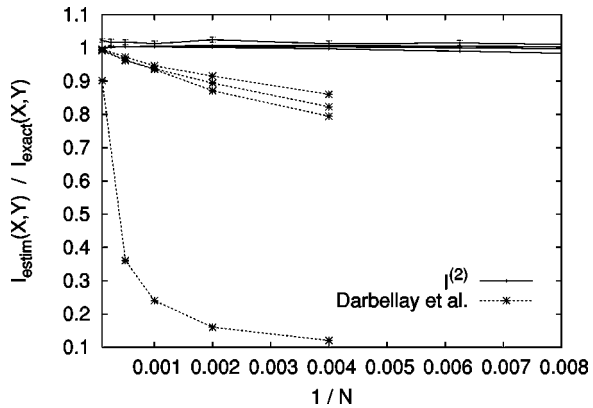

FIG. 11. Ratios $I(X,Y)_{\text{estim}}/I_{\text{exact}}(X,Y)$ for the Γ-exponential distribution, plotted against $1/N$. Full lines are from estimator $I^{(2)}$, dashed lines are from [29]. Our data were obtained with $k=1$ after a transformation to logarithms. The four curves correspond to $\theta=0.1, 0.3, 2.0$, and 100.0 (from bottom to top for our data, from top to bottom for the data of [29]).

$$I(X,Y)_{\text{exact}} = \begin{cases} \log\dfrac{2\theta}{1-2\theta} + \psi\left(\dfrac{1}{1-2\theta}\right) - \psi(1), & \theta < \dfrac{1}{2} \\ -\psi(1), & \theta = \dfrac{1}{2} \\ \log\dfrac{2\theta-1}{\theta} + \psi\left(\dfrac{2\theta}{2\theta-1}\right) - \psi(1), & \theta > \dfrac{1}{2}. \end{cases} \tag{36}$$

Mutual information estimates using $I^{(2)}(X,Y)$ with $k=1$ are shown in Fig. 12. Again we transformed $(x_i,y_i) \to (\log x_i, \log y_i)$ since this improved the accuracy, albeit not as much as for the Γ-exponential distribution. More precisely, we plot $I^{(2)}(X,Y)/I(X,Y)_{\text{exact}}$ against $1/N$ for the same four values of θ studied also in [29], and we plot also the estimates obtained in [29]. We see that MI was severely underestimated in [29], in particular for large θ where the MI is small **(**for $\theta \to \infty$, one has $I(X,Y) \approx [\psi'(1)-1]/2\theta = 0.32247/\theta$**)**. Our estimates are also too low, but much less so. It is clearly seen that $I^{(2)}(X',Y')/I(X,Y)_{\text{exact}}$ decreases for $\theta \to \infty$ in contradiction to the above conjecture. This represents the only case where the conjecture does not hold numerically. As we already said, we do not know which feature of the ordered Weinman exponential distribution is responsible for this difference.

C. Higher dimensions

In higher dimensions we shall only discuss applications of our estimators to m correlated Gaussians, because as in the case of two dimensions this is easily compared to analytic results [Eq. (31)] and to previous numerical results [30]. As already mentioned in the Introduction and as shown above for 2D distributions (Fig. 7), our estimates seem to be exact for independent random variables. We choose the same one-parameter family of 3D Gaussian distributions with all the correlation coefficients equal to r as in [30]. In Fig. 13 we show the behavior of the *relative* systematic errors of both proposed estimators. One can easily see that the data converge for $r \to 0$, i.e., when all three Gaussians become independent. This supports the conjecture made in the previous subsection. In addition, in Fig. 13 one can see the difference between the estimators $I^{(1)}$ and $I^{(2)}$. For intermediate numbers of points, $N \sim 100-200$, the "cubic" estimator has lower systematic error. Apart from that, $I^{(2)}$ evaluated for N is roughly equal to $I^{(1)}$ evaluated for $2N$, reflecting the fact that $I^{(2)}$ effectively uses smaller length scales as discussed already for $d=2$.

To compare our results in high dimension with those presented in [30], we shall calculate not the high-dimensional redundancies $I(X_1,X_2,\ldots,X_m)$ but the MI I**(**$(X_1,X_2,\ldots,X_{m-1}),X_m$**)** between two variables, namely an $(m-1)$-dimensional vector and a scalar. For estimation of this MI we can use the formulas as for the 2D case [Eqs. (8) and (9), respectively] where n_x would be defined as the number of points in the $(m-1)$-dimensional stripe of the (hyper) cubic cross section. Using directly Eq. (A3) would increase the errors in estimation [see the Appendix for the relation

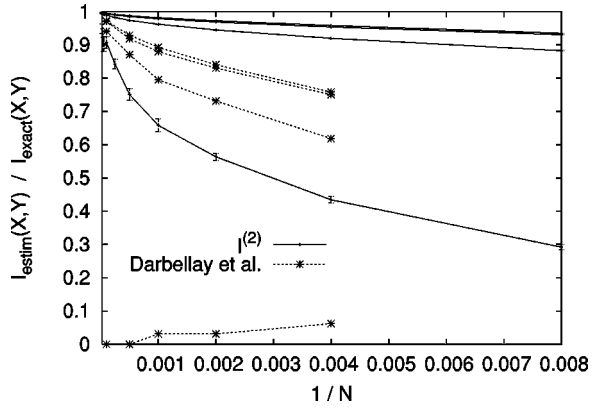

FIG. 12. Ratios $I(X,Y)_{\mathrm{estim}}/I_{\mathrm{exact}}(X,Y)$ for the ordered Weinman exponential distribution, plotted against $1/N$. Full lines are from estimator $I^{(2)}$, dashed lines are from [29]. Our data were obtained with $k=1$ after a transformation to logarithms. The four curves correspond to $\theta=0.1, 0.3, 1.0$, and 100.0 (from top to bottom).

between $I(X_1,X_2,\ldots,X_m)$ and $I((X_1,X_2,\ldots,X_{m-1}),X_m)$].

In Fig. 14 we show the average values of $I^{(1,2)}$. They are in very good agreement with the theoretical ones for all three values of the correlation coefficient r and all dimensions tested here (in contrast, in [30] the estimators of MI significantly deviate from the theoretical values for dimensions $\geqslant 6$). It is impossible to distinguish (on this scale) between estimates $I^{(1)}$ and $I^{(2)}$.

In Fig. 15, statistical errors of our estimate are presented as a function of the number of neighbors k. More precisely, we plotted the standard deviation of $I^{(1)}$ multiplied by $\sqrt{N}/m$ against k for the case where all correlation coefficients are $r=0.9$. Each curve corresponds to a different dimension m. The data scale roughly as $\sim m/\sqrt{N}$ for large dimension. Moreover, these statistical errors seem to converge to finite values for $k\to\infty$. This convergence becomes faster for increasing dimensions. The same behavior is observed for $I^{(2)}$.

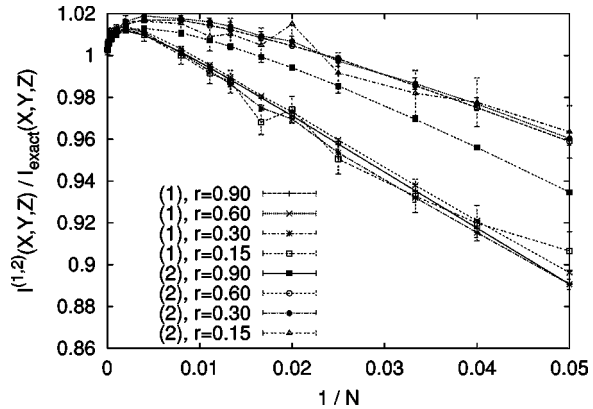

FIG. 13. Ratios $I^{(1,2)}(X,Y,Z)/I_{\mathrm{exact}}(X,Y,Z)$ for $k=1$ plotted against $1/N$, for four different values of r. All Gaussians have unit variance and all nondiagonal elements in the correlation matrix $\sigma_{i,k}, i\neq k$ (correlation coefficients) take the value r.

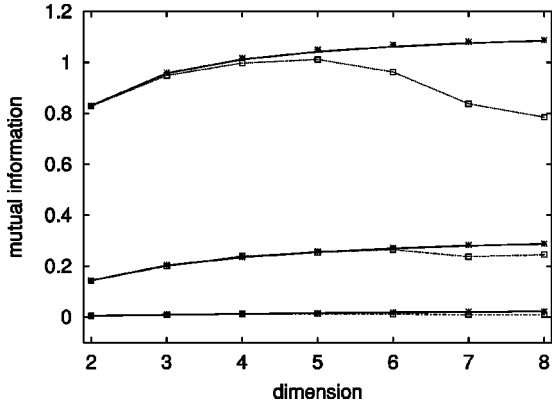

FIG. 14. Averages of $I^{(1,2)}(X_1,(X_2\cdots X_m))$ for $k=1$ plotted against m for three different values of $r=0.1, 0.5, 0.9$. The sample size is 50 000; averaging is done over 100 realizations (same parameters as in [30], Fig. 1). Full lines indicate theoretical values, pluses (+) are for $I^{(1)}$, and crosses ($\times$) are for $I^{(2)}$. Squares and dotted lines are read off from Fig. 1 of Ref. [30].

IV. APPLICATIONS: GENE EXPRESSION DATA AND INDEPENDENT COMPONENT ANALYSIS

A. Gene expression

In the first application to real world data, we study the gene expression ratios from [31], and compare our MI estimators to kernel density estimators (KDE) used in [6]. The authors of [31] considered $N=300$ closely related yeast genomes obtained by one or at most a few mutations from wild type, and indexed by $i=1,\ldots,N$. The measured raw data are expression ratios r_{im} of $M\approx 6000$ genes [open reading frames (ORFs) labeled by index $m=1,\ldots,M$] for each of the genomes. These data form an $N\times M$ matrix which can be interpreted either as a set of N vectors $\mathbf{X}_i$, each of dimension M and characterizing the expression activity of one genome, or as M 300-dimensional vectors $\mathbf{Y}_m$, each characterizing one ORF.

According to these two points of view, we can consider two types of mutual information. Mutual information between two genomes i and i', quantifying the similarities of their expression profiles, can be obtained by forming the M two-dimensional vectors $\mathbf{y}_m=(r_{im},r_{i'm})$ which can be understood as 2D projections of $\mathbf{Y}_m$, and estimating the MI of this cloud of M 2D points. Alternatively, one can estimate similarities between two ORFs m and m' by forming the N vectors $\mathbf{x}_i=(r_{im},r_{im'})$ and estimating the MI of the distribution represented by them. These MIs can then be used instead of covariance matrices to improve cluster analyses.

In the following, we shall only follow the second alternative, i.e., we only estimate MIs between ORFs, simply because we want to compare our results with those of [6] where the authors also considered only the MI between ORFs. Biologically of interest are both alternatives. We shall not discuss the subsequent cluster analysis, since this can be done with standard algorithms [31] (a clustering algorithm specific to MI used as a (dis) similarity measure will be discussed elsewhere [32]). In [6] it was found that kernel density estimators performed much better than estimators based on bin-

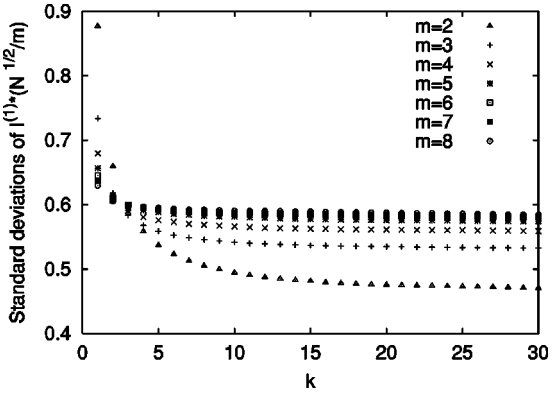

FIG. 15. Standard deviations of the estimate $I^{(1)}$ for Gaussian deviates with unit variance and covariance $r=0.9$, multiplied by $\sqrt{N}/m$ and plotted against k. Each curve corresponds to a fixed value of dimension m. Number of samples is $N=10\,000$.

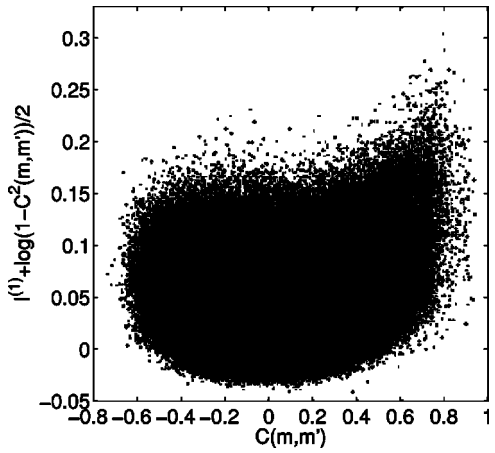

FIG. 16. Estimates $I^{(1)}(m,m')-(1/2)\ln[1-C^2(m,m')]$ for all pairs (m,m') of ORFs, plotted against $C^2(m,m')$. According to Eq. (A5), this should be positive, which gives an indication of the errors involved in the estimation.

ning, but that the estimated MIs were so strongly correlated to linear correlation coefficients that they hardly carried more useful information.

Let us first reinvestigate the MI estimates of the four ORF pairs "A" to "D" shown in Figs. 3, 5, and 7 of [6]. The claim that KDE was superior to binning was based on a surrogate analysis. For surrogates consisting of completely independent pairs, KDE was able to show that all four pairs were significantly dependent, while binning-based estimators could disprove the null hypothesis of independence only for two pairs. In addition, KDE had both smaller statistical and systematic errors. Both KDE and binning estimators were applied to rank-ordered data [6].

In KDE, the densities are approximated by sums of N Gaussians with fixed prescribed width h centered at the data points. In the limit $h\to 0$ the estimated MI diverges, while it goes to zero for $h\to\infty$. Our main criticism of [6] is that the authors used a very large value of h (roughly $\frac{1}{2}$ to $\frac{1}{3}$ of the total width of the distribution). This is recommended in the literature [33], since both statistical and systematic errors would become too large for smaller values of h. But with such a large value of h one is insensitive to finer details of the distributions, and should not be surprised to find hardly anything beyond linear correlations.

With our present estimators $I^{(1)}$ and $I^{(2)}$ we found indeed considerably larger statistical errors, when using small values of k ($k<10$, say). But when using $k\approx 50$ (corresponding to $\sqrt{k/N}\approx 0.4$, similar to the ratio h/σ used in [6]), the statistical errors were comparable to those in [6]. Systematic errors could be estimated by using the exact inequality Eq. (A5) given in the Appendix [when applying this, one has of course to remember that the estimate of the correlation coefficient also contains errors which lead to systematic overestimation of the right-hand side of Eq. (A5) [8]]. For instance, for pair "B" one finds $I>1.1$ from Eq. (A5). While this is satisfied for $k<5$ within the expected uncertainty, it is violated both by the estimate of [6] ($I\approx 0.9$) and by our estimate for $k=50$ ($I\approx 0.7$). With our method and with $k\approx 50$, we could also show that none of the four pairs is independent, with roughly the same significance as in [6].

Thus the main advantage of our method is that it does not deteriorate as quickly as KDE does for high resolution. In addition, it seems to be faster, although the precise CPU time depends on the accuracy of the integration needed in KDE. In [6] also a simplified algorithm is given [Eq. (33) of [6]] where the integral is replaced by a sum. Although it is supposed to be faster than the algorithm involving numerical integration (on which were based the above estimates), it is much slower than our present estimators [it is $O(N^2)$ and involves the evaluation of $3N^2$ exponential functions]. This simplified algorithm (which is indeed just a generalized correlation sum with the Heaviside step function replaced by Gaussians) gives also rather big systematic errors, e.g., $I=0.66$ for pair "B."

Only this simplified algorithm was used in [6] to estimate the MIs between all $M(M-1)/2$ pairs of ORFs. When plotted against the (estimated) correlation coefficients $C(m,m')$, this gave a narrow half-moon-shaped distribution whose width was not significantly larger than the estimated uncertainty (see Fig. 8 of [6]). In Fig. 16 we show our own results. We used the estimator $I^{(1)}$ with $k=30$. Since the experimental data contained some outliers, we first transformed to uniform density by rank-ordering the data. Without that, both $I^{(1)}$ and also the linear correlation would have been heavily biased for some pairs. In view of the inequality Eq. (A5) we actually plot $I^{(1)}(m,m')+\frac{1}{2}\ln[1-C^2(m,m')]$.

From Fig. 16 we see several things: First of all, if the ORFs m and m' were independent, we should have $I^{(1)}\approx 0$ on average. This is not the case, even for $C(m,m')=0$. Secondly, the average of $I^{(1)}+\frac{1}{2}\ln[1-C^2(m,m')]$ for fixed $C(m,m')$ is positive for all $C(m,m')$. Thus MI is in general not uniquely given by $C(m,m')$, and MI carries more information than linear correlations do. Third, from the violation of the inequality $I^{(1)}(m,m')+\frac{1}{2}\ln[1-C^2(m,m')]\geqslant 0$ one can estimate statistical errors. They are ≈ 0.03. Finally, while $I^{(1)}(m,m')+\frac{1}{2}\ln[1-C^2(m,m')]$ is roughly constant for $C(m,m')<0.3$, it grows sharply for large positive correlations. This effect seems not to be due to systematic or statistical errors. Indeed, systematic errors (which increase with k) would bring these points down, and the effect would not be

visible for $k>50$. It would be interesting to see what these highly correlated ORF pairs are and why their MI is even higher than suggested by linear correlations, but we shall not pursue this here.

B. ICA

Independent component analysis (ICA) is a statistical method for transforming an observed multicomponent data set (e.g., a multivariate time series comprising n measurement channels) $\mathbf{x}(t)=(x_1(t),x_2(t),\ldots,x_n(t))$ into components that are statistically as independent from each other as possible [4]. In the simplest case, $\mathbf{x}(t)$ could be a linear superposition of n independent sources $\mathbf{s}(t)=(s_1(t),s_2(t),\ldots,s_n(t))$,

$$\mathbf{x}(t)=\mathbf{A}s(t), \tag{37}$$

where $\mathbf{A}$ is a nonsingular $n\times n$ "mixing" matrix. In that case, we know that a decomposition into independent components is possible, since the inverse transformation

$$\mathbf{s}(t)=\mathbf{W}\mathbf{x}(t) \quad \text{with} \quad \mathbf{W}=\mathbf{A}^{-1} \tag{38}$$

does exactly this. If Eq. (37) does not hold, then no decomposition into strictly independent components is possible by a linear transformation like Eq. (38), but one can still search for the least dependent components. In a slight misuse of notation, this is still called ICA.

But even if Eq. (37) does hold, the problem of blind source separation (BSS), i.e., finding the matrix $\mathbf{W}$ without explicitly knowing $\mathbf{A}$, is not trivial. Basically, it requires that $\mathbf{x}$ is such that all superpositions $\mathbf{s}'=\mathbf{W}'\mathbf{x}$ with $\mathbf{W}'\neq\mathbf{W}$ are not independent. Since linear combinations of Gaussian variables are also Gaussian, BSS is possible only if the sources are not Gaussian. Otherwise, any rotation (orthogonal transformation) $\mathbf{s}'=\mathbf{R}\mathbf{s}$ would again lead to independent components, and the original sources $\mathbf{s}$ could not be uniquely recovered.

This leads to basic performance tests for any ICA problem:

(i) How independent are the found "independent" components?

(ii) How unique are these components?

(iii) How robust are the estimated *dependences* against noise, against statistical fluctuations, and against outliers?

(iv) How robust are the estimated *components*?

Different ICA algorithms can then be ranked by how well they perform, i.e., whether they find indeed the most independent components, whether they declare them as unique if and only if they indeed are, and how robust are the results. While questions (ii) and (iv) have often been discussed in the ICA literature (for a particularly interesting recent study, see [34]), the first (and most basic, in our opinion) test has not attracted much interest. This might seem strange since MI is an obvious candidate for measuring independence, and the importance of MI for ICA was noticed from the very beginning. We believe that the reason was the lack of good MI estimators. We propose to use our MI estimators not only for testing the actual independence of the components found by standard ICA algorithms, but also to use them for testing for uniqueness and robustness. We will also show how our estimators can be used for improving the decomposition obtained from a standard ICA algorithm, i.e., for finding components which are more independent. Algorithms which use our estimators for ICA from scratch will be discussed elsewhere.

It is useful to decompose the matrix $\mathbf{W}$ into two factors, $\mathbf{W}=\mathbf{R}\mathbf{V}$, where $\mathbf{V}$ is a prewhitening that transforms the covariance matrix into $\mathbf{C}'=\mathbf{V}\mathbf{C}\mathbf{V}^T=\mathbf{1}$, and $\mathbf{R}$ is a pure rotation. Finding and applying $\mathbf{V}$ is just a principal component analysis (PCA) together with a rescaling, so the core of the ICA problem reduces to finding a suitable rotation after having the data prewhitened. In the following we always assume that the prewhitening (PCA) step has already been done.

Any rotation can be represented as a product of rotations which act only in some 2×2 subspace, $\mathbf{R}=\Pi_{i,j}\,\mathbf{R}_{ij}(\phi)$, where

$$\mathbf{R}_{ij}(\phi)(x_1,\cdots x_i\cdots x_j\cdots x_n)=(x_1\cdots x_i'\cdots x_j'\cdots x_n) \tag{39}$$

with

$$x_i'=\cos\phi x_i+\sin\phi x_j, \quad x_j'=-\sin\phi x_i+\cos\phi x_j. \tag{40}$$

For such a rotation one has (see the Appendix)

$$I(\mathbf{R}_{ij}(\phi)\mathbf{X})-I(\mathbf{X})=I(X_i',X_j')-I(X_i,X_j), \tag{41}$$

i.e., the change of $I(X_1\cdots X_n)$ under any rotation can be computed by adding up changes of two-variable MIs. This is an important numerical simplification. It would not hold if MI is replaced by some other similarity measure, and it indeed is not strictly true for our estimates $I^{(1)}$ and $I^{(2)}$. But we found the violations to be so small that Eq. (41) can still be used when minimizing MI.

Let us illustrate the application of our MI estimates to a fetal ECG recorded from the abdomen and thorax of a pregnant woman (eight electrodes, 500 Hz, 5 s). We chose this data set because it was analyzed by several ICA methods [34,35] and is available on the web [37]. In particular, we will use both $I^{(1)}$ and $I^{(2)}$ to check and improve the output of the JADE algorithm [36] (which is a standard ICA algorithm and was more successful with these data than TDSEP [38]; see [34]).

The output of JADE for these data, i.e., the supposedly least dependent components, is shown in Fig. 17. Obviously channels 1–3 are dominated by the heartbeat of the mother, and channel 5 by that of the child. Channels 4 and 6 still contain large heartbeat components (of mother and child, respectively), but look much more noisy. Channels 7 and 8 seem to be dominated by noise, but with rather different spectral composition. The pairwise MIs of these channels are shown in Fig. 18 (left panel) [39]. One sees that most MIs are indeed small, but the first three components are still highly interdependent. This could be a failure of JADE, or it could mean that the basic model does not apply to these components. To decide between these possibilities, we minimized $I(X_1\cdots X_8)$ by means of Eqs. (39)–(41). For each pair (i,j) with $i,j=1\cdots 8$ we found the angle which minimized

FIG. 17. Estimated independent components using JADE.

$I(X_i',X_j')-I(X_i,X_j)$, and repeated this altogether ≈ 10 times. We did this both for $I^{(1)}$ and $I^{(2)}$, with $k=1$. We checked that $I(X_1 \ldots X_8)$, calculated directly, indeed decreased (from $I^{(1)}_{\mathrm{JADE}}=1.782$ to $I^{(1)}_{\mathrm{min}}=1.160$ and from $I^{(2)}_{\mathrm{JADE}}=2.264$ to $I^{(2)}_{\mathrm{min}}=1.620$).

The resulting components are shown in Fig. 19. The first two components look now much cleaner; all the noise from the first three channels seems now concentrated in channel 3. But otherwise things have not changed very much. The pairwise MI after minimization is shown in Fig. 18 (right panel). As suggested by Fig. 19, channel 3 is now much less dependent on channels 1 and 2. But the latter are still very strongly interdependent, and a linear superposition of independent sources as in Eq. (37) can be ruled out. This was indeed to be expected: In any oscillating system there must be at least two mutually dependent components involved, and generically one expects both to be coupled to the output signal.

To test for the uniqueness of the decomposition, we computed the variances

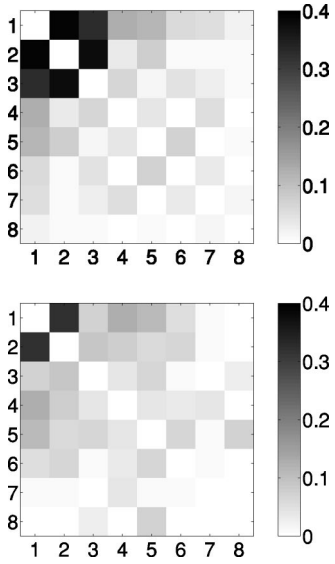

FIG. 18. Left panel: pairwise MIs between all ICA components obtained by JADE, estimated with $I^{(1)},k=1$. The diagonal is set to zero. Right panel: pairwise MIs between the optimized channels shown in Fig. 19.

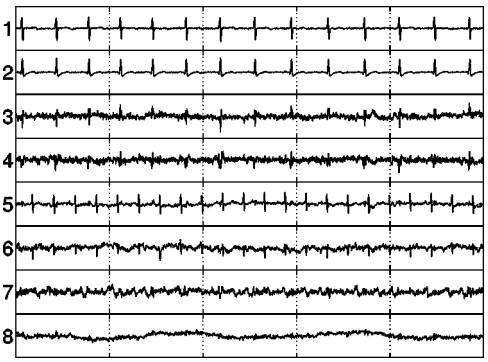

FIG. 19. Estimated independent components after minimizing I^1.

$$\sigma_{ij}=\frac{1}{2\pi}\int_0^{2\pi} d\phi[I(\mathbf{R}(\phi)(X_i,X_j))-\overline{I(X_i,X_j)}]^2, \qquad (42)$$

where

$$\overline{I(X_i,X_j)}=\frac{1}{2\pi}\int_0^{2\pi} d\phi I(\mathbf{R}(\phi)(X_i,X_j)). \qquad (43)$$

If σ_{ij} is large, the minimum of the MI with respect to rotations is deep and the separation is unique and robust. If it is small, however, BSS cannot be achieved since the decomposition into independent components is not robust. Results for the JADE output are shown in Fig. 20 (left panel), and those for the optimized decomposition are shown in the right panel of Fig. 20. The most obvious difference between them is that the first two channels have become much more clearly distinct and separable from the rest, while channel 3 is less separable from the rest (except from channel 5). This makes sense, since channels 3, 4, 7, and 8 now contain mostly Gaussian noise, which is featureless and thus rotation invariant after whitening. Most of the signals are now contained in channel 5 (fetus) and in channels 1 and 2 (mother).

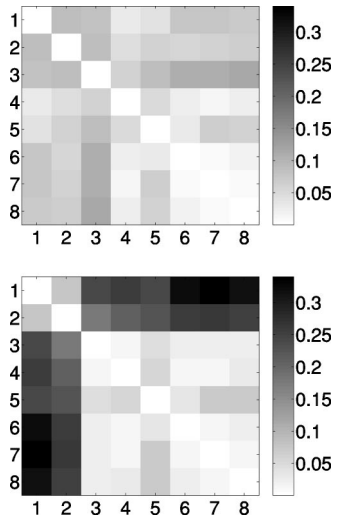

FIG. 20. Square roots of variances, $\sqrt{\sigma_{ij}}$, of $I^{(1)}[(X_i,X_j)]$ (with $k=1$) from JADE output (left panel) and after minimization of MI (right panel). Again, elements on the diagonal have been set to zero.

These results are in good agreement with those of [34], but are obtained with less numerical effort and can be interpreted more straightforwardly.

V. CONCLUSION

We have presented two closely related families of mutual entropy estimators. Each family is parametrized by an integer $k \geqslant 1$ and uses kth neighbor distance statistics in the joint space. In general they perform very similarly, as far as CPU times, statistical errors, and systematic errors are concerned. Choosing small k reduces in general systematic errors, while large k leads to smaller statistical errors. The choice of the particular estimator depends thus on the size of the data sample and on whether bias or variance is to be minimized.

Their biggest advantage seems to be in vastly reduced systematic errors (in particular for small k) when compared to previous estimators. This allows us to use them on very small data sets (even fewer than 30 points gave good results). It also allows us to use them in independent component analyses to estimate absolute values of mutual dependences. Traditionally, contrast functions have been used in ICA which allow us to minimize MI but not to estimate its absolute value. We expect that our estimators will also become useful in other fields of time series and pattern analysis. One large class of problems is interdependences in physiological time series, such as breathing and heartbeat, or in the output of different EEG channels. The latter is particularly relevant for diseases characterized by abnormal synchronization, such as epilepsy or Parkinson's disease. In the past, various measures of interdependence have been used, including MI. But the latter was not employed extensively (see, however, [40]), mainly because of the supposed difficulty in estimating it reliably. We hope that the present estimators might change this situation.

ACKNOWLEDGMENTS

One of us (P.G.) wants to thank Georges Darbellay for extensive and very fruitful e-mail discussions. We also want to thank Ralph Andrzejak, Thomas Kreuz, and Walter Nadler for numerous fruitful discussions, and for critically reading the manuscript.

APPENDIX

We collect here some well-known facts about MI, in particular for higher dimensions, and some immediate consequences. The first important property of $I(X,Y)$ is its independence with respect to reparametrizations. If $X'=F(X)$ and $Y'=G(Y)$ are homeomorphisms (smooth and uniquely invertible maps), and $J_X=\|\partial X/\partial X'\|$ and $J_Y=\|\partial Y/\partial Y'\|$ are the Jacobi determinants, then

$$\mu'(x',y') = J_X(x')J_Y(y')\mu(x,y) \tag{A1}$$

and similarly for the marginal densities, which gives

$$\begin{aligned} I(X',Y') &= \int\int dx'dy'\mu'(x',y')\log\frac{\mu'(x',y')}{\mu'_x(x')\mu'_y(y')} \\ &= \int\int dxdy\mu(x,y)\log\frac{\mu(x,y)}{\mu_x(x)\mu_y(y)} = I(X,Y). \end{aligned} \tag{A2}$$

The next important property, checked also directly from the definitions, is

$$I(X,Y,Z) = I((X,Y),Z) + I(X,Y). \tag{A3}$$

This is analogous to the additivity axiom for Shannon entropies [1], and says that MI can be decomposed into hierarchical levels. By iterating it, one can decompose $I(X_1\cdots X_n)$ for any $n>2$ and for any partitioning of the set $(X_1\cdots X_n)$ into the MI between elements within one cluster and MI between clusters.

Let us now consider a homeomorphism $(X',Y')=F(X,Y)$. By combining Eqs. (A2) and (A3), we obtain

$$\begin{aligned} I(X',Y',Z) &= I((X',Y'),Z) + I(X',Y') = I((X,Y),Z) + I(X',Y') \\ &= I(X,Y,Z) + [I(X',Y') - I(X,Y)]. \end{aligned} \tag{A4}$$

Thus, changes of high-dimensional redundancies under reparametrization of some subspace can be obtained by calculating MIs in this subspace only. Although this is a simple consequence of well-known facts about MI, it seems to have not been noticed before. It is numerically extremely useful, and would not hold in general for other interdependence measures. Again it generalizes to any dimension and to any number of random variables.

It is well known that Gaussian distributions maximize the Shannon entropy for given first and second moments. This implies that the Shannon entropy of any distribution is bounded from above by $(1/2)\log\det\mathbf{C}$, where $\mathbf{C}$ is the covariance matrix. For MI one can prove a similar result: For any multivariate distribution with joint covariance matrix $\mathbf{C}$ and variances $\sigma_i=C_{ii}$ for the individual (scalar) random variables X_i, the redundancy is bounded from below,

$$I(X_1,\cdots,X_m) \geqslant \frac{1}{2}\log\frac{\det\mathbf{C}}{\sigma_1\cdots\sigma_m}. \tag{A5}$$

The right-hand side of this inequality is just the redundancy of the corresponding Gaussian, and to prove Eq. (A5) we must show that the distribution minimizing the MI is Gaussian.

In the following we sketch only the proof for the case of two variables X and Y, the generalization to $m>2$ being straightforward. We also assume without loss of generality that X and Y have zero mean. To prove Eq. (A5), we set up a minimization problem where the constraints [correct normalization and correct second moments; consistency relations $\mu_x(x)=\int dy\ \mu(x,y)$ and $\mu_y(y)=\int dx\ \mu(x,y)$] are taken

into account by means of Lagrangian multipliers. The "Lagrangian equation" $\delta L/\delta\mu(x,y)=0$ leads then to

$$\mu(x,y)=\frac{1}{Z}\mu_x(x)\mu_y(y)e^{-ax^2-by^2-cxy}, \tag{A6}$$

where Z, a, b, and c are constants fixed by the constraints. Since the minimal MI decreases when the variances $\sigma_x=C_{xx}$ and $\sigma_y=C_{yy}$ increase with C_{xy} fixed, the constants a and b are non-negative. Equation (A6) is obviously consistent with $\mu(x,y)$ being a Gaussian. To prove uniqueness, we integrate Eq. (A6) over y and set $x=-iz/c$ to obtain

$$Ze^{-az^2/c^2}=\int dy\, e^{izy}[\mu_y(y)e^{-by^2}]. \tag{A7}$$

This shows that $e^{-by^2}\mu_y(y)$ is the Fourier transform of a Gaussian, and thus $\mu_y(y)$ is also Gaussian. The same holds true of course for $\mu_x(x)$, showing that the minimizing $\mu(x,y)$ must be Gaussian, QED.

Finally, we should mention some possibly confusing notations. First, MI is often also called transinformation or redundancy. Secondly, what we call higher-order redundancies are called higher-order MIs in the ICA literature. We did not follow that usage in order to avoid confusion with cumulant-type higher-order MIs [41].

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).
[2] A. Renyi, *Probability Theory* (North Holland, Amsterdam, 1971).
[3] *Independent Component Analysis: Principles and Practice*, edited by S. Roberts and R. Everson (Cambridge Univ. Press, Cambridge, 2001).
[4] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis* (Wiley, New York, 2001).
[5] Y.-I. Moon, B. Rajagopalan, and U. Lall, Phys. Rev. E **52**, 2318 (1995).
[6] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, Bioinformatics **18** (Suppl. 2), S231 (2002).
[7] A. M. Fraser and H. L. Swinney, Phys. Rev. A **33**, 1134 (1986).
[8] G. A. Darbellay and I. Vajda, IEEE Trans. Inf. Theory **45**, 1315 (1999).
[9] P. Grassberger, Phys. Lett. A **128**, 369 (1988).
[10] M. S. Roulston, Physica D **125**, 285 (1999).
[11] R. L. Dobrushin, Theor. Probab. Appl. **3**, 462 (1958).
[12] O. Vasicek, J. R. Stat. Soc. Ser. B. Methodol. **38**, 54 (1976).
[13] A. Kaiser and T. Schreiber, Physica D **166**, 43 (2002).
[14] E. S. Dudewicz and E. C. van der Meulen, J. Am. Stat. Assoc. **76**, 967 (1981).
[15] B. van Es, Scand. J. Stat. **19**, 61 (1992).
[16] N. Ebrahimi, K. Pflughoeft, and E. S. Soofi, Stat. Probab. Lett. **20**, 225 (1994).
[17] J. C. Correa, Commun. Stat: Theory Meth. **24**, 2439 (1995).
[18] A. B. Tsybakov and E. C. van der Meulen, Scand. J. Stat. **23**, 75 (1996).
[19] R. Wieczorkowski and P. Grzegorzewksi, Commun. Stat.-Simul. Comput. **28**, 541 (1999).
[20] L. F. Kozachenko and N. N. Leonenko, Probl. Inf. Transm. **23**, 95 (1987).
[21] P. Grassberger, Phys. Lett. **107A**, 101 (1985).
[22] R. L. Somorjai, *Methods for Estimating the Intrinsic Dimensionality of High-Dimensional Point Sets*, in Dimensions and Entropies in Chaotic Systems, edited by G. Mayer-Kress (Springer, Berlin, 1986).
[23] J. D. Victor, Phys. Rev. E **66**, 051903 (2002).
[24] G. A. Darbellay and I. Vajda, IEEE Trans. Inf. Theory **46**, 709 (2000).
[25] W. H. Press *et al.*, *Numerical Recipes* (Cambridge Univ. Press, New York, 1993).
[26] P. Grassberger, Phys. Lett. A **148**, 63 (1990).
[27] R. Hegger, H. Kantz, and T. Schreiber, TISEAN: Nonlinear Time Series Analysis Software Package (ULR: www.mpipks-dresden.mpg.de/~tisean).
[28] G. A. Darbellay, Comput. Stat. Data Anal. **32**, 1 (1999).
[29] G. A. Darbellay and I. Vajda, Inst. of Information Theory and Automation, Technical Rep. No. 1921 (1998), to be obtained from http://siprint.utia.cas.cz/darbellay/
[30] G. A. Darbellay, *3rd IEEE European Workshop on Computer-intensive Methods in Control and Data Processing, Prague, 1998* (IEEE, Piscataway, NJ, 1999), p. 83.
[31] T. R. Hughes *et al.*, Cell **102**, 109 (2000); see also www.rii.com/publications/2000/cell_hughes.htm
[32] H. Stögbauer, A. Kraskov, S. A. Astakhov, and P. Grassberger, e-print physics/0405044.
[33] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, London, 1986).
[34] F. Meinecke, A. Ziehe, M. Kawanabe, and K.-R. Mller, IEEE Trans. Biomed. Eng. **49**, 1514 (2002).
[35] J.-F. Cardoso, *Multidimensional Independent Component Analysis, Proceedings of ICASSP '98* (IEEE, Piscataway, NJ, 1998).
[36] J.-F. Cardoso and A. Souloumiac, IEE Proc. F, Radar Signal Process. **140**, 362 (1993).
[37] Daisy: Database for the identification of systems, edited by B. L. R. De Moor, www.esat.kuleuven.ac.be/sista/daisy (1997).
[38] A. Ziehe and K.-R. Mller, TDSEP—*An Efficient Algorithm for Blind Separation Using Time Structure*, in Proceedings of the 8th International Conferences on Artificial Neural Networks, ICANN'98, edited by L. Niklasson *et al.* (Springer, Berlin, 1998), p. 675.
[39] In Figs. 18–20 we used $k=1$, since the time sequences were

sufficiently long to give very small statistical errors. To find components as independent as possible, we should have used much larger k, since this would reduce statistical errors at the cost of increased but nevertheless very small systematic errors. We checked that basically the same results were obtained with k up to 100.

[40] B. Pompe, P. Blidh, D. Hoyer, and M. Eiselt, IEEE Eng. Med. Biol. Mag. **17**, 32 (1998).
[41] H. Matsuda, Phys. Rev. E **62**, 3096 (2000).
[42] See http://siprint.utia.cas.cz/timeseries/