Estimation of mutual information using kernel density estimators

Young-Il Moon, Balaji Rajagopalan, and Upmanu Lall
Utah Water Research Laboratory, Utah State University, Logan, Utah 84322-8200

(Received 8 May 1995)

Mutual information is useful for investigating the dependence between two experimental time series. It is often used to establish an appropriate time delay in phase-portrait reconstruction from time-series data. A histogram based approach has been used so far to estimate the probabilities. It is shown here that kernel density estimation of the probability density functions needed in estimating the average mutual information across two coordinates can be more effective than the histogram method of Fraser and Swinney [Phys. Rev. A **33**, 1134 (1986)].

PACS number(s): 05.45.+b, 02.50.Rj

I. INTRODUCTION

Mutual information [1] provides a general measure of dependence between two variables. Consequently, it is an important statistic when analyzing experimental time series from nonlinear systems. The two variables of interest may be lagged copies of the scalar time series of the same observable or two coordinates in a multivariate time series. Let us denote the time series of the two variables as $s_1, s_2, \dots, s_i, \dots, s_n$, and $q_1, q_2, \dots, q_j, \dots, q_n$. where n is the record length and the sampling rate ∂t is fixed. The *mutual information* between observations s_i and q_j is defined in bits as

$$I_{s,q}(s_i,q_j)=\log_2\left[\frac{P_{s,q}(s_i,q_j)}{P_s(s_i)P_q(q_j)}\right], \tag{1}$$

where $P_{s,q}(s_i,q_j)$ is the joint probability density of s and q evaluated at (s_i,q_j) and $P_s(s_i)$ and $P_q(q_j)$ are the marginal probability densities of s and q evaluated at s_i and q_j respectively.

If s and q are independent, their joint probability density $P_{s,q}(s_i,q_j)$ will simply be the product of their marginal probability densities $P_s(s_i)$ and $P_q(q_j)$ and $I_{s,q}(s_i,q_j)$ is zero. On the other hand, if s_i is completely determined by q_j, then $I_{s,q}(s_i,q_j)$ will tend to infinity. The mutual information measure is symmetric, i.e., $I_{s,q}(s_i,q_j)=I_{q,s}(q_j,s_i)$. Information theoretic and entropy based interpretations of mutual information exist [1].

Where the overall dependence between the two series is of interest, one can define (analogously to linear correlation) the *average mutual information* $\bar{I}_{s,q}$ as

$$\bar{I}_{s,q}=\sum_{i,j}P_{s,q}(s_i,q_j)\log_2\left[\frac{P_{s,q}(s_i,q_j)}{P_s(s_i)P_q(q_j)}\right]. \tag{2}$$

This measure is useful for identifying components in multivariate sampling that seem to be related or independent. A particular recent use [2–4] is the choice of an appropriate delay parameter while reconstructing a state space from an experimental time series.

There has been growing interest in state space reconstruction from time series data in fields as diverse as hydrology [5], hydrodynamics [6], epidemiology [7], and chemistry [8]. The state space is constructed by developing a d-dimensional embedding in terms of a vector time series $\mathbf{x}(t)$ using time delays τ: $\mathbf{x}(t)=\{x_t, x_{t-\tau}, x_{t-2\tau}, \dots, x_{t-(d-1)\tau}\}$. For short and noisy data the quality of the reconstruction depends on the value chosen for τ [8]. If τ is too small, the reconstructed attractor is restricted to the diagonal of the reconstruction space because x_t and $x_{t-\tau}$ will be nearly the same. On the other hand, if τ is chosen too large, and the system is chaotic, then all information to properly reconstruct the attractor may be lost, since neighboring trajectories have diverged and averaging in time and/or space is no longer useful. Consequently, a good choice of τ is one where the coordinates are first nearly independent.

Fraser and Swinney [1] proposed the use of the first minimum of the average mutual information ($I_{x_t,x_{t-\tau}}$) as a criterion for choosing τ. This choice may be better than choosing τ using the autocorrelation function (ACF) as a criteria on [9–11] since the ACF only measures the linear dependence, while $I_{x_t,x_{t-\tau}}$ measures the nonlinear dependence of two variables. However, neither criterion always gives the best possible result [12,13] in a given situation. Nevertheless, $I_{s,q}$ is useful for investigating the dependence between coordinates and also for identifying other variables that may be useful for attractor construction.

Fraser and Swinney [1] used a multivariate histogram for the estimation of the probabilities $P(\ )$ needed for estimating the average mutual information $\bar{I}_{x_t,x_{t-\tau}}$. Here we propose the use of kernel density estimators instead of histograms. Our investigations show that this is particularly advantageous with small data sets. A brief overview of kernel density estimation as used here follows. Examples that illustrate the improvement possible using these estimators are then provided.

II. KERNEL DENSITY ESTIMATION OF $\bar{I}_{x_t,x_{t-\tau}}$

Kernel density estimation (KDE) is a nonparametric method for estimating probability densities. We learn from the statistical literature [14–16] that kernel density estimates can be superior to the histogram in terms of (i) a better mean square error rate of convergence of the estimate to the underlying density, (ii) an insensitivity to the choice of origin, and (iii) the ability to specify more sophisticated window shapes than the rectangular window for "binning" or frequency counting. The latter can be exploited in multivariate settings for significantly improved density estimates.

Given an origin $\mathbf{y}_0$ and a bin width h, the bins of the histogram are defined through the hypercubes formed by intervals $[\mathbf{y}_0+mh,\ \mathbf{y}_0+(m+1)h]$ for integers M. The histogram is defined by

$$\hat{p}(\mathbf{y})=(\text{No. of } \mathbf{y}_i \text{ in same bin as } \mathbf{y})/(nh^d) \ . \tag{3}$$

While the histogram is easy to comprehend, it has several drawbacks. It is discontinuous and changes with the choice of the origin and bin width. Silverman [14] illustrates these problems graphically. Histogram construction is such a routine process that many fail to realize that even when using identical bin widths, different origin choices may change the histograms significantly. Clearly, it may be desirable to choose h differently by coordinate in the multivariate setting. In situations where the underlying data lie essentially in a subspace of dimension smaller than d (because some of the coordinates are strongly dependent), it may be desirable to construct histograms with orientation dictated by such a subspace.

Fraser and Swinney's [1] algorithm is based on an adaptive partitioning of the data such that each bin constructed has a nearly uniform distribution of points. Uniformity of points in each bin is checked using a χ^2 test at a specified level of significance. If the test for uniformity fails, that bin is partitioned along a coordinate. Consequently, the size of the bin used can vary over the state space. This adaptive binning strategy is quite sophisticated for histogram estimation.

One can free the histogram of sensitivity to the choice of origin rather easily. Define a "bin" (or hypercube) of width h centered at the point of estimate $\mathbf{y}$. This leads to a moving window rather than a fixed window estimator. Now one can still invoke the definition in (3) for a valid density estimate. The purpose of "binning" is to develop a "local" estimate (i.e., in the neighborhood of $\mathbf{y}$) of the relative frequencies of events. There is no formal reason to stay with hypercubes as bins. One could use other shapes that still lead to a valid estimate of the probability density. When a generalized weight or kernel function is used the resulting estimator is called a kernel density estimator, given [14] as

$$\hat{p}(\mathbf{y})=\frac{1}{n}\sum_{i=1}^{B} K(u) \ , \tag{4}$$

where

$$u=\frac{(\mathbf{y}-\mathbf{y}_i)^T S^{-1}(\mathbf{y}-\mathbf{y}_i)}{h^2} \ , \tag{5}$$

$K(u)$ is a multivariate kernel function, $\mathbf{y}=[y_1,y_2,\ldots,y_d]^T$ is the d-dimensional random vector whose density is being estimated, $\mathbf{y}_i=[y_{1i},y_{2i},\ldots,y_{di}]^T$, $i=1-n$ are the n sample vectors, h is the kernel bandwidth, and S is the covariance matrix on the $\mathbf{y}_i$. The kernel function $K(u)$ is required to be a valid probability density function. In this case we use the multivariate Gaussian probability density function for $K(u)$, which is given as

$$K(u)=\frac{1}{(2\pi)^{d/2}h^d\det(S)^{1/2}}\exp(-u/2) \ . \tag{6}$$

An evaluation of $K(u)$ represents the weight given to an observation $\mathbf{y}_i$, which is based on the distance between $\mathbf{y}$ and $\mathbf{y}_i$. The distance used here is the Euclidean distance modified to recognize the covariance in the coordinates. We can see from (4) that the kernel estimator is a local weighted average of the relative frequency of observations in the neighborhood of the point of estimate. The kernel function $K(\ )$ prescribes the relative weights and H prescribes the range of data values over which the average is computed. The role of the covariance matrix S is to recognize possible linear dependence among the coordinates. Its use allows one to appropriately orient the resulting kernel function and vary the bin width in proportion to the scale of variation in the rotated coordinates. Singular value decomposition is used to evaluate the inverse in (5) and to recognize rank deficiency in S. It may sometimes be desirable (if computational resources permit) to use a local (based on, say, k nearest neighbors of $\mathbf{y}$) covariance matrix $S_k(\mathbf{y})$. In this work S was based on the full sample available.

There are many methods for choosing the bandwidth h. Some of the best ones in the statistical literature are due to Sheather and Jones [17], for $d=1$, and Wand and Jones [18], for $d=2$. The computational burden associated with these and other data driven, automatic bandwidth selectors can be formidable. Here we made an expedient choice of the bandwidth as the one that minimizes the mean integrated square error in $\hat{p}(\mathbf{y})$ if the underlying distribution is assumed to be multivariate Gaussian. While this is not a theoretically satisfying choice, its performance in our tests was comparable and the computation time was orders of magnitude lower than the more rigorous choices. The "optimal" Gaussian bandwidth corresponding to the kernel choice in (6) is given by Silverman [14] as

$$h=\left\{\frac{4}{(d+2)}\right\}^{1/(d+4)} n^{-1/(d+4)} \ . \tag{7}$$

The application of this algorithm to the estimation of $\bar{I}_{x_t,x_{t-\tau}}$ proceeds as follows.

(i) For a given τ, form the n_τ data pairs $(x_t,x_{t-\tau})_i$ and the series $x_{t,i}$ and $x_{t-\tau,i}$, $i=1,\ldots,n_\tau$.

(ii) For $i=1,\ldots,n_\tau$, estimate the probabilities $P_{x_t}(x_{t,i})$, $P_{x_{t-\tau}}(x_{t-\tau,i})$, and $P_{x_t,x_{t-\tau}}((x_t,x_{t-\tau})_i)$ at the sample point, using Eqs. (4)–(7), with appropriate values of d and sample estimates of $S(x_t)$, $S(x_{t-\tau})$, and $S(x_t,x_{t-\tau})$, respectively.

(iii) Form the estimate $\bar{I}_{x_t,x_{t-\tau}}$ as in (2).

III. DATA SETS USED FOR COMPARISON

In order to demonstrate the application of the KDE for the estimation of $\overline{I}_{x_t,x_{t-\tau}}$ and the subsequent choice of a useful delay time τ, four simulated time series are chosen. They are (1) 400 data points from the sine wave $x_t = \sin(0.02\pi t)$, $t = 1, \ldots, 400$; 500 data points from an autoregressive (AR) model

$$x_t = \rho x_{t-1} + \sqrt{1-\rho^2}N(0,1) \quad (\text{where } \rho = 0.85)\ ; \qquad (8)$$

(2) 4096 data points from the Lorenz attractor [19] for the variable x with samples every 0.05 s (12 parts per orbit) given by the system of three differential equations

$$\begin{aligned} dx/dt &= 16.0(y-x)\ , \\ dy/dt &= x(45.92-z)-y\ , \qquad (9) \\ dz/dt &= xy(-4.0z)\ ; \end{aligned}$$

and (3) 2048 data points from the Rossler attractor [20] for the variable x given by the system of three ordinary differential equations

$$\begin{aligned} dx/dt &= -y-z\ , \\ dy/dt &= x+0.15y\ , \qquad (10) \\ dz/dt &= 0.2z(x-10)\ . \end{aligned}$$

$\overline{I}_{x_t,x_{t-\tau}}$ is also estimated using the histogram method of Fraser and Swinney [1] for a comparison with the KDE approach. We tried three versions of the Fraser and Swinney algorithm. They are INTER (1986), MUTINFO (19 June 1992), and MREDUND (10 July 1992). The results of MREDUND (called FSH here) seem to be better than the rest. Therefore, we present comparative results of the KDE and FSH only.

IV. RESULTS AND DISCUSSION

The average mutual information is calculated up to lag 100 for each of the data sets using both the KDE and FSH. Of primary interest is the success of each algorithm in computing $\overline{I}_{x_t,x_{t-\tau}}$ and its impact on the subsequent choice of a first local minimum of $\overline{I}_{x_t,x_{t-\tau}}$ with respect to τ. We find that this lag τ^* can vary quite a bit across methods, with the choice from FSH often being poor. For selected cases, it was possible to analytically compute the requisite probabilities and use them to derive the expected sample estimates of $\overline{I}_{x_t,x_{t-\tau}}$. In these cases, we found that the KDEs were numerically quite close to those from the analytical expressions. We do not fully understand why the FSH numbers were significantly different. No simple scaling factors were able to explain the difference. These results are now itemized.

(i) Figure 1 shows the results for the sine data set. The ACF of this series, $\cos(0.02\pi t)$, has a minimum (-1) at lag 50 and is zero at lags 25 and 75. The $\overline{I}_{x_t,x_{t-\tau}}$ estimated from the KDE has a minimum at lags 25 and 75 and at lag 50, $\overline{I}_{x_t,x_{t-\tau}}$ is infinite (which is to be expected from the sine data with a period of 50). The lags of maximum $\overline{I}_{x_t,x_{t-\tau}}$ are the same for the KDE and FSH. However, the estimated τ^* value for FSH is 2, which is obviously not the correct one (25) that is indeed obtained using the KDE. The "rough" character of the FSH estimates is disturbing. Often investigators [2] find it necessary to smooth the resulting $\overline{I}_{x_t,x_{t-\tau}}(\tau)$ with respect to τ to determine an appropriate τ^*. While this is clearly necessary given the FSH estimate, it can lead to a choice of τ^* that is an artifact of the smoothing method. The KDE estimates do not suffer from these problems because the smoothing is done at the density estimation stage. The numerical estimates of $\overline{I}_{x_t,x_{t-\tau}}$ at lags 25 and 75 from the KDE are 43.7 and 35.7. These compare favorably with the analytical estimates at these lags, which are 48.0 and 42.0. FSH reports 3.2 and 4.2.

(ii) The results for the AR data are shown in Fig. 2. Note that for an AR model the joint and marginal densities $P_{x_t,x_{t-\tau}}(\ )$, $P_{x_t}(\ )$, and $P_{x_{t-\tau}}(\ )$ are all Gaussians and hence $\overline{I}_{x_t,x_{t-\tau}}$ can be calculated directly by fitting Gaussian distributions to the data. From Fig. 2 we observe that there is little difference in the analytical estimates and the KDEs of $\overline{I}_{x_t,x_{t-\tau}}$, while FSH once again significantly underestimates. The lag τ^*would be selected as 11 from the KDE and from the analytical expression, while it would be 3 from FSH.

(iii) For the Lorenz attractor and Rossler attractor (see Figs. 3 and 4), both the KDE and FSH give similar values of τ^*. However, the numerical values of $\overline{I}_{x_t,x_{t-\tau}}$ are once again quite different. The sample sizes used for these data sets were larger than for the first two. Increasing the sample size with the first two data sets did not improve FSH performance appreciably.

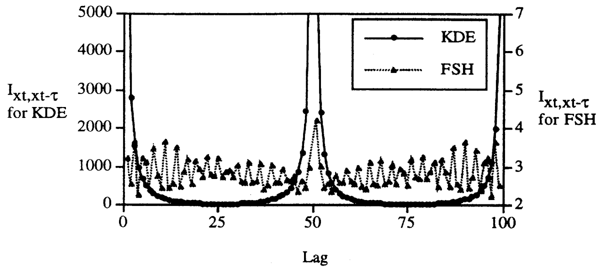

FIG. 1. $\overline{I}_{x_t,x_{t-\tau}}$ from the KDE ($\tau^*=25$) and FSH ($\tau^*=2$) for $\sin(0.02\pi t)$.

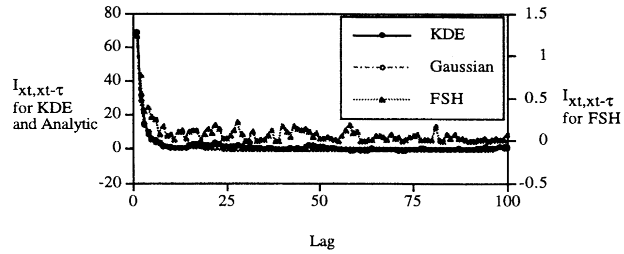

FIG. 2. $\overline{I}_{x_t,x_{t-\tau}}$ from the KDE ($\tau^*=11$), fitted Gaussian densities ($\tau^*=11$), and FSH ($\tau^*=3$) for the AR data.

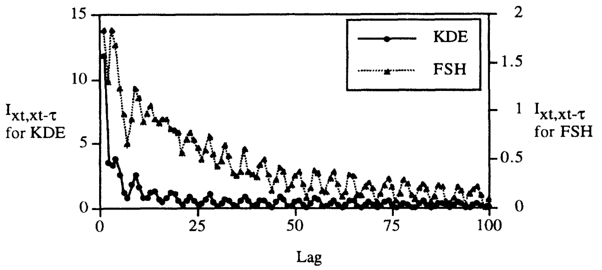

FIG. 3. $\bar{I}_{x_t,x_{t-\tau}}$ from the KDE ($\tau^*=3$) and FSH ($\tau^*=2$) for Lorenz data.

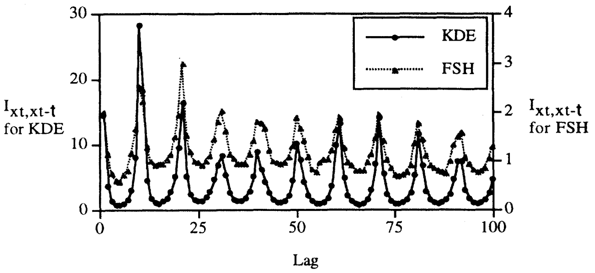

FIG. 4. $\bar{I}_{x_t,x_{t-\tau}}$ from the KDE ($\tau^*=5$) and FSH ($\tau^*=5$) for Rossler data.

Given the results here, it is clear that the KDE provides an attractive alternative to the FSH method for estimating the average sample mutual information. The implications for coordinate selection in reconstructing a phase space are also clear. We were frankly surprised to see the large difference between the FSH estimate and the KDEs (or analytically based estimates) for our test cases. FSH is a rather clever adaptive histogram estimation strategy, whereas the KDE we used is rather naive (e.g., no attempts are made to optimize the bandwidth using a data driven procedure). Extensions to the estimation of marginal and total mutual information for multiple coordinates follow directly. An adaptive kernel density estimation strategy for the multivariate probability density estimation has been developed by Lall and Bosworth [21]. They partition the data using a *k-d* tree [22] and then estimate local covariance matrices S_k for each partition k rather than using a global covariance matrix in (6). This can easily be adapted for the estimation of mutual information and would provide another alternative within the nonparametric framework.

ACKNOWLEDGMENTS

Partial support of this work by the USGS Grant No. 1434-92-G-226 and NSF Grant No. EAR-9205727 is acknowledged. We thank Thomas-Martin Kruel at the University of Wuerzburg, Germany for providing us with the code for FSH.

[1] A. M. Fraser and H. L. Swinney, Phys. Rev. A **33**, 1134 (1986).

[2] J. M. Martinerie, A. M. Albano, A. I. Mees, and P. E. Rapp, Phys. Rev. A **45**, 7058 (1992).

[3] H. D. I. Abarbanel, T. A. Carroll, L. M. Pecora, J. J. Sidorowich, and L. S. Tsimring, Phys. Rev. E **49**, 1840 (1994).

[4] J. Gao and Z. Zheng, Phys. Rev. E **49**, 3807 (1994).

[5] U. Lall, Y.-I. Moon, and K. Bosworth, Water Resour. Res. **29**, 1003 (1993).

[6] A. Brandstater, J. Swift, H. L. Swinney, A. Wolf, D. Farmer, E. Jen, and J. Crutchfield, Phys. Rev. Lett. **51**, 1442 (1983).

[7] W. M. Schaffer and M. Kot, Theor. Biol. **112**, 403 (1985).

[8] J.-C. Roux, R. H. Simoyi, and H. L. Swinney, Physica D **8**, 257 (1993).

[9] J. Holzfuss and G. Mayer-Kress, in *An Approach to Error-Estimation in the Application of Dimension Algorithms, Dimensions and Entropies in Chaotic Systems,* edited by G. Mayer-Kress (Springer-Verlag, Berlin, 1986), p. 114.

[10] A. A. Tsonis and J. B. Elsner, Nature **333**, 545 (1988).

[11] K. E. Graf and T. Elbert, in *Dimensional Analysis of the Waking EEG, Chaos in Brain Function,* edited by Erol Basar (Springer-Verlag, Berlin, 1990), p. 135.

[12] W. Li, J. Stat. Phys. **60**, 823 (1990).

[13] H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, and L. S. Tsimring, Rev. Mod. Phys. **65**, 1331 (1993).

[14] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, New York, 1986).

[15] L. Devroye and L. Györfi, *Nonparametric Density Estimation: The L1 View* (Wiley, New York, 1985).

[16] D. W. Scott, *Multivariate Density Estimation* (Wiley, New York, 1992).

[17] S. J. Sheather and M. C. Jones, J. R. Stat. Soc. B **53**, 683 (1991).

[18] M. P. Wand and M. C. Jones, Comput. Stat. **9**, 97 (1994).

[19] E. N. Lorenz, J. Atmos. Sci. **20**, 130 (1963).

[20] O. E. Rossler, Phys. Lett. **57A**, 397 (1976).

[21] U. Lall and K. Bosworth, in *Stochastic and Statistical Methods in Hydrology and Environmental Engineering, Waterloo, Time Series Analysis and Forecasting,* edited by K. Hipel (Kluwer, Dordrecht, 1993).

[22] J. H. Friedman, Smooth. Tech. Curve Estimat. **757**, 5 (1979).