Author(s)
Shiraj Khan, Sharba Bandyopadhyay, Auroop R. Ganguly, Sunil Saigal, David J. Erickson III, Vladimir Protopopescu, and George Ostrouchov

Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data

Shiraj Khan,[1,2] Sharba Bandyopadhyay,[3] Auroop R. Ganguly,[1,*] Sunil Saigal,[2]
David J. Erickson III,[4] Vladimir Protopopescu,[1] and George Ostrouchov[4]
[1]*Computational Sciences and Engineering, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA*
[2]*Civil and Environmental Engineering, University of South Florida, Tampa, Florida 33620, USA*
[3]*Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21218, USA*
[4]*Computer Science and Mathematics, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA*

(Received 6 February 2007; revised manuscript received 17 May 2007; published 14 August 2007)

Commonly used dependence measures, such as linear correlation, cross-correlogram, or Kendall's τ, cannot capture the complete dependence structure in data unless the structure is restricted to linear, periodic, or monotonic. Mutual information (MI) has been frequently utilized for capturing the complete dependence structure including nonlinear dependence. Recently, several methods have been proposed for the MI estimation, such as kernel density estimators (KDEs), k-nearest neighbors (KNNs), Edgeworth approximation of differential entropy, and adaptive partitioning of the XY plane. However, outstanding gaps in the current literature have precluded the ability to effectively automate these methods, which, in turn, have caused limited adoptions by the application communities. This study attempts to address a key gap in the literature—specifically, the evaluation of the above methods to choose the best method, particularly in terms of their robustness for short and noisy data, based on comparisons with the theoretical MI estimates, which can be computed analytically, as well with linear correlation and Kendall's τ. Here we consider smaller data sizes, such as 50, 100, and 1000, and within this study we characterize 50 and 100 data points as *very short* and 1000 as *short*. We consider a broader class of functions, specifically linear, quadratic, periodic, and chaotic, contaminated with artificial noise with varying noise-to-signal ratios. Our results indicate KDEs as the best choice for *very short* data at relatively high noise-to-signal levels whereas the performance of KNNs is the best for *very short* data at relatively low noise levels as well as for *short* data consistently across noise levels. In addition, the optimal smoothing parameter of a Gaussian kernel appears to be the best choice for KDEs while three nearest neighbors appear optimal for KNNs. Thus, in situations where the approximate data sizes are known in advance and exploratory data analysis and/or domain knowledge can be used to provide *a priori* insights into the noise-to-signal ratios, the results in the paper point to a way forward for automating the process of MI estimation.

DOI: 10.1103/PhysRevE.76.026209 PACS number(s): 05.45.−a

I. INTRODUCTION

In nonlinear systems, the understanding of underlying nonlinear processes and their interactions is very important for predictive modeling as well as for generating bounds on predictability. However, data analysis methods based on nonlinear dynamical approaches are typically not robust when applied to short and noisy data [1]. The definition of what constitutes short and noisy, in terms of data sizes and noise-to-signal ratios, may be application and context specific. A consideration of data availability scenarios in a couple of domains, specifically the earth sciences and biomedical engineering, in conjunction with the literature on mutual information (MI) estimation methods, suggest that a critical gap continues to exist in our understanding of situations where the length of data sets is short, particularly of the order of 100 or 1000 data points.

Physically based definitions for what constitutes *long* versus *short* data sizes need to follow from a comparison of sampling coverage time-span in relation to the characteristic time of the dynamical system under consideration. The characteristic time can be, for example, one full seasonal cycle for purely seasonal observations or a complete span of the attractor for a chaotic system. If the sample size is large, but the sampling coverage is restricted to a small portion of the cycle or the attractor, then observations are still not representative of the population. In this sense, the data size must still be considered *short* in a physical sense because it does not have the coverage necessary to make the relevant inferences from the data. While samples with greater coverage are more representative of the population, the trade-off, especially for a limited number of samples, is that the sampling frequency needs to be adequate to capture the features of the dynamical system and make appropriate inferences from the observations. In this sense, even if the sampling coverage is large but the frequency is inadequate, the data size must still be considered *short* from a physical perspective. Thus, the Nyquist frequency on the one hand and the characteristic period of the dynamical system under consideration on the other provide guidelines for the definitions of *long* versus *short* data sizes and indeed provide a physical basis for such definitions. However, in real-world situations, knowledge of the characteristic period of the dynamical system or the signal bandwidth may not necessarily be known *a priori* and, in some cases, may be difficult to estimate if the data are con-

[*] Corresponding author. gangulyar@ornl.gov

taminated with nonrepeatable patterns, measurement errors, or other forms of noise. Thus, for such systems, there is a need for caution before making a claim that a set of observations is *short* or, perhaps more important, *long* enough. This paper is concerned with simulated data, where we have knowledge of the system, and generates noise sequences from independent and identically distributed processes. Here we implicitly define the characteristic time (basic period) of the system as equal to unity; thus, the number of data points is a natural measure for our examples. In this study, a data size of 50–100 is referred to as *very short* whereas a data size of 1000 is considered *short*.

We use the term noise in a generic sense to include variability in measurement errors as well as any inherent, but nonrepeatable, randomness that may be present in complex systems. Indeed, noise levels encountered in real-world data may vary considerably depending on the domain, data collection methods, measurement accuracy, and inherent randomness in the observables, as well as other factors. Here we consider noise-to-signal ratios that range all the way from zero, which implies no noise, to unity, which implies that the noise is as important as the underlying signal itself. For this study, we call a noise-to-signal ratio of zero to about one-half as *low noise* and higher ratios as *high noise*.

In general, linear correlation may not be an adequate measure of dependence even for simple nonlinear functional forms. This can be simply shown in the case of two variables (X,Y) where $(Y=X^2)$ and X is uniformly distributed in the interval $(-1,1)$. The theoretical covariance and hence the linear correlation reduce to zero even though the variable Y is completely specified once X is known. The situation gets even more problematic when the nonlinear interactions get more complex. One key question is whether nonlinear dependence measures and corresponding estimation procedures can be developed to capture complete dependence, including the linear and nonlinear components thereof. However, the application of nonlinear dynamical and/or information theoretic measures of dependence can be a challenge, especially when short and noisy data are available. Thus, the second key question is whether nonlinear dependence estimation techniques can be made robust to noisy and limited data. For example, the identification of the underlying nonlinear dynamical component via the correlation dimension is known to be a difficult problem for geophysical [2] or electroencephalographic (EEG) [3,4] signals. Similarly, the detection of the underlying interactions among variables characterizing a complex system becomes a difficult task [5]. The inherent difficulty of numerical estimation as well as perceived problems with model parsimony or overfitting have resulted in a relatively limited use of nonlinear approaches, even when the underlying processes are known to be nonlinear. The problem exists in certain biomedical applications [6–8], but grows more acute in domains like geophysics [2,9] where the data collection and generation processes are often not repeatable. Our definition of what constitutes short and noisy data is motivated by problems in these domains. The references cited earlier show that *very short* and *short* data sets, as well as *low-noise* and *high-noise* conditions, do exist for real-world problems. Thus, there is a clear need to investigate methods, which are robust to short and noisy data, for the determination of nonlinear multivariate interactions. However, the methodologies need to be rigorously tested such that well-known problems in nonlinear statistics like overfitting do not yield misleading correlations.

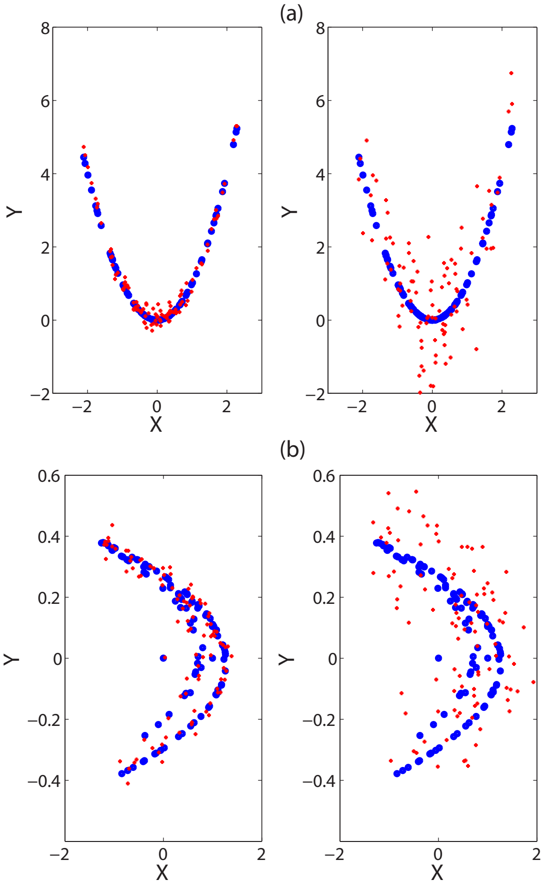

FIG. 1. (Color Online) Plot of 100 points with different noise-to-signal ratios (shown by plus) and with zero noise level (shown by dots). Noise-to-signal ratios on the left and right figures are 0.1 and 0.5, respectively. (a) $X\sim N(0,1)$, $Y:y_i=x_i^2+\varepsilon_i$, where $\varepsilon \sim N(0,\sigma_\varepsilon)$ is the Gaussian noise with zero mean and σ_ε standard deviation. (b) $X:x_i=H_{x_i}+\varepsilon x_i$, $Y:y_i=H_{y_i}+\varepsilon y_i$, where H_X and H_Y are the X and Y components of the Henon map, respectively. $\varepsilon x \sim N(0,\sigma_{H_X})$ and $\varepsilon y \sim N(0,\sigma_{H_Y})$, where σ_{H_X} and σ_{H_Y} are the standard deviations of H_X and H_Y, respectively.

The problem of detecting excessive spurious dependence or missing existing dependence structures among nonlinear signals is exacerbated for short and noisy data. The degree to which even a small amount of noise can obscure the underlying dependence structure is evident from Fig. 1 which shows two cases, such as quadratic and Henon, based on simulations with 100 points each. In both cases, the simulated data are contaminated with Gaussian noise with zero mean and standard deviation given by σ_ϵ/σ_s, which denotes the noise-to-signal ratio. The variables σ_ϵ and σ_s are the standard deviations of the noise and signal, respectively. Visual inspection reveals that the dependence structure departs significantly from the underlying true dependence structure as the noise-to-signal ratio increases. Robust measures for nonlinear dependence would need to capture the dependence structure even when the latter is obscured by noise. Previous studies designed to compare existing or newly proposed methods for nonlinear dependence with each other, as well as with a standard method, have been limited in scope. The classic algorithm was proposed by Fraser and Swinney [10], which was compared with the kernel density estimation (KDE) method given by Moon *et al.* [11]. The comparison utilized the following combination of data sizes and simulations: 400 for a sinusoidal curve, 500 for an autoregressive process, 4096 for data sets generated from the Lorenz system, and 2048 for Rössler, where the last two are chaotic. Later, KDE was refined and validated on real-world geophysical data sets [12]. Kraskov *et al.* [13] compared two k-nearest-neighbor (KNN) estimators with simulations from correlated Gaussians for data sizes of 125, 250, 500, 1000, 2000, 4000, 10 000, and 20 000, as well as with simulations from the exponential distribution. In addition, they tested their methods on gene expression data. The Edgeworth approximation of differential entropy proposed by Hulle [14] was compared against the KNN method and Parzen density estimator. For the comparisons, data sets of size 1000 and 10 000 were generated from the Gaussian and exponential distributions. Cellucci *et al.* [15] focused on statistical evaluation of mutual information estimation by comparing the MI estimates with linear correlations and the rank-based correlations from Kendall's τ. In addition, they proposed a new algorithm (henceforth referred to as Cellucci) based on adaptive partitioning and compared it with the Fraser-Swinney method given in [10]. Their comparisons utilized simulations from the Gaussian distribution and linear and quadratic functions contaminated with artificial noise, as well as the chaotic systems, such as Lorenz and Rössler. The data sizes utilized for the comparison were 4096, 8192, 10 000, 65 536, and 100 000. Cellucci *et al.* [15] mentioned that when they initiated their research, the KNN method by Kraskov *et al.* [13] had not been published yet. Indeed, they also suggested the need of an expanded future research effort to compare and contrast their adaptive partitioning method with the KNN and KDE methods.

From these discussions, it is clear that a thorough comparison of the various methods for the estimation of MI, specifically, KDE, KNN, Edgeworth, and adaptive partitioning, do not exist in the literature. Furthermore, detailed comparisons have not been attempted across a wide class of simulated functional forms. In addition, the MI estimation methods have not been compared with base-line approaches like linear correlation and Kendall's τ, other than the specific comparisons presented in the study by Cellucci *et al.* [15]. Finally, a clear gap exists in terms of detailed comparisons of the various MI estimation methods for short and noisy data.

Methods for the estimation of MI proposed in recent years include KDE [11], adaptive partitioning of the observation space [16], Parzen window density estimator [17], KNN [13], Edgeworth approximation of differential entropy (Edgeworth) [14], mutual information carried by the rank sequences [18], and adaptive partitioning of the XY plane [15]. The goal of this study is to investigate and compare recently developed MI estimation methods, specifically KDE, KNN, Edgeworth, and Cellucci, based on simulated data generated from linear, quadratic, periodic, and chaotic data contaminated artificially with various levels of Gaussian noise. We generate 50, 100, and 1000 points for our analysis. As mentioned earlier, the motivation for the data sizes comes from a specific geophysical application (the relationship of the interannual climate index known as ENSO with the variability of tropical riverflows [9]) and a specific biomedical application (dependence among EEG signals [6,8]). The simulated data allow us to compare the relative performance of the MI estimation methods across an order of magnitude in terms of data sizes and noise-to-signal ratios ranging from 0 to 1 in increments of 0.1. Uncertainties on the MI estimates are obtained through bootstrapping and provided as 90% confidence bounds. The total number of bootstraps used for 50, 100, and 1000 points are 200, 100, and 10, respectively, reflecting a pragmatic trade-off between the need for accuracy and computational tractability. However, such trade-offs may not be required in more efficient or higher-performance computational implementations. The performances of the MI estimation methods are compared against each other and against base lines comprising a linear correlation coefficient (CC) obtained from linear regression (LR) and rank-based CCs from Kendall's τ. We have also used theoretical MI values from linear, quadratic, and periodic functions, which can be computed analytically, for comparing the performance of different MI estimation methods. The purpose of the above comparisons is to identify the one MI estimation method or combination of MI estimation methods in terms of robustness to short and noisy data, at least for the illustrations considered here, whose estimation values are closest to the theoretical MI values and significantly different from linear estimates in that their confidence bounds do not intersect.

The rest of the paper is organized as follows. In Sec. II, the MI and its estimation methods are described. The MI is defined in Sec. II A while we outline the four MI estimation methods—namely, KDE, KNN, Edgeworth, and Cellucci—in Sec. II B. In Sec. III, the description of simulated data sets to be analyzed is provided. We present and discuss the results obtained using four MI estimation methods, LR, and Kendall's τ in Sec. IV. In Sec. V, the conclusion and discussion are presented.

II. MUTUAL INFORMATION AND ITS ESTIMATION METHODS

Several dependence measures, such as linear correlation, cross-correlogram, Kendall's τ, and MI, have been utilized to capture the dependence structure between a pair of variables (X,Y). However, while the first three measures can only capture linear, periodic, or monotonic dependence, MI can describe the full dependence structure including nonlinear dependence if any [19]. In addition, MI reduces to the linear dependence when the data are indeed linearly related. In an information-theoretic sense, MI quantifies the information stored in one variable about another variable. MI has several satisfying theoretical properties and analogous relations with the linear correlation. While the linear CC can be used to calculate the prediction mean squared errors (MSEs) from linear regression, MI can be used to compute a bound on the achievable prediction MSEs based on the information content in the independent variables about the dependent variables. MI has been shown to have traditional analysis-of variance- (ANOVA-) like interpretations [20]. For time serial data, MI can be computed as a function of temporal lags to obtain nonlinear versions of the auto- or cross-correlation functions (ACF or CCF). The information-theoretic properties of MI, which make it a reliable measure of the statistical dependence, have been described by Cover and Thomas [21]. The applicability of MI for feature, parameter, and model selection problems has been described by Brillinger [20]. Besides the direct use of MI in the computation of nonlinear dependence [20,22], MI has indicated a value in areas ranging from optimal time delay embeddings during phase-space reconstructions [10] to extracting causal relationships among variables [23,24,15].

A. Definitions of mutual information

For the bivariate random variables (X,Y), the MI is defined as

$$I(X;Y) = \int_Y \int_X p_{XY}(x,y) \ln \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} dxdy, \qquad (1)$$

where $p_{XY}(x,y)$ is the joint probability density function (PDF) between X and Y and $p_X(x)$ and $p_Y(y)$ are the marginal PDFs [13]. The unit of MI is defined corresponding to the base of the logarithm in Eq. (1): i.e., nats for log, bits for $\log_2$, and Hartleys for $\log_{10}$. MI is positive and symmetrical—i.e., $I(X;Y)=I(Y;X)$. It is also invariant under one-to-one transformations—i.e., $I(X;Y)=I(U;V)$, where $u=f(x)$, $v=f(y)$, and f is invertible. If X and Y are independent, the joint PDF is equal to the product of marginal PDFs leading to $I(X;Y)=0$. If there exists perfect dependence between X and Y, MI approaches infinity [21].

MI between random variables X and Y can also be defined in terms of information entropies as

$$I(X;Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y), \qquad (2)$$

where $H(X)$ and $H(Y)$ are called the marginal information entropies which measure the information content in X and Y, respectively, $H(Y|X)$ is the entropy of Y conditional on X which measures the information content remaining in Y if the information content in X is known completely, and $H(X,Y)$ is the joint information entropy which measures the information content in a pair of random variables X and Y [21]. The bivariate case is considered here for simplicity.

The linear CC between two variables X and Y denoted by $\rho(X,Y)$ is a measure of the strength of the linear dependence between the variables and varies from 0 to 1. The estimation of the most likely value and the corresponding uncertainties is relatively straightforward. However, the estimation of the mean and uncertainty bounds, for a MI-based dependence measure that is normalized to scale between 0 to 1, is an area of ongoing research.

If (X,Y) is bivariate normal, the MI and linear CC are related as $I(X;Y)=-0.5\ln[1-\rho(X,Y)^2]$ [25]. Joe [26] proposed a linear CC-like measure for MI, which scales from 0 to 1, given as

$$\hat{\lambda}(X,Y) = \sqrt{1-\exp[-2\hat{I}(X;Y)]}, \qquad (3)$$

where $\hat{\lambda}(X,Y)$ and $\hat{I}(X;Y)$ are the estimated nonlinear CC and MI, respectively. Later Granger and Lin [27] used the same measure to estimate nonlinear CC from the MI. While this study utilizes nonlinear CC based solely on MI, other bases for nonlinear CC suggested in the literature include mutual nonlinear prediction [28] and nonlinear association analysis [29]. A detailed comparison of the various definitions of nonlinear CC and their relative performances is left as areas for future research. In order to estimate the predictability of Y given X, once the MI is known, Brillinger [20] proposed an equation which provides a lower bound on the prediction MSE. This equation, which is analogous to the MSE for linear regression obtained from the linear correlation coefficient, is given as

$$\Delta(Y) \geqslant \frac{1}{2\pi e} \exp[2\{\hat{H}(Y) - \hat{I}(X;Y)\}],$$

where $\hat{H}(Y)$ is the estimated information entropy of Y and $\Delta(Y)$ gives a lower bound on MSEs from the MI and measures the predictability of Y based on the information content in X.

B. Mutual information estimators

1. Kernel density estimators

The MI in Eq. (1) for any bivariate data set (X,Y) of size n can be estimated as

$$\hat{I}(X;Y) = \frac{1}{n}\sum_{i=1}^{n} \ln \frac{\hat{p}_{XY}(x_i,y_i)}{\hat{p}_X(x_i)\hat{p}_Y(y_i)}, \qquad (4)$$

where $\hat{p}_{XY}(x_i,y_i)$ is the estimated joint PDF and $\hat{p}_X(x_i)$ and $\hat{p}_Y(y_i)$ are the estimated marginal PDFs at (x_i,y_i).

For the multivariate data set $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$, where each $\boldsymbol{x}$ is in a d-dimensional space, the multivariate kernel density estimator with kernel K is defined by

$$\hat{p}(\boldsymbol{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\boldsymbol{x}-\boldsymbol{x}_i}{h}\right), \qquad (5)$$

where h is the smoothing parameter [30]. We choose the standard multivariate normal kernel defined by

$$K(\boldsymbol{x}) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T\boldsymbol{x}\right). \qquad (6)$$

Using Eqs. (5) and (6), the probability density function is defined as

$$\hat{p}(\boldsymbol{x}) = \frac{1}{nh^d}\sum_{i=1}^{n} \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{S}|}} \exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{x}_i)^T\boldsymbol{S}^{-1}(\boldsymbol{x}-\boldsymbol{x}_i)}{2h^2}\right), \qquad (7)$$

where **S** is the covariance matrix and $|\mathbf{S}|$ is the determinant of **S**. For a normal kernel, Silverman [30] suggested an optimal smoothing parameter or Gaussian bandwidth given as

$$h_o = \left(\frac{4}{d+2}\right)^{1/(d+4)} n^{-1/(d+4)}. \tag{8}$$

Moon *et al.* [11] presented the same procedure and utilized Eq. (7) for estimating marginal probability densities—i.e., $\hat{p}_X$ and $\hat{p}_Y$—and the joint probability density—i.e., $\hat{p}_{XY}$—and substituted these densities into Eq. (4) to estimate MI.

2. k-nearest neighbors

If $X=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$, where each $\boldsymbol{x}$ is in d-dimensional space, is a continuous random variable, the Shannon entropy of X, which is restricted to random variable taking discrete values, defined as

$$H(X) = -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x},$$

can be estimated by

$$\hat{H}(X) = -\frac{1}{n}\sum_{i=1}^{n} \ln \hat{p}(\boldsymbol{x}_i), \tag{9}$$

where $\hat{p}(\boldsymbol{x}_i)$ is the estimated marginal PDF at $\boldsymbol{x}_i$. Kraskov *et al.* [13] expanded Eq. (9) as

$$\hat{H}(X) = -\frac{1}{n}\sum_{i=1}^{n} \psi(n_x(i)) - \frac{1}{k} + \psi(n) + \ln c_{d_X} + \frac{d_X}{n}\sum_{i=1}^{n} \ln \epsilon(i), \tag{10}$$

where n and k are the number of data points and nearest neighbors, respectively, d_X is the dimension of $\boldsymbol{x}$, and c_{d_X} is the volume of the d_X-dimensional unit ball. For two random variables X and Y, let $\epsilon(i)/2$ be the distance between (x_i,y_i) and its kth neighbor denoted by (kx_i,ky_i). Let $\epsilon_x(i)/2$ and $\epsilon_y(i)/2$ be defined as $\|x_i-kx_i\|$ and $\|y_i-ky_i\|$, respectively. $n_x(i)$ is the number of points x_j such that $\|x_i-x_j\| \leqslant \epsilon_x(i)/2$. $\psi(x)$ is the digamma function, $\psi(x)=\Gamma(x)^{-1}d\Gamma(x)/dx$, where $\psi(x+1)=\psi(x)+1/x$ and $\Gamma(x)$ is the ordinary gamma function. The function $\psi(y)$ satisfies the relation $\psi(1)=-C$, where $C=0.577\,215\,664\,9$ is the Euler-Mascheroni constant. Similarly, $\hat{H}(Y)$ can be derived by replacing x with y in Eq. (10). In the similar way, the estimated joint entropy between X and Y can be given as

$$\hat{H}(X,Y) = -\psi(k) - \frac{1}{k} + \psi(n) + \ln(c_{d_X}c_{d_Y}) + \frac{d_X+d_Y}{n}\sum_{i=1}^{n} \ln \epsilon(i),$$

where d_Y is the dimension of $\boldsymbol{y}$, and c_{d_Y} is the volume of the d_Y-dimensional unit ball. Substituting $\hat{H}(X)$, $\hat{H}(Y)$, and $\hat{H}(X,Y)$ in Eq. (2), the MI can be estimated as

$$\hat{I}(X;Y) = \psi(k) - \frac{1}{k} - \frac{1}{n}\sum_{i=1}^{n}[\psi(n_x(i)) + \psi(n_y(i))] + \psi(n),$$

where $n_y(i)$ is the number of points y_j such that $\|y_i-y_j\| \leqslant \epsilon_y(i)/2$ [13].

3. Edgeworth approximation of differential entropy

If $X=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)$, where each $\boldsymbol{x}$ is in a d-dimensional space, the Edgeworth expansion of the density $p(\boldsymbol{x})$ after ignoring higher-order terms is given by

$$p(\boldsymbol{x}) \approx \phi_p(\boldsymbol{x})\left(1 + \frac{1}{3!}\sum_{i,j,k} \kappa^{i,j,k} h_{i,j,k}(\boldsymbol{x})\right), \tag{11}$$

where $\phi_p(\boldsymbol{x})$ is the normal distribution with the same mean and covariance matrix as p, (i,j,k) is the input dimension where $(i,j,k)\in(1,\ldots,d)$, $\kappa^{i,j,k}$ is the standardized cumulant—i.e., $\kappa^{i,j,k}=\frac{\kappa^{ijk}}{\sigma_i\sigma_j\sigma_k}$, where κ^{ijk} is the cumulant for input dimensions (i,j,k) and σ is the standard deviation—for a large number of points, and $h_{i,j,k}$ is the ijkth Hermite polynomial [14].

Let $p(\boldsymbol{x})$ be defined in a set $\mathbb{X}$. The differential entropy of X which is analogous to the Shannon entropy and could be thought of as its extension to the domain of real numbers is defined as

$$H(X) = -\int_{\mathbb{X}} p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x}.$$

In terms of the density—i.e., $p(\boldsymbol{x})$, defined in Eq. (11)—the differential entropy of X can also be defined as

$$H(p) = H(\phi_p) - J(p) = H(\phi_p) - \int_{\mathbb{X}} p(\boldsymbol{x}) \ln \frac{p(\boldsymbol{x})}{\phi_p(\boldsymbol{x})} d\boldsymbol{x}, \tag{12}$$

where $H(\phi_p)=0.5\ln|\mathbf{S}|+\frac{d}{2}\ln 2\pi+\frac{d}{2}$ is the d-dimensional entropy of normal estimate ϕ_p, where $|\mathbf{S}|$ is the determinant of a covariance matrix $\mathbf{S}$, and $J(p)$ is called negentropy, which measures the distance to normal distribution [14]. From Eq. (11), $p(\boldsymbol{x})=\phi_p(\boldsymbol{x})[1+Z(\boldsymbol{x})]$, where $Z(\boldsymbol{x})=\frac{1}{3!}\sum_{i,j,k}\kappa^{i,j,k}h_{i,j,k}(\boldsymbol{x})$. Substituting $p(\mathbf{x})$ in Eq. (12) leads to

$$H(p) \approx H(\phi_p) - \int_{\mathbb{X}} \phi_p(\boldsymbol{x})[Z(\boldsymbol{x}) + 0.5Z(\boldsymbol{x})^2] d\boldsymbol{x}.$$

Using $\int_{\mathbb{X}}\phi_p(\boldsymbol{x})Z(\boldsymbol{x})d\boldsymbol{x}=0$ and the orthogonal properties of Hermite polynomials—i.e., $\int_{-\infty}^{\infty}\phi_p(\boldsymbol{x})h_n(\boldsymbol{x})h_m(\boldsymbol{x})d\boldsymbol{x}=n!\,\delta_{nm}$, where δ_{nm} is the Kronecker delta—Hulle [14] obtained an approximate expression for $H(p)$:

$$H(p) \approx H(\phi_p) - \frac{1}{12}\sum_{i=1}^{d}(\kappa^{i,i,i})^2 - \frac{1}{4}\sum_{i,j=1,i\neq j}^{d}(\kappa^{i,i,j})^2 - \frac{1}{72}\sum_{i,j,k=1,i<j<k}^{d}(\kappa^{i,j,k})^2. \tag{13}$$

We utilize Eq. (13) for the estimation of $\hat{H}(X)$, $\hat{H}(Y)$, and $\hat{H}(X,Y)$ and substitute in Eq. (2) to get the MI estimates.

4. Adaptive partitioning of the XY plane

Cellucci *et al.* [15] developed a procedure for estimating MI such that the null hypothesis—i.e., H_0: X and Y are statistically independent—is rejected. They used an adaptive partitioning of the XY plane to estimate the joint probability density: i.e., $\hat{p}_{XY}$. The XY plane is nonuniformly partitioned in such a way that the Cochran criterion on $E_{XY}(i,j)$—i.e., $E_{XY}(i,j)\geqslant 5$ for at least 80% of all elements—is satisfied, where $E_{XY}(i,j)$ is the expected number of points in the (i,j)th element of the XY partition given the assumption of X and Y being statistically independent is valid. The whole procedure of Cellucci *et al.* [15] is described below.

Let x and y axes be partitioned into equal number of elements denoted by N_E which leads to

$$\hat{p}_X(i)=\hat{p}_Y(j)=\frac{n/N_E}{n},\quad \text{for } i,j=1,\ldots,n,$$

where n is the total number of points and $\hat{p}_X(i)$ and $\hat{p}_Y(j)$ are the marginal densities at the ith element of the x axis and jth element of the y axis, respectively. Under the null hypothesis that X and Y are statistically independent, the expected number of points in the (i,j)th element of the XY partition is given as

$$E_{XY}(i,j)=n\hat{p}_X(i)\hat{p}_Y(j)=\frac{n}{{N_E}^2}.$$

N_E is computed from a more conservative criterion—i.e., $E_{XY}(i,j)=n/{N_E}^2\geqslant 5$ for all elements—rather than the Cochran criterion. After computing N_E, N_E partitions in the x axis and N_E partitions in the y axis are used for the estimation of joint probability density at the (i,j)th element of the XY partition: i.e., $\hat{p}_{XY}(i,j)$. The MI is estimated by substituting $\hat{p}_X$, $\hat{p}_Y$, and $\hat{p}_{XY}$ in the equation given as

$$\hat{I}(X;Y)=\sum_{i=1}^{N_E}\sum_{j=1}^{N_E}\hat{p}_{XY}(i,j)\ln\frac{\hat{p}_{XY}(i,j)}{\hat{p}_X(i)\hat{p}_Y(j)}.$$

III. DETAILS OF THE DATA

We analyze simple examples of linear, quadratic, and periodic functions, as well as a chaotic system, specifically the Henon map, contaminated with different levels of artificial Gaussian noise.

Linear. A simple linear function with Gaussian noise can be generated as

$$X\sim N(0,1),\quad Y{:}y_i=x_i+\varepsilon_i,$$

where $i=1,\ldots,n$ and X is independent and identically distributed (iid). $\varepsilon\sim N(0,\sigma_\varepsilon)$ is the Gaussian noise with zero mean and standard deviation σ_ε. In this case, σ_ε gives the noise level. ε is iid and independent of X.

Quadratic. We generate a simple quadratic, with artificial Gaussian noise, in the following manner:

$$X\sim N(0,1),\quad Y{:}y_i={x_i}^2+\varepsilon_i,$$

where $i=1,\ldots,n$; X is iid and $\varepsilon\sim N(0,\sigma_\varepsilon)$ is the Gaussian noise with zero mean and standard deviation σ_ε. ε is iid and independent of X.

Periodic. We consider a simple periodic function, specifically the sine function, contaminated with Gaussian noise in the following way:

$$X\sim \text{uniform}(-\pi,\pi),\quad Y{:}y_i=\sin(x_i)+\varepsilon_i,$$

where $i=1,\ldots,n$ and X is uniformly distributed between $-\pi$ and π. $\varepsilon\sim N(0,\sigma_\varepsilon)$ is the Gaussian noise with zero mean and standard deviation σ_ε. ε is iid and independent of X.

Chaotic. We consider the Henon map given as

$$H_X{:}H_{x_{i+1}}=1-\alpha {H_{x_i}}^2+H_{y_i},$$

$$H_Y{:}H_{y_{i+1}}=\beta H_{x_i},$$

where $i=1,\ldots,n$, $\alpha=1.4$, $\beta=0.3$, and $(H_{x_1},H_{y_1})=(0.0,0.0)$. The Henon map contaminated with Gaussian noise is generated as

$$X{:}x_i=H_{x_i}+\varepsilon_{x_i},\quad Y{:}y_i=H_{y_i}+\varepsilon_{y_i},$$

where $\varepsilon_x\sim N(0,\sigma_{H_X})$ and $\varepsilon_y\sim N(0,\sigma_{H_Y})$ are iid and independent of H_X and H_Y, respectively. σ_{H_X} and σ_{H_Y} are the standard deviations of H_X and H_Y, respectively.

Theoretical MI can be computed analytically for the linear, quadratic, and periodic cases as shown in Appendix A.

IV. RESULTS

We first estimate MI from KDE, KNN, Edgeworth, and Cellucci and then substitute in Eq. (3) to get the nonlinear CC estimates. Linear CCs are obtained from LR whereas rank-based CCs are estimated from Kendall's τ. The mean of CCs and its 90% confidence bounds are evaluated using bootstrapping. The total number of bootstrap samples used for 50, 100, and 1000 data points are 200, 100, and 10, respectively. The correlation coefficient presented here is the mean of bootstrap samples. The lower and upper bounds of 90% confidence bounds are given as 5% and 95% quantiles of bootstrap samples, respectively.

A. Performance of linear and nonlinear dependence measures

Numerical recipes for the estimation of nonlinear dependence can be evaluated based on how well the estimators capture the complete dependence structure, including nonlinear dependence if any, how well they can refrain from capturing spurious nonlinear dependence when the dependence structure is known to be linear. In order to compare the performance of different methods, we compare nonlinear CCs from KDE, KNN, Edgeworth, and Cellucci with linear CCs obtained from LR. If the confidence bounds of nonlinear CCs overlap with the bounds of linear CCs, it means here that nonlinear correlations are not different from linear correlations at 90% confidence level. Nonlinear CCs obtained from the MI estimation methods are compared with theoretical CCs derived from the theoretical MI values which can be computed analytically for three out of four test cases considered here: namely, linear, quadratic, and periodic. The performance of the MI estimation methods is also compared with a rank-based correlation measure, specifically the Kendall's τ. Plots of normal and kernel density estimates for linear, quadratic, periodic, and chaotic cases are shown in Figs. 7–10 of Appendix B.

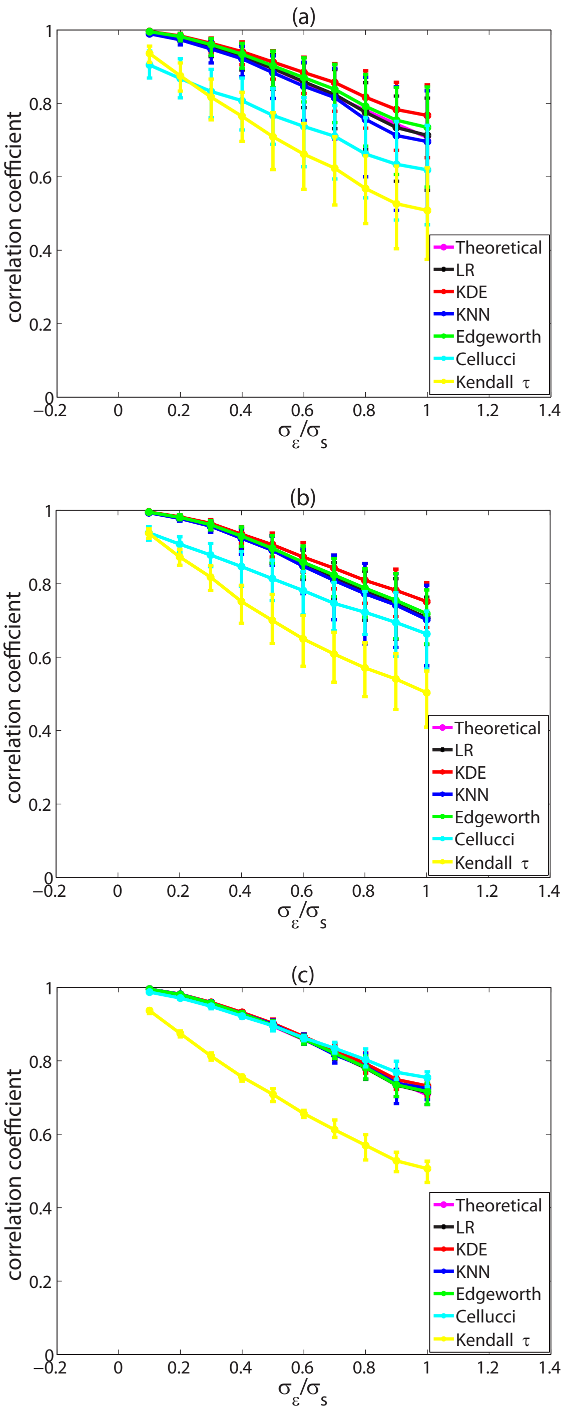

FIG. 2. (Color online) Linear: comparisons between linear CCs from LR and nonlinear CCs from KDE, KNN, Edgeworth, Cellucci, and Kendall's τ, at different noise-to-signal ratios (σ_ϵ/σ_s) for (a) 50 points, (b) 100 points, and (c) 1000 points.

1. Linear

Linear and nonlinear CCs with 90% confidence bounds are shown in Fig. 2. The theoretical CC, which is computed analytically, is expected to be identical to the linear CC. As noise levels increase, linear and nonlinear CCs decrease and their corresponding variances increase for both *very short* and *short* data. The complete description of results obtained from KDE, KNN, Edgeworth, Cellucci, and Kendall's τ for *very short* and *short* data at low and high noise is given in Table I. For *very short* data, KNN appears to be a better choice at low noise because it has no bias, overlaps with theoretical CCs, and has narrow confidence bounds [Figs. 2(a) and 2(b)]. At high noise, KDE is positively biased but it appears to be a better choice given that the others have wider confidence bounds. Thus, for *very short* data, KNN may be utilized at low noise but at high noise, KDE seems to be the best choice. For *short* data, Kendall's τ is the worst whereas Edgeworth is better than KDE because it overlaps exactly with theoretical CCs [Fig. 2(c)]. LR and KNN stand out among the rest since they have very small bias, overlap exactly with theoretical CCs, and have narrow bounds at all noise levels. Thus, for *short* data, either KNN or LR may be utilized at all noise levels.

TABLE I. Linear: description of results where each entry consists of three columns given as (1) Column 1: 0, −, or +, where "0," "−," and "+" mean nonlinear CCs are zero, negatively, and positively biased with respect to theoretical CCs, respectively. (2) Column 2: Y or N, where "Y" and "N" mean 90% confidence bounds of nonlinear CCs overlap and do not overlap with theoretical CCs, respectively. (3) Column 3: Y or N, where "Y" and "N" mean 90% confidence bounds of nonlinear CCs overlap and do not overlap with linear CCs, respectively. Bold and slanted entries indicate the best and second best methods for each case specified in the top headings of the table, respectively.

	Very short data		Short data	
	Low noise	High noise	Low noise	High noise
KDE	+ *Y Y*	**+ Y Y**	0 Y Y	+ Y Y
KNN	**0 Y Y**	−Y Y	**0 Y Y**	**0 Y Y**
Edgeworth	+ Y Y	+ *Y Y*	*0 Y Y*	*0 Y Y*
Cellucci	−N N	−Y Y	−Y Y	+ Y Y
Kendall's τ	−N N	−N Y	−N N	−N N

2. Quadratic

LR and Kendall's τ fail to capture the nonlinear dependence as shown by near zero CC in Fig. 3. The variance increases for KDE, KNN, Edgeworth, and Cellucci as the noise level increases at all noise levels. Table II gives the complete description of results obtained from LR, KDE, KNN, Edgeworth, Cellucci, and Kendall's τ for *very short* and *short* data at low and high noise. For *very short* data, as the noise level increases, the bias increases for KNN and Cellucci and decreases for KDE and Edgeworth [Figs. 3(a) and 3(b)]. At low noise, only KNN and Edgeworth overlap with theoretical CCs but KNN is more closer to theoretical than Edgeworth. At high noise, the performance of KDE is the best because it is closer to theoretical CCs, does not intersect with linear CCs, and has narrow confidence bounds as compared to that from KNN, Edgeworth, and Cellucci. Thus, for *very short* data, KNN and KDE may be utilized at low and high noise, respectively. For *short* data, KNN is the best because it overlaps exactly with theoretical CCs and has

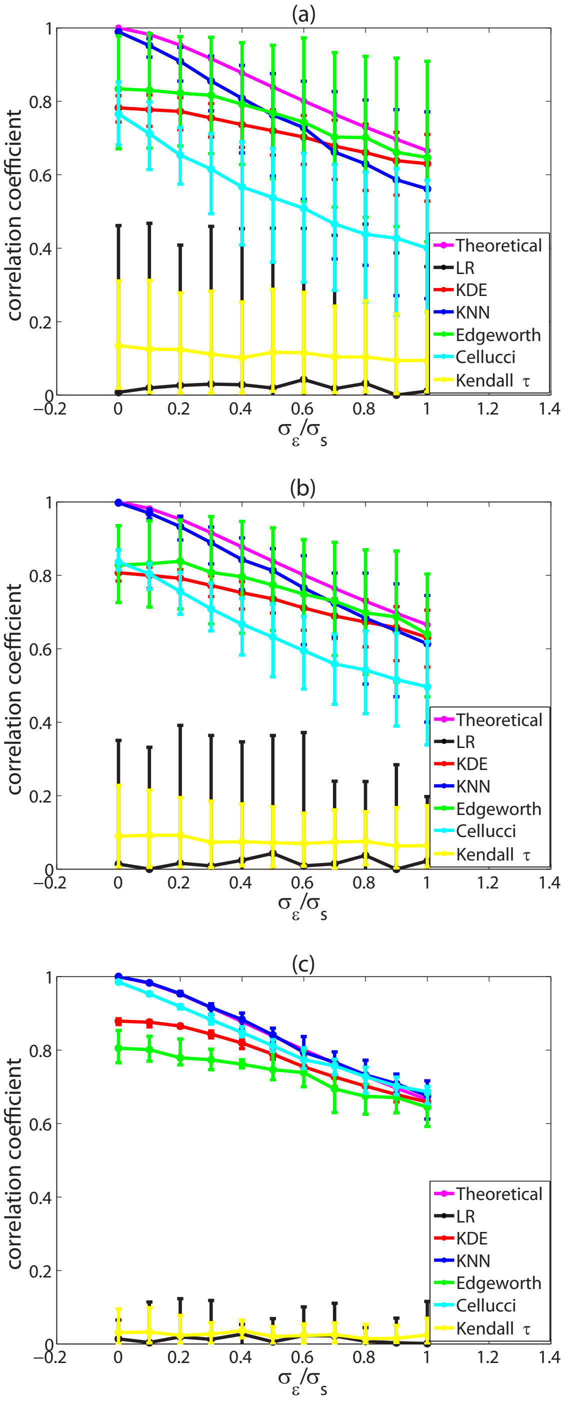

FIG. 3. (Color online) Quadratic: comparisons between linear CCs from LR and nonlinear CCs from KDE, KNN, Edgeworth, Cellucci, and Kendall's τ, at different noise-to-signal ratios (σ_ϵ/σ_s) for (a) 50 points, (b) 100 points, and (c) 1000 points.

narrow confidence bounds [Fig. 3(c)]. Cellucci is closer to theoretical CCs than KDE and Edgeworth. Thus, KNN seems to be the best choice for *short* data. KDE may be further improved at low noise by choosing a smaller value of the smoothing parameter.

TABLE II. Quadratic: description of results where each entry consists of three columns given as (1) Column 1: 0, −, or +, where "0," "−," and "+" mean nonlinear CCs are zero, negatively, and positively biased with respect to theoretical CCs, respectively. (2) Column 2: Y or N, where "Y" and "N" mean 90% confidence bounds of nonlinear CCs overlap and do not overlap with theoretical CCs, respectively. (3) Column 3: Y or N, where "Y" and "N" mean 90% confidence bounds of nonlinear CCs overlap and do not overlap with linear CCs, respectively. Bold and slanted entries indicate the best and second best methods for each case specified in the top headings of the table, respectively.

	Very short data		Short data	
	Low noise	High noise	Low noise	High noise
KDE	− N N	− **Y N**	− N N	− Y N
KNN	− **Y N**	− Y Y	**0 Y N**	**0 Y N**
Edgeworth	− *Y N*	− *Y N*	− N N	− Y N
Cellucci	− N N	− Y Y	− *N N*	− *Y N*
Kendall's τ	− N Y	− N Y	− N Y	− N Y

3. Periodic

Correlation coefficients and their 90% confidence bounds obtained from LR, KDE, KNN, Edgeworth, Cellucci, and Kendall's τ are shown in Fig. 4. KNN overlaps with theoretical CCs for both *very short* and *short* data at all noise levels except for the fact that at high noise it produces wide confidence bounds. The performance of Kendall's τ is the worst at all noise levels. Edgeworth appears to capture only the linear correlation or the linear component of the overall dependence, and produces wide confidence bounds. In this case the density of Y is bimodal, which causes Edgeworth estimates to be incorrect. The results obtained from LR, KDE, KNN, Edgeworth, Cellucci, and Kendall's τ are described in Table III for *very short* and *short* data at low and high noise. For *very short* data, the variances from all the methods are small at low noise but increase as the noise level increases [Figs. 4(a) and 4(b)]. KNN and KDE have the lowest variances at low and high noise, respectively. KNN overlaps exactly with theoretical CCs and has narrow and wide confidence bounds at low and high noise, respectively. Thus, KNN is a better choice at low noise. At high noise, KDE and KNN overlap with theoretical CCs as well as with linear CCs but KDE has the smallest confidence bounds. Thus, for *very short* data, KNN and KDE may be utilized at low and high noise, respectively. For *short* data, there is not much difference in the variances from all methods with Cellucci having the lowest variance [Fig. 4(c)]. The performances of KNN and Cellucci are better than the rest. Cellucci overlaps with theoretical CCs for only few noise levels whereas KNN overlaps exactly with theoretical CCs and has narrow bounds. Thus, KNN has an edge over all other methods considered here for *short* data across all noise levels.

4. Chaotic

For *very short* and *short* data, linear CCs between X and Y components of the Henon map are negative for all noise

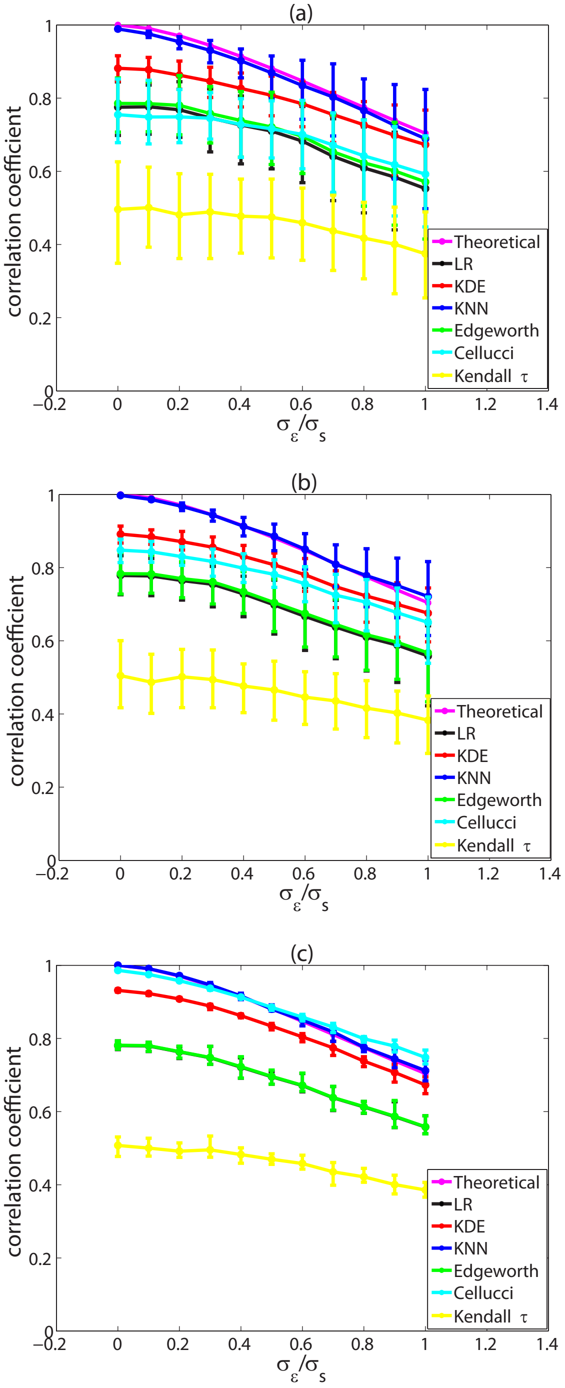

FIG. 4. (Color online) Periodic: comparisons between linear CCs from LR and nonlinear CCs from KDE, KNN, Edgeworth, Cellucci, and Kendall's τ, at different noise-to-signal ratios (σ_ϵ/σ_s) for (a) 50 points, (b) 100 points, and (c) 1000 points. In (c), LR overlaps exactly with Edgeworth.

TABLE III. Periodic: description of results where each entry consists of three columns given as (1) Column 1: 0, −, or +, where "0," "−," and "+" mean nonlinear CCs are zero, negatively, and positively biased with respect to theoretical CCs, respectively. (2) Column 2: Y or N, where "Y" and "N" mean 90% confidence bounds of nonlinear CCs overlap and do not overlap with theoretical CCs, respectively. (3) Column 3: Y or N, where "Y" and "N" mean 90% confidence bounds of nonlinear CCs overlap and do not overlap with linear CCs, respectively. Bold and slanted entries indicate the best and second best methods for each case specified in the top headings of the table, respectively.

	Very short data		Short data	
	Low noise	High noise	Low noise	High noise
KDE	− *N Y*	− **Y Y**	− N N	− *N N*
KNN	− **Y N**	− *Y Y*	**0 Y N**	**0 Y N**
Edgeworth	− N Y	− N Y	− N Y	− N Y
Cellucci	− N Y	− Y Y	− *N N*	+ N N
Kendall's τ	− N N	− N Y	− N N	−N N

levels. Since nonlinear CCs from the MI estimation methods do not have directionality, the absolute values of linear CC are considered here. Note that theoretical CCs for the Henon map could not be computed analytically and were not found in the literature. However, given the dynamical relation between X and Y, it is reasonable to expect that the theoretical nonlinear CCs will be greater than linear CCs at all noise levels. However, the performance of the numerical recipes to estimate the dependence need to be evaluated, especially in terms of their ability to capture additional dependence beyond linear correlation. Nonlinear and linear CCs decay as noise level increases.

For *very short* data, KNN estimates higher CCs than all other methods when σ_ϵ/σ_s is less than around 0.5 after which KDE yields higher values compared to all other methods [Figs. 5(a) and 5(b)]. The performance of Kendall's τ is the worst since it cannot even capture the linear portion of the dependence which is estimated by the linear correlation for the majority of noise levels. At low noise, both Edgeworth and Cellucci are ruled out because they are lower than KNN and KDE and have wide confidence bounds. Thus, KNN seems to be a better choice at low noise since KDE is negatively biased. As the noise level increases, the confidence bounds from all methods increase. At high noise, the confidence bounds from KNN, Edgeworth, and Cellucci overlap with linear CCs. KDE seems to have an edge over the other methods since it has narrow confidence bounds and does not overlap with linear CCs. Thus, KNN and KDE may be utilized for *very short* data at low and high noise, respectively. For *short* data, Cellucci differs completely from the other estimators at high noise [Fig. 5(c)]. KNN is a better choice at low noise because it appears to be the most consistent. At high noise, KNN and Edgeworth are ruled out because they overlap with linear CCs due to their wide confidence bounds. KDE overlaps with KNN but it stands out due to its ability to capture more correlation than purely linear correlation. Thus, for *short* series, KNN and KDE may be utilized at low and high noise, respectively.

B. Performance of KDE and KNN with different parameter values

In the case of KDE, the amount of smoothing defined by smoothing parameter, h in Eq. (5), is very important for the

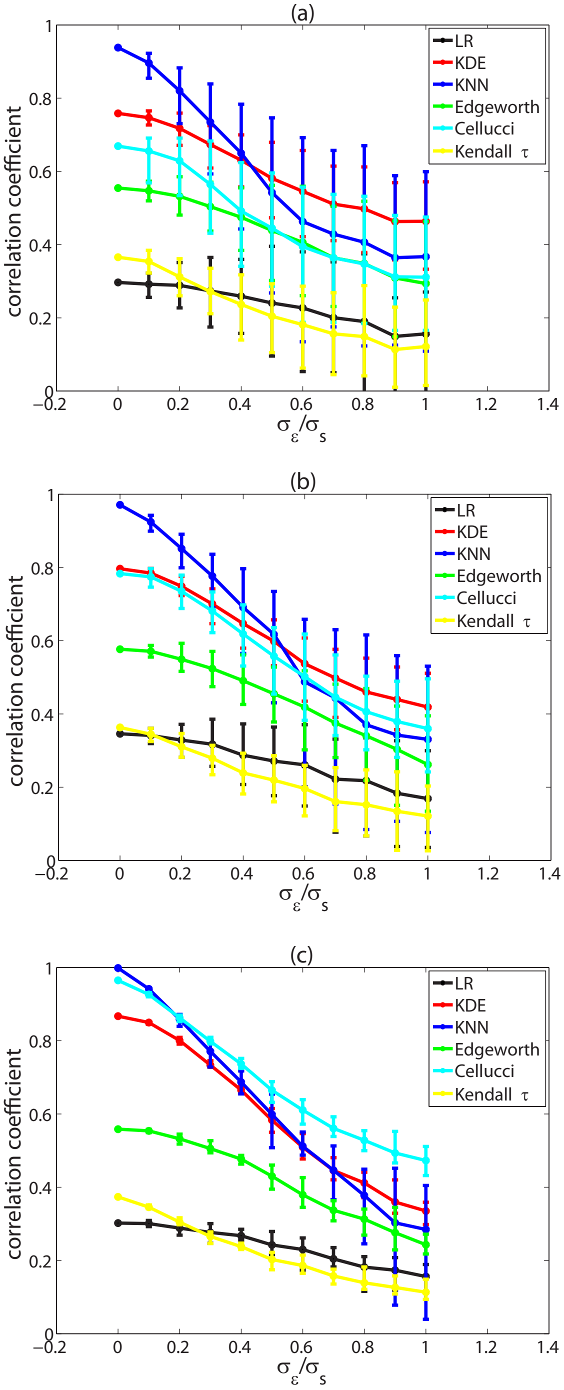

FIG. 5. (Color online) Chaotic: comparisons between linear CCs from LR and nonlinear CCs from KDE, KNN, Edgeworth, Cellucci, and Kendall's τ, at different noise-to-signal ratios (σ_ϵ/σ_s) for (a) 50 points, (b) 100 points, and (c) 1000 points.

density estimation, which, in turn, influences the MI estimates. The selection of appropriate smoothing parameter needs to be guided by the end use of the density estimates. Here we use the optimal smoothing parameter for a Gaussian kernel (h_o) with KDE given in Eq. (8). We investigate the effects of h on nonlinear CC estimates from KDE by selecting different values of h around h_o. For KNN, the number of nearest neighbors (k) governs the overall amount of smoothing in the densities which are subsequently used in entropy estimation given in Eq. (10). Small values of k lead to small bias and large variance whereas large k results in large bias and small variance. Thus, the bias-variance trade-off, which is a common issue encountered in statistical estimation procedures, is also important here. Kraskov *et al.* [13] warned against using large k since the decrease in variance is outweighed by the increase in bias. They proposed k ranging from 2 to 4. Here we use k as 3 for KNN. We evaluate the effects of k on nonlinear CC estimates from KNN by selecting different k values. The results presented here are obtained for two cases: specifically, quadratic and periodic.

For *very short* data, the bias and variance from KDE increase with the increase of h at low noise and all noise levels, respectively [Fig. 6(a)]. At low noise, KDE does not overlap with theoretical CCs. However, KDE with $h=0.75h_o$ and $h=h_o$ performs better at high noise since their 90% confidence bounds overlap with theoretical CCs. The bias and variance from KNN increase as the number of nearest neighbors increase across all noise levels [Fig. 6(b)]. At low noise, the performance of KNN with $k=3$ is the best of all the cases considered here since it has small bias and its confidence bounds overlap with theoretical CCs. At high noise, KNN has large bias and variance for all k. If KNN needs to be used at high noise, $k=3$ appears to be a better choice since it is closer to theoretical CCs as compared to the others and the variances from all k are comparable. Thus, for *very short* data, KDE with $h=0.75h_o$ or $h=h_o$ may be utilized at high noise whereas KNN with $k=3$ seems to be a better choice at low noise.

For *short* data, KNN with $k=3$ performs better at low noise since it has small bias and variance [Fig. 6(c)]. As k increases, the bias increases and the variance decreases at high noise. KNN with all k considered here performs better at high noise but the selection of appropriate k needs to be guided by the acceptable levels of bias and variance. Thus, for *short* data, KNN with $k=3$ is the best since it overlaps exactly with theoretical CCs and its variance does not differ significantly from the others.

V. CONCLUSION AND DISCUSSION

Our results indicate that two MI estimation methods, specifically KDE and KNN, outperform the other methods and estimation procedures in terms of their ability to capture the dependence structure including nonlinear dependence where present. We find that KNN is the best estimator for *very short* data with relatively low noise while KDE works better for *very short* data when the noise levels are higher. A visual examination of the density plots may help in explaining the relative performance of KDE and KNN (Figs. 7–10 in Appendix B). For *short* data, KNN is the best choice for capturing the nonlinear dependence across all noise levels except when the data are generated from chaotic dynamics, where KDE is a better choice at higher noise levels. We surmise that the relative performance of KDE and KNN with respect to various noise levels is a consequence of the bias-variance trade-off. Previous literature suggests that KDE es-

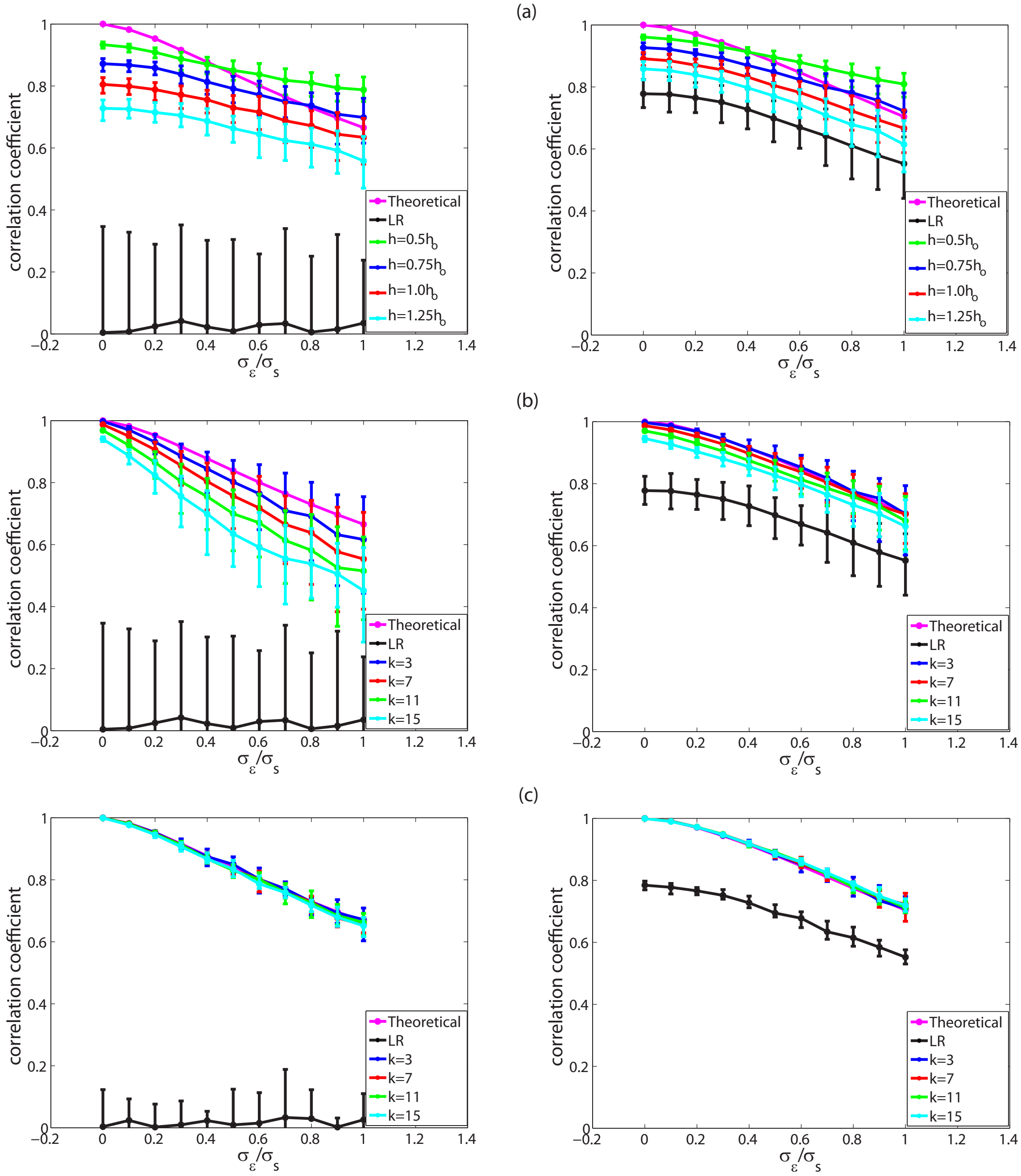

FIG. 6. (Color online) Performance of KDE and KNN with different values of smoothing parameter (h) and number of nearest neighbors (k), respectively. The results from quadratic and periodic functions are presented in the left and right, respectively. (a) KDE with 100 points, (b) KNN with 100 points, and (c) KNN with 1000 points. In (a), h_o is the optimal smoothing parameter for a Gaussian kernel given in Eq. (8).

timates can often be highly biased if the particular KDE recipe used here is followed [12], while KNN estimates can have significant variance when the number of nearest neighbors (k) is set to low values—e.g., $k=3$ as used in this study. The bias in the KDE estimates dominates the variance of the estimates for low noise-to-signal ratios. The KNN performs relatively better for low noise levels since its bias and variance are lower than that from KDE. However, the converse is true for high noise-to-signal ratios, and hence the KDE performs relatively better. For high noise, the variance dominates because of the noise in the data but the variance associated with $k=3$ for KNN increases dramatically. One way to address the large variance from KNN is to use a much larger value of k, but it would also increase the bias.

In general, the above discussions and pointers appear to suggest that the results for nonlinear dependence obtained from KDE and KNN could reflect the lower bounds of what may be potentially achievable through improvements or intelligent combinations of KNN and KDE. Specifically, both the KDE and KNN estimates can be potentially improved by

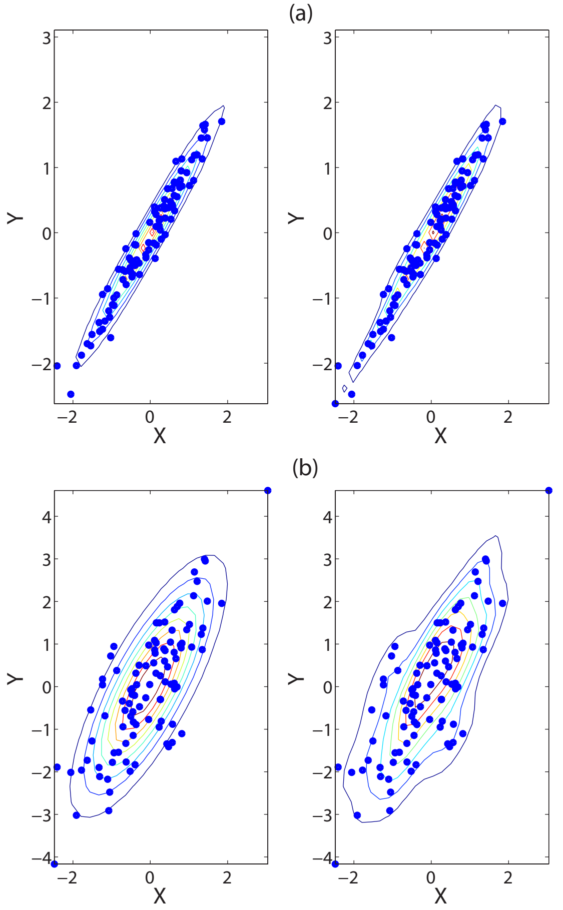

FIG. 7. (Color online) Linear: normal (left) and kernel (right) densities with different noise-to-signal ratios $(\sigma_\epsilon/\sigma_s)$ with 100 points. For kernel density, a Gaussian kernel with optimal smoothing parameter h_o given in Eq. (8) is used. (a) $\sigma_\epsilon/\sigma_s=0.2$. (b) $\sigma_\epsilon/\sigma_s=0.9$. The linear dependence structure can be seen clearly in (a) but cannot be readily identified in (b) based on an eye estimation.

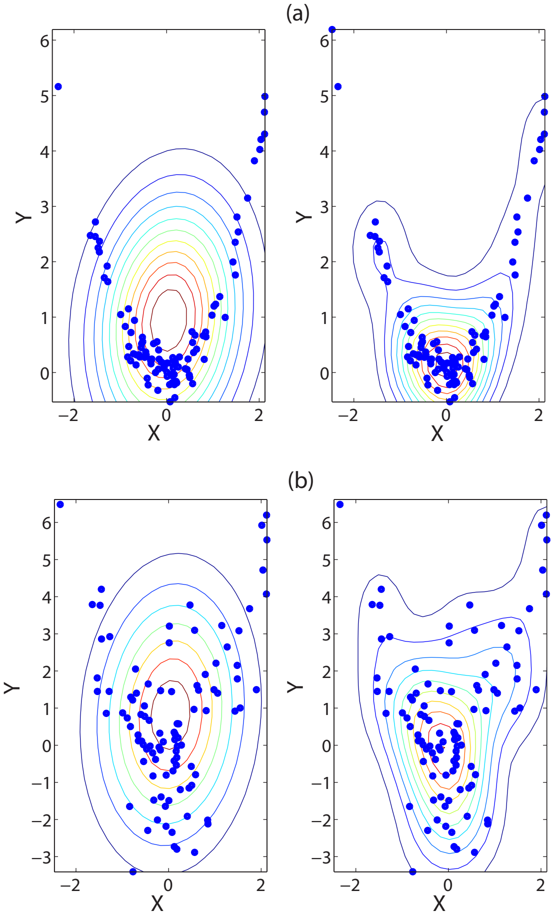

FIG. 8. (Color online) Quadratic: normal (left) and kernel (right) densities with different noise-to-signal ratios $(\sigma_\epsilon/\sigma_s)$ with 100 points. For kernel density, a Gaussian kernel with optimal smoothing parameter h_o given in Eq. (8) is used. (a) $\sigma_\epsilon/\sigma_s=0.2$. (b) $\sigma_\epsilon/\sigma_s=0.9$. At low noise, such as in (a), the nonlinear dependence can be clearly seen as shown by the kernel density. However, at high noise, such as in (b), the dependence structure is not readily discernible visually from the kernel density.

utilizing a plug-in method for kernel, smoothing parameter (h), or k selection. Such plug-in procedures would cause additional estimation variance but may reduce the overall MSE of estimation. However, the development or utilization of procedures for the selection of optimal kernels, smoothing parameters, or nearest neighbors may be rather involved and hence is an area of future research.

We have presented preliminary justifications for the relative performance of the MI estimation methods based on considerations like the bias-variance trade-off and the nature of the approximations underlying the estimation procedures. Our evaluation suggests that the development of guidance for the use of the most suitable estimation procedure may be possible and would depend on known data or domain characteristics and exploratory data analysis. If such guidance can indeed be provided, this could lead to the development of automated or semiautomated procedures for the choice of the most appropriate estimation procedure and the corresponding parameters. However, significant future research on multiple test cases comprising simulated and real data may be necessary before such procedures can be deployed in real-world settings.

ACKNOWLEDGMENTS

This research was partially funded by the Laboratory Directed Research and Development Program of the Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. The authors are thankful to Dr. Alexander Kraskov for providing the KNN-based mutual information code. The authors would like to thank two anonymous reviewers for their helpful suggestions which significantly improved the quality of the paper. S.K. was partially funded by the Department of Civil and Environmental Engineering (CEE) at the University of South Florida (USF) in Tampa, FL. A.R.G. acknowledges partial funding from the "Sensor Net" research program at ORNL. The authors would

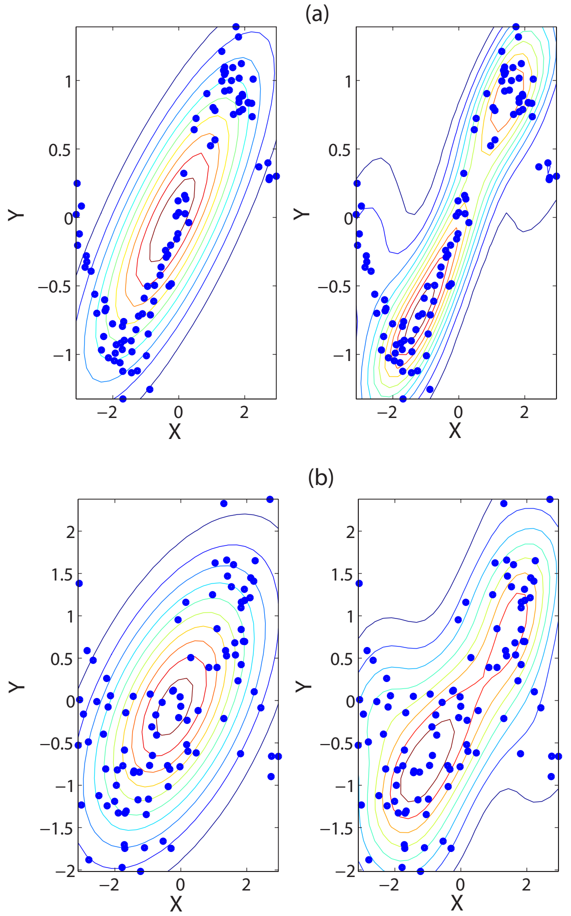

FIG. 9. (Color online) Periodic: normal (left) and kernel (right) densities with different noise-to-signal ratios (σ_ϵ/σ_s) with 100 points. For kernel density, a Gaussian kernel with optimal smoothing parameter h_o given in Eq. (8) is used. (a) $\sigma_\epsilon/\sigma_s=0.2$. (b) $\sigma_\epsilon/\sigma_s=0.9$. With increasing noise levels, the nonlinear dependence structure cannot be identified visually as shown by the kernel density plots.

like to acknowledge Dr. Gabriel Kuhn, Dr. Olufemi Omitaomu, and Dr. Ranga Raju Vatsavai, of ORNL, for their helpful comments.

APPENDIX A: COMPUTATIONS OF THEORETICAL MUTUAL INFORMATION

In this study, we consider four different types of simulations: i.e., linear, quadratic, periodic, and chaotic systems. For the linear, quadratic, and periodic cases, the exact MIs as defined by Eq. (2) can be computed as shown below.

1. Linear

Let $X\sim N(0,1), Y: y_i=x_i+\varepsilon_i$, where $i=1,\ldots,n$; X is iid; and $\varepsilon\sim N(0,\sigma_\varepsilon)$, where σ_ε is the noise level and is iid and independent of X. Let $Z=\varepsilon$, so $Y=X+Z$. Therefore, $H(Y|X)$ can be obtained as

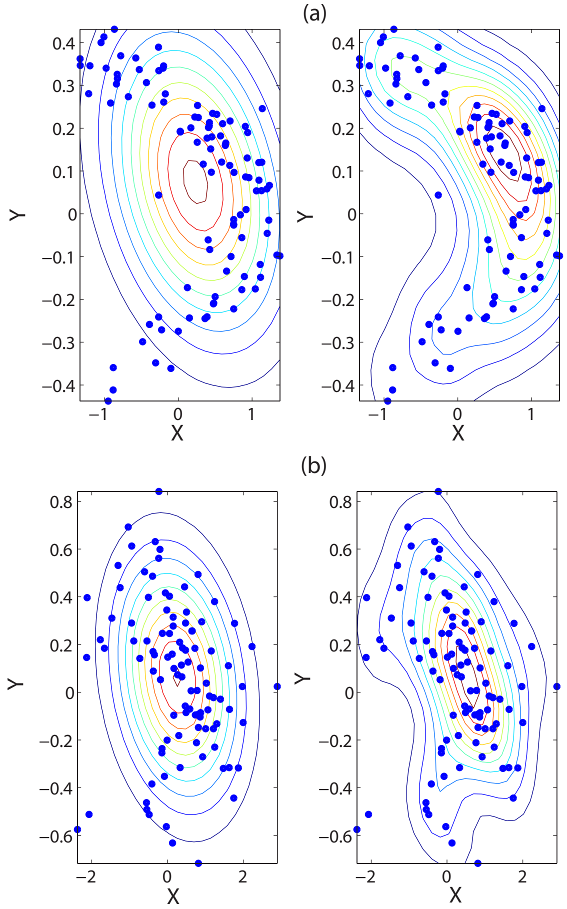

FIG. 10. (Color online) Chaotic: normal (left) and kernel (right) densities with different noise-to-signal ratios (σ_ϵ/σ_s) with 100 points. For kernel density, a Gaussian kernel with optimal smoothing parameter h_o given in Eq. (8) is used. (a) $\sigma_\epsilon/\sigma_s=0.2$. (b) $\sigma_\epsilon/\sigma_s=0.9$. Kernel density plot shows the Henon attractor in (a). However, the Henon attractor cannot be readily distinguished visually in (b).

$$H(Y|X) = H(Z) = 0.5\ \ln(2\pi e {\sigma_\varepsilon}^2).$$

The PDF of Z is

$$p_Z(z) = (2\pi)^{-1/2}(\sigma_\varepsilon)^{-1}\exp\left(\frac{-z^2}{2{\sigma_\varepsilon}^2}\right).$$

The PDF of X is given as

$$p_X(x) = (2\pi)^{-1/2}(\sigma_X)^{-1}\exp\left(\frac{-x^2}{2{\sigma_X}^2}\right),$$

where σ_X, which is called the signal level, is the standard deviation of X.

In order to compute $H(Y)$, the PDF of Y—i.e., $p_Y(y)$—is needed. Since $Y=X+Z$ and X and Z are independent, $p_Y(y)$ can be obtained through the convolution of the PDFs of X and Z given as

$$p_Y(y) = \int_{-\infty}^{\infty} p_X(x) p_Z(y-x) dx. \tag{A1}$$

Solving Eq. (A1), we get

$$p_Y(y) = (2\pi)^{-1/2}({\sigma_X}^2 + {\sigma_\varepsilon}^2)^{-1/2} \exp\left(\frac{-x^2}{2({\sigma_X}^2 + {\sigma_\varepsilon}^2)}\right).$$

Therefore, $H(Y)$ can be given as

$$H(Y) = \int p_Y(y) \ln p_Y(y) dy = 0.5 \ln[2\pi e({\sigma_X}^2 + {\sigma_\varepsilon}^2)].$$

Substituting $H(Y|X)$ and $H(Y)$ in Eq. (2), we get

$$I(X;Y) = 0.5 \ln\left(1 + \frac{{\sigma_X}^2}{{\sigma_\varepsilon}^2}\right).$$

2. Quadratic

Let $X \sim N(0,1), Y: y_i = x_i^2 + \varepsilon_i$, where $i = 1, \ldots, n$; X is iid, and $\varepsilon \sim N(0, \sigma_\varepsilon)$, where σ_ε is the noise level and is iid and independent of X. Let $U = X^2$ and $Z = \varepsilon$, so $Y = U + Z$. Therefore, $H(Y|X)$ can be obtained as

$$H(Y|X) = H(Z) = 0.5 \ln(2\pi e {\sigma_\varepsilon}^2).$$

The PDF of Z is given as

$$p_Z(z) = (2\pi)^{-1/2}(\sigma_\varepsilon)^{-1} \exp\left(\frac{-z^2}{2{\sigma_\varepsilon}^2}\right).$$

The PDF of U is given as

$$p_U(u) = \begin{cases} (2\pi)^{-1/2}(u)^{-1/2} \exp\left(\dfrac{-u}{2}\right), & u > 0, \\ 0, & \text{otherwise}. \end{cases}$$

In order to compute $H(Y)$, the PDF of Y—i.e., $p_Y(y)$—is needed. Since $Y = U + Z$ and U and Z are independent, $p_Y(y)$ can be obtained through the convolution of the PDFs of U and Z given as

$$p_Y(y) = \int_{-\infty}^{\infty} p_U(u) p_Z(y-u) du. \tag{A2}$$

$H(Y)$ is computed as $H(Y) = \int p_Y(y) \ln p_Y(y) dy$, where $p_Y(y)$ in Eq. (A2) is solved using numerical integration for different values of σ_ε. We obtain $I(X;Y)$ by substituting $H(Y|X)$ and $H(Y)$ in Eq. (2).

3. Periodic

Let $X \sim \text{uniform}(-\pi, \pi), Y: y_i = \sin(x_i) + \varepsilon_i$, where $i = 1, \ldots, n$; X is uniformly distributed between $-\pi$ and π; and $\varepsilon \sim N(0, \sigma_\varepsilon)$, where σ_ε is the noise level and is iid and independent of X. Let $V = \sin(X)$ and $Z = \varepsilon$, so $Y = V + Z$. Therefore, $H(Y|X)$ can be obtained as

$$H(Y|X) = H(Z) = 0.5 \ln(2\pi e {\sigma_\varepsilon}^2).$$

The PDF of Z is given as

$$p_Z(z) = (2\pi)^{-1/2}(\sigma_\varepsilon)^{-1} \exp\left(\frac{-z^2}{2{\sigma_\varepsilon}^2}\right).$$

The PDF of V is given as

$$p_V(v) = (\pi)^{-1}(1 - v^2)^{-1/2} \quad \text{for } 0 \leqslant v < 1.$$

In order to compute $H(Y)$, the PDF of Y—i.e., $p_Y(y)$—is needed. Since $Y = V + Z$ and V and Z are independent, $p_Y(y)$ can be obtained through the convolution of the PDFs of V and Z given as

$$p_Y(y) = \int_{-\infty}^{\infty} p_V(v) p_Z(y-v) dv. \tag{A3}$$

$H(Y)$ is computed as $H(Y) = \int p_Y(y) \ln p_Y(y) dy$, where $p_Y(y)$ in Eq. (A3) is solved using numerical integration for different values of σ_ε. We obtain $I(X;Y)$ by substituting $H(Y|X)$ and $H(Y)$ in Eq. (2).

APPENDIX B: FIGURES SHOWING NORMAL AND KERNEL DENSITY PLOTS

Figures showing normal and kernel density plots are given in Figs. 7–10.

[1] M. Barahona and C.-S. Poon, Nature (London) **381**, 215 (1996).

[2] S. Khan, A. R. Ganguly, and S. Saigal, Nonlinear Processes Geophys. **12**, 41 (2005).

[3] J. Theiler and P. E. Rapp, Electroencephalogr. Clin. Neurophysiol. **98**, 213 (1996).

[4] K. Lehnertz and C. E. Elger, Phys. Rev. Lett. **80**, 5019 (1998).

[5] D. A. Smirnov and B. P. Bezruchko, Phys. Rev. E **68**, 046209 (2003).

[6] N. Nicolaou and S. J. Nasuto, Phys. Rev. E **72**, 063901 (2005).

[7] R. Quian Quiroga, A. Kraskov, and P. Grassberger, Phys. Rev. E **72**, 063902 (2005).

[8] R. Quian Quiroga, A. Kraskov, T. Kreuz, and P. Grassberger, Phys. Rev. E **65**, 041903 (2002).

[9] S. Khan, A. R. Ganguly, S. Bandyopadhyay, S. Saigal, D. J. Erickson III, V. Protopopescu, and G. Ostrouchov, Geophys. Res. Lett. **33**, L24402 (2006).

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

[11] Y. I. Moon, B. Rajagopalan, and U. Lall, Phys. Rev. E **52**, 2318 (1995).

[12] B. Rajagopalan, U. Lall, and D. Tarboton, Stochastic Environ. Res. Risk Assess. **11**, 523 (1997).

[13] A. Kraskov, H. Stögbauer, and P. Grassberger, Phys. Rev. E **69**, 066138 (2004).

[14] M. M. V. Hulle, Neural Comput. **17**, 1903 (2005).

[15] C. J. Cellucci, A. M. Albano, and P. E. Rapp, Phys. Rev. E **71**, 066208 (2005).

[16] G. A. Darbellay and I. Vajda, IEEE Trans. Inf. Theory **45**, 1315 (1999).

[17] N. Kwak and C.-H. Choi, IEEE Trans. Pattern Anal. Mach. Intell. **24**, 1667 (2002).

[18] Q. Wang, Y. Shen, and J. Q. Zhang, Physica D **200**, 287 (2005).

[19] G. A. Darbellay, Comput. Stat. Data Anal. **32**, 1 (1999).

[20] D. R. Brillinger, Braz. J. Probab. Stat. **18**, 163 (2004).

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

[22] R. Steur, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, Bioinformatics **81**, S231 (2002).

[23] J. Xu, Z.-R. Liu, R. Liu, and Q.-F. Yang, Physica D **106**, 363 (1997).

[24] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).

[25] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes* (Holden-Day, San Francisco, 1964).

[26] H. Joe, J. Am. Stat. Assoc. **84**, 157 (1989).

[27] C. Granger and J. Lin, J. Time Ser. Anal. **15**, 371 (1994).

[28] S. J. Schiff, P. So, T. Chang, R. E. Burke, and T. Sauer, Phys. Rev. E **54**, 6708 (1996).

[29] H. K. M. Meeren, J. P. M. Pijn, E. L. J. M. V. Luijtelaar, A. M. L. Coenen, and F. H. L. da Silva, J. Neurosci. **22**, 1480 (2002).

[30] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall/CRC, London, 1986).