Correntropy-based partial directed coherence for testing multivariate Granger causality in nonlinear processes

Rohit Kannan and Arun K. Tangirala*

Department of Chemical Engineering, Indian Institute of Technology Madras, Chennai 600004, India

(Received 1 October 2012; revised manuscript received 23 March 2014; published 30 June 2014)

Identification of directional influences in multivariate systems is of prime importance in several applications of engineering and sciences such as plant topology reconstruction, fault detection and diagnosis, and neurosciences. A spectrum of related directionality measures, ranging from linear measures such as partial directed coherence (PDC) to nonlinear measures such as transfer entropy, have emerged over the past two decades. The PDC-based technique is simple and effective, but being a linear directionality measure has limited applicability. On the other hand, transfer entropy, despite being a robust nonlinear measure, is computationally intensive and practically implementable only for bivariate processes. The objective of this work is to develop a nonlinear directionality measure, termed as KPDC, that possesses the simplicity of PDC but is still applicable to nonlinear processes. The technique is founded on a nonlinear measure called *correntropy*, a recently proposed generalized correlation measure. The proposed method is equivalent to constructing PDC in a kernel space where the PDC is estimated using a vector autoregressive model built on correntropy. A consistent estimator of the KPDC is developed and important theoretical results are established. A permutation scheme combined with the sequential Bonferroni procedure is proposed for testing hypothesis on absence of causality. It is demonstrated through several case studies that the proposed methodology effectively detects Granger causality in nonlinear processes.

DOI: 10.1103/PhysRevE.89.062144 PACS number(s): 02.50.Ey, 02.50.Sk, 05.45.Tp

I. INTRODUCTION

Identification of causal relationships in multivariate systems is an important problem in many scientific areas. Identifying whether the influence between a pair of signals is along a direct or an indirect path is a key step in reconstructing the structure of a process. Determining the process structure using process flow sheets is usually a tedious process and, thus, it is practical to adopt a data-driven approach. Furthermore, a data-driven analysis also reveals the strength of connectivities. The problem of connectivity reconstruction appeals to a diverse set of scientific areas and applications, namely, plant topology reconstruction [1,2], fault detection and diagnosis [3], econometrics [4,5], neurosciences [6,7], and climatology [8]. Since most physical systems are nonlinear and multivariable in nature, it is necessary to work with an efficient causality detection method that can handle nonlinearities in a multivariable framework.

A variety of data-driven causality detection measures have emerged over the past two decades for analyzing interrelationships in multivariate processes. The majority of these measures rely on the concept of Granger causality (GC). Among them, linear measures, particularly partial directed coherence (PDC) [7] and directed transfer function (DTF) [6] (frequency domain), have proven to be effective in detecting direct and total couplings between variables. Partial directed coherence and directed transfer function are normalized measures of the direct and total influence exerted by a source in the jth channel on a variable in the ith channel of a multivariate process. In reconstructing the connectivity, PDC is preferred to DTF since the latter measures both the direct and indirect influences between a pair of variables [7,9]. However, the non-normalized directed transfer function can be split into direct and indirect transfer functions from which one can obtain a causal measure [10]. It is shown in [10] that the squared magnitude of the direct transfer function, called the direct energy transfer, is an exact measure of the connectivity strength unlike PDC which offers a qualitative measure. Although PDC and DTF are theoretically applicable only to linear systems, they have been applied to nonlinear systems with some success [11,12]. It has been observed that linear Granger causality detection measures work well when a good linear approximation of the nonlinear system is available in the working regime [13,14]. However, the success of linear measures in detecting nonlinear causal relationships depends on the extent of nonlinearity. The use of PDC for nonlinear systems can lead to spurious links [15,16] as shown in Sec. IV. While nonparametric versions of PDC exist [17], they require a comparatively larger amount of data and a reliable significance level analysis to avoid spurious detections.

A comprehensive review of causality measures for nonlinear systems is presented in [18]. A predominant number of these methods such as transfer entropy [19], correction, and partitioning methods fall under the class of information-theoretic approaches [18], while others such as neural network [20] and kernel-based methods [2,21,22] are primarily based on the construction of nonlinear predictor structures for the variables of interest. Information-theoretic approaches rely on estimates of entropy, which can be obtained in a parametric or a non-parametric fashion. A comprehensive review of the various commonly employed information-theoretic causality detection methods is presented in [18]. While the linear causality measures have gained wide acceptance for multivariable systems, nonlinear multivariate measures are associated with a great deal of computational burden. An added impediment is the lack of tractable statistical methods for computing errors in the resulting estimates, a standard difficulty with nonlinear methods. Transfer entropy, introduced by Schreiber

*arunkt@iitm.ac.in

[19], is a popular nonparametric method for nonlinear causality detection. Transfer entropy is a Kullback-Leibler divergence of conditional probability density functions (pdfs) [18]. Thus the estimation of transfer entropy requires estimates of the conditional pdfs concerning the variables in a nonparametric fashion. Schreiber [19] suggested the use of correlation integrals and kernel density estimators for conditional pdf estimation. The estimation of transfer entropy requires the specification of the number of past observations (based on the order of the system) of the cause and effect variables to be used in the construction of the joint pdfs. For multivariable systems, the number of arguments in the joint pdfs is larger than that for bivariate processes. The size of data and the computational effort required for an accurate estimate of transfer entropy increase exponentially with the dimensionality of the joint pdf [23], leading to what is called the *curse of dimensionality* [18]. A practical application of transfer entropy is thus restricted to bivariate processes, while an extension to trivariate processes is explored in [24]. Further, it is observed that the kernel width used in density estimation significantly affects the quality of the pdf estimates [25,26].

Diks and DeGoede [27] introduced another measure of Granger causality in nonlinear systems using the concept of correlation integrals, and showed that their measure is closely related to the information-theoretic measures. Their proposed method essentially involves the computation of conditional entropy using correlation integrals, which provide a nonparametric estimate of GC. Paluš *et al.* [28] use conditional mutual information as a measure for inferring causal relations, which is essentially equivalent to using transfer entropy. Estimation procedures of several other entropy-based methods are discussed in [18].

In a study by Marinazzo *et al.* [21], a kernel-based method for GC in nonlinear systems is proposed and simulations are performed on rat EEG data using polynomial and Gaussian kernels. Marinazzo *et al.* have also employed kernel based techniques for the analysis of dynamical networks [2] and neural data [22]. However, the authors in [22] report certain cases where linear Granger causality measures outperform their kernel-based GC approach. Ancona *et al.* [20] have addressed the problem of bivariate nonlinear causality detection by constructing a radial basis function network over the time series data. A serious drawback of most of the above approaches is that the analysis is restricted to bivariate systems and an extension to the multivariate case has not been outlined. From an overall viewpoint, existing nonlinear measures are quite restrictive in terms of practical implementation.

Santamaria *et al.* [26] proposed a generalized correlation measure named *correntropy*, which combines both the time structure and the statistical distribution of the random variables. Unlike correlation, correntropy captures the higher-order moments of the probability density function (pdf) of a random variable and can thus identify nonlinear characteristics. Motivated by the computational simplicity and the ability of correntropy to detect nonlinearities, Park and Príncipe [29] adopted an approach for causality detection in nonlinear processes wherein they employ the standard time domain GC detection approach using correntropy as a measure of correlation in the kernel space. By constructing an autoregressive model of an appropriate order in the kernel space, the authors in [29] check for causal links by looking at the estimates of variances of the error terms of the predictions. However, the authors restrict their analysis to bivariate processes.

The major contribution of this work is an efficient measure for causality detection in multivariate nonlinear processes. This article explores the extension of PDC to nonlinear systems using the definition of a generalized correlation function (correntropy). An efficient method for the implementation of the proposed measure is provided and important theoretical results are established.

The organization of this work is as follows. Section II presents a brief review of the fundamental concepts necessary for the developments in this article. Section III formulates the problem of estimation of KPDC and provides an efficient implementation procedure along with guidelines for choosing key parameter values. The simulation results in Sec. IV demonstrate the ability of kernel PDC when applied to several nonlinear multivariable systems of varied nature. This article ends with a few concluding remarks in Sec. V.

II. BACKGROUND

In this section, a brief review of PDC and its estimation procedure from a VAR model is provided. A brief introduction to correntropy along with interpretations is also included.

A. Partial directed coherence (PDC)

Coherence is a commonly used frequency domain measure of the linear relationship between two variables [30] defined by

$$C_{ij}(\omega) = \frac{|S_{ij}(\omega)|^2}{S_{ii}(\omega)S_{jj}(\omega)}, \tag{1}$$

where $\mathbf{S}(\omega)$ denotes the cross-spectral power-density matrix. The multivariate counterpart of coherence is partial coherence, which measures the coherence between two signals after accounting for the effects of the other signals (confounding) in the process. Coherence measures the total association between two variables, which could be due to direct, indirect, or that arising from the confounding by other variables, while partial coherence is a measure of the direct association. However, both coherence and partial coherence, being symmetric, fail to provide the directionality information.

The directionality information was first addressed by Saito and Harashima [31], who proposed the notion of directed coherence using information theoretic arguments for bivariate series. Directed coherence is a decomposition of coherence into components involving directed influences along feedforward and feedback pathways. The directed coherence [7] between the source variable x_j and the effect variable x_i, for diagonal covariance matrices Σ_e, is given by

$$\tilde{\bar{\gamma}}_{ij}(\omega) = \frac{\sigma_{jj} h_{ij}(\omega)}{\sqrt{S_{ii}(\omega)}}, \tag{2}$$

where the quantities $h_{ij}(\omega)$ and $S_{ij}(\omega)$ denote the ijth elements of the transfer function matrix $\mathbf{H}(\omega)$ and the spectral matrix $\mathbf{S}(\omega)$, respectively (see Appendix B).

For nondiagonal noise covariance matrices, a directed coherence function is defined as

$$\tilde{\gamma}_{ij}(\omega) = \frac{h_{ij}(\omega)}{\sqrt{h_{i.}(\omega)\Sigma_e h_{i.}^*(\omega)}} = \frac{h_{ij}(\omega)}{S_{ii}(\omega)}, \tag{3}$$

where $h_{i.}$ refers to the ith row of $\mathbf{H}(\omega)$.

Similarly, a partial directed coherence function, based on the decomposition of partial coherence, is defined as

$$\tilde{\pi}_{ij}(\omega) = \frac{\bar{a}_{ij}(\omega)}{\sqrt{\bar{a}_{.j}^*(\omega)\Sigma_e^{-1}\bar{a}_{.j}(\omega)}} = \frac{(H)_{ij}^{-1}(\omega)}{\sqrt{(S)_{jj}^{-1}(\omega)}}, \tag{4}$$

where $\bar{a}_{ij}$ and $\bar{a}_{.j}$ denote the ijth element and jth column of $\bar{\mathbf{A}}(\omega) = \mathbf{H}^{-1}(\omega)$.

On forsaking the covariance term in the partial directed coherence function, one obtains PDC [7]

$$\pi_{ij}(\omega) = \frac{\bar{a}_{ij}(\omega)}{\bar{a}_{.j}^*(\omega)\bar{a}_{.j}(\omega)} = \frac{\bar{a}_{ij}(\omega)}{\sum_{i=1}^m |\bar{a}_{ij}(\omega)|^2}. \tag{5}$$

The estimation of the above quantities are carried out by a VAR modeling of the jointly stationary process (see Appendix A). From the expression for PDC in Eq. (5), one can see that the normalization is with respect to the effect variable unlike in the directed transfer function where the normalization is with respect to the source:

$$\sum_{i=1}^{m} |\pi_{ij}(\omega)|^2 = 1. \tag{6}$$

Although PDC provides structural information about the system, it is not a quantitative measure of the level of interaction unlike the direct energy transfer [10]. Further, PDC provides consistent estimates only in the cases when the errors in the variables are uncorrelated and there is no instantaneous causality.

B. Correntropy

Most common measures of similarity between two random variables, such as correlation (in the time domain) and coherence (in the frequency domain), are restricted to second-order statistics. These statistics are easy to estimate and implement. The effectiveness of these statistics, however, depends heavily on the assumptions of Gaussianity and linearity. Various information theoretic measures such as mutual information exist, which capture information pertaining to the higher-order moments of the probability density functions (pdfs) describing the random variables.

Physical processes of interest are composed of two principal characteristics: statistical distribution and time structure. Existing measures of similarity for stochastic processes either incorporate information about the time structure or the statistical distribution, but not both. Nonlinearity of signals is generally associated with non-Gaussianity, since most measures do not have the ability to distinguish between statistical distributions. Inspired by information theoretic learning methods, Santamaria *et al.* [26] developed a generalized correlation measure named correntropy, which incorporates information from higher-order statistics apart from the time structure. Correntropy is broadly defined in terms of the inner product (correlation) of two vectors in a higher-dimensional space, which is efficiently described by a reproducing kernel. The definition of the generalized correlation function proposed by Santamaria *et al.* [26] makes use of the Gaussian kernel as the kernel function.

The kernel function is a function which satisfies Mercer's theorem, inducing a nonlinear mapping ϕ which transforms data from the input space to a higher-dimensional reproducing kernel Hilbert space (RKHS) $\mathbf{F}$,

$$k_\sigma(x,y) = \langle \phi(x), \phi(y) \rangle_{\mathbf{F}}, \tag{7}$$

where σ is a parameter of the kernel function. For the translation-invariant Gaussian kernel, which is used in the definition of correntropy, we have

$$k_\sigma(x,y) = k_\sigma(x-y) = \frac{1}{\sqrt{2\pi}\sigma}\exp - \left(\frac{\|x-y\|^2}{2\sigma^2}\right), \tag{8}$$

where σ is the kernel width.

A general form of correntropy between two scalar random variables (also called cross correntropy) X and Y is defined as [26]

$$V(X,Y) = E[k(X-Y)]. \tag{9}$$

On using the Taylor's series expansion for the Gaussian function, the autocorrentropy function reduces to

$$\begin{aligned} V(t_1,t_2) &= E[k_\sigma(x[t_1]-x[t_2])] \\ &= \frac{1}{\sqrt{2\pi}\sigma}\sum_{n=0}^{\infty}\frac{(-1)^n}{2^n\sigma^{2n}n!}E[\|x[t_1]-x[t_2]\|^{2n}]. \end{aligned} \tag{10}$$

Thus, for correntropy to be a function of the lag $(t_1 - t_2)$ alone, we require the process to be strictly stationary on all the even moments. This is a weaker requirement compared to strict stationarity and a stronger requirement compared to wide-sense stationarity, and shall be assumed throughout our analysis. An estimate of autocorrentropy can be obtained from

$$\hat{V}[l] = \frac{1}{N-l+1}\sum_{n=l}^{N} k(x[n]-x[n-l]). \tag{11}$$

The above estimator of autocorrentropy is both unbiased and consistent. A similar expression holds for the estimator for cross correntropy,

$$\hat{V}(X,Y)[l] = \frac{1}{N-l+1}\sum_{n=l}^{N} k(x[n]-y[n-l]). \tag{12}$$

Since correntropy incorporates information about the higher-order moments of the random variable, it is able to distinguish between processes with different underlying distributions. Further, since correntropy is a (linear) measure of similarity (correlation) in a higher-dimensional space, it is observed that correntropy is able to capture the nonlinearities in the system unlike correlation, which is a linear measure in the input space.

From the expression for correntropy in Eq. (10), one can see that, as the kernel width σ increases, the higher-order moments decay and correntropy reduces to the conventional correlation function. In order to obtain a meaningful interpretation of

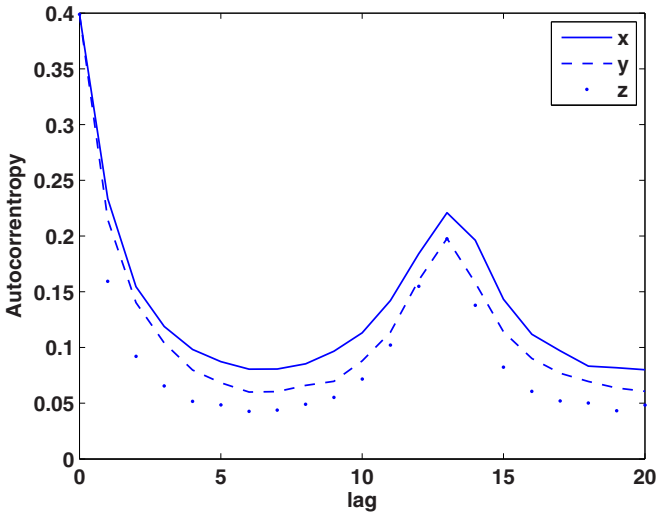

FIG. 1. (Color online) Autocorrentropy estimates for the system defined in Eq. (13) for $R = 28$, $\sigma = 10$, and $b = \frac{10}{3}$.

similarity, the kernel width is tuned using Silverman's rule [32,33] as a guideline.

An example of correntropy estimation for the chaotic Lorenz attractor system [26] described by

$$\dot{x} = \sigma(y - x), \quad \dot{y} = -y - xz + Rx, \quad \dot{z} = xy - bz \quad (13)$$

is shown in Fig. 1. It is observed that the correntropy estimates are able to capture the nonlinear coupling information embedded in the time structure of the process unlike the estimates of correlation, shown in Fig. 2. The reader is directed to [26,34–36] for further discussions on correntropy.

III. KERNEL PDC

In this section, we present the extension of PDC to nonlinear systems using the kernel trick and correntropy. Since PDC and correntropy are efficient indicators of Granger causality and nonlinearity, respectively, the proposed method is efficient in detecting structural and directional relationships in nonlinear systems. As outlined in Sec. II, PDC requires the estimation of a VAR model from data. While an estimate of PDC in the input space is only an indicator of linear Granger causality, estimates of PDC in the kernel space can handle nonlinear processes as well. The kernel space is described by the Gaussian kernel, which is an implicit mapping kernel (because it induces an infinite dimensional nonlinear transformation), i.e., an explicit representation of the data in the kernel feature space is not possible. Let ϕ denote the nonlinear transformation, induced by the Gaussian kernel, from the input space to the kernel space. PDC is estimated in the kernel space as

$$\phi(\mathbf{x}[k]) = -\sum_{r=1}^{p} \mathbf{A}_r^{\phi}\, \phi(\mathbf{x}[k-r]) + \mathbf{v}[k], \quad (14)$$

$$\pi_{ij}^{\phi}(\omega) = \frac{\bar{a}_{ij}^{\phi}(\omega)}{(\bar{a}^{\phi})_{\cdot j}^{*}(\omega)\bar{a}_{\cdot j}^{\phi}(\omega)} = \frac{\bar{a}_{ij}^{\phi}(\omega)}{\sum_{i=1}^{m}\left|\bar{a}_{ij}^{\phi}(\omega)\right|^2}, \quad (15)$$

where

$$\bar{\mathbf{A}}^{\phi}(\omega) = \mathbf{I} - \mathbf{A}^{\phi}(\omega), \quad \mathbf{A}^{\phi}(\omega) = \sum_{r=1}^{p} A_r^{\phi} z^{-r}\Big|_{z=e^{-j\omega}}. \quad (16)$$

Since ϕ is not known explicitly, the transformed values of the random signals are not computable. However, in order to estimate the coefficients of the VAR model in Eq. (14), it is sufficient to know the covariance between the variables in the kernel space at various lags. Thus, using an estimator of correntropy as detailed in Eq. (12), a VAR model can be estimated in the kernel space using data in the input space. A brief outline of the estimation procedure is presented.

A. Estimation of kernel PDC

Consider an m-dimensional multivariate process, denoted by the vector $\mathbf{x} = [x_1\ x_2\ \cdots\ x_m]^{\mathrm{T}}$. Let ϕ be the nonlinear transformation induced by the Gaussian kernel, which transforms data vectors from the input space to the kernel feature space. Let the transformed vector be denoted by $\mathbf{x}^{\phi}[=\phi(\mathbf{x})]$. A VAR model (to be estimated) in the kernel feature space is expressed by

$$\mathbf{x}^{\phi}[k] = -\sum_{r=1}^{p} \mathbf{A}_r^{\phi}\mathbf{x}^{\phi}[k-r] + \mathbf{v}[k], \quad (17)$$

where $\mathbf{v}$ is the driving force (innovations) and the negative sign is added for notational convenience.

Denote by $\mathcal{V}[l]$ the correntropy matrix at lag l, i.e., $\mathcal{V}[l] = [V(X_i, X_j)[l]]_{i,j=1,\ldots,m}$. The coefficient matrices $\mathbf{A}_r$ can be estimated using the Yule-Walker equations [37] in the kernel space,

$$\sum_{r=0}^{p} \mathbf{A}_r^{\phi}\mathcal{V}[r] = \Sigma_v,$$

$$\sum_{r=0}^{q-1} \mathbf{A}_r^{\phi}\mathcal{V}^{\mathrm{T}}[q-r] + \sum_{r'=q}^{p} \mathbf{A}_{r'}^{\phi}\mathcal{V}[r'-q] = \mathbf{0},$$

$$q = 1, \ldots, p. \quad (18)$$

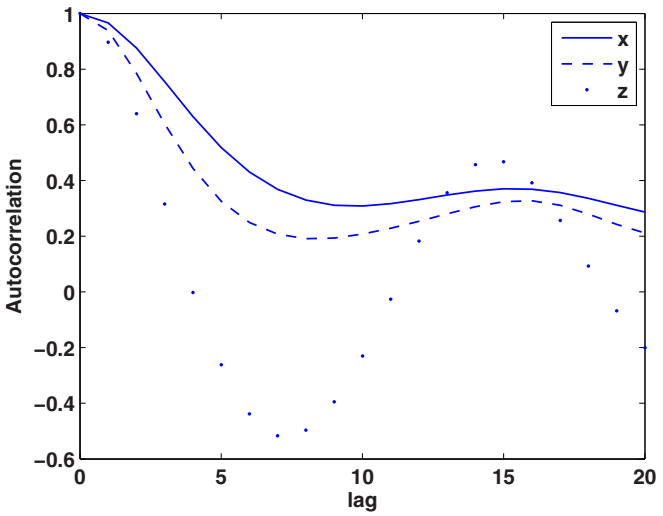

FIG. 2. (Color online) Autocorrelation estimates for the system defined in Eq. (13) for $R = 28$, $\sigma = 10$, and $b = \frac{10}{3}$.

After the estimation of the coefficient matrices, PDC in the kernel space, hereafter referred to as kernel PDC (KPDC), is estimated in a manner similar to PDC using Eqs. (A3) and (5):

$$\mathbf{A}^{\phi}(\omega) = \sum_{r=1}^{p} A_r^{\phi} z^{-r}\Big|_{z=e^{-j\omega}},$$
$$\bar{\mathbf{A}}^{\phi}(\omega) = \mathbf{I} - \mathbf{A}^{\phi}(\omega),$$
$$\Gamma_{ij}(\omega) = \frac{\bar{a}_{ij}^{\phi}(\omega)}{\bar{a}_{.j}^{\phi^*}(\omega)\bar{a}_{.j}^{\phi}(\omega)} = \frac{\bar{a}_{ij}^{\phi}(\omega)}{\sum_{i=1}^{m}\left|\bar{a}_{ij}^{\phi}(\omega)\right|^2}. \quad (19)$$

In order to obtain a meaningful VAR model in the kernel feature space, one has to choose an appropriate value of the kernel width parameter. The kernel width plays an important role in the performance of KPDC (Γ) since it determines the scale at which similarity is quantified. A heuristic to determine a suitable value of the kernel width is given by Silverman's rule [33]

$$\sigma^* = 0.9AN^{-0.2}, \quad (20)$$

where

$$A = \min(\sigma_d, 1.34\times R_{\mathrm{IQ}}).$$

σ_d is the sample standard deviation, R_{IQ} is the interquartile range, and N is the sample size. It is noteworthy that when the kernel width is large ($>20\sigma^*$), estimates of KPDC reduce to those of PDC [34], ensuring that "ordinary PDC" is contained in the KPDC. The associated lemma and its proof are presented at the end of this section.

For a mean-centered estimate, one has to use centered correntropy for estimating KPDC akin to the use of centered correlation for estimating PDC. The centered correntropy U between two random variables X and Y is estimated using [29]

$$\hat{U}(X,Y)[l] = \frac{1}{N-l+1}\sum_{n=l}^{N} k(x[n]-y[n-l]) - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} k(x[i]-y[j]). \quad (21)$$

The latter term in the definition of centered correntropy is called the cross information potential (CIP), which when computed exactly has a computational complexity of $O(N^2)$. When the sample size N is large, computing the exact cross information potential may be infeasible. Seth and Principe [38] suggest the use of an incomplete Cholesky decomposition to approximate the cross information potential when there exists a sufficiently accurate lower-dimensional representation of the kernel gram matrix $\mathcal{K}$ (in other words, only a few eigenvalues of $\mathcal{K}$ are significant). The kernel gram matrix $\mathcal{K}$ is defined as

$$\mathcal{K} = [k_{ij}]_{i,j=1,\ldots,N}, \quad (22)$$

where $k_{ij} = k(x_i, x_j)$. An estimate of the cross information potential can be succinctly represented using the kernel gram matrix as [38]

$$\hat{\mathcal{P}}(X,Y) = \frac{1}{N^2}\mathbf{1}^{\mathrm{T}}\mathcal{K}_{XY}\mathbf{1}, \quad (23)$$

where $\mathcal{P}$ is the cross information potential and $\mathbf{1}$ is a column vector of ones. In order to reduce the computational effort, a Cholesky decomposition of $\mathcal{K}$ is carried out [39]:

$$\mathcal{K} = GG^{\mathrm{T}}. \quad (24)$$

Here G is a lower triangular matrix with positive diagonal entries. If only d out of the N eigenvalues of $\mathcal{K}$ are significant, $\mathcal{K}$ can be approximated using $\bar{G}\bar{G}^{\mathrm{T}}$ where $\bar{G}$ is an approximation of G and is of dimensions $N \times d$. Since $\bar{G}$ can be computed in $O(Nd^2)$ time [38], this approximation can be effectively used to reduce the computational time of correntropy, and hence KPDC.

Lemma 1. For *large* σ, estimates of KPDC asymptotically reduce to that of PDC (i.e., $\hat{\Gamma}_{ij} \overset{N\to\infty}{\longrightarrow} \hat{\pi}_{ij}$ for $\sigma \gg 1$).

Proof. Consider two processes $\{x[k]\}_{k=0}^{N}$ and $\{y[k]\}_{k=0}^{N}$ with zero means and variances σ_x^2 and σ_y^2. Denote by X and Y the random variables generating $\{x[k]\}$ and $\{y[k]\}$. An estimate of the centered correntropy is obtained as

$$\hat{U}_{XY}[l] \equiv \hat{U}(X,Y)[l] = \frac{1}{N-l}\sum_{n=l}^{N-1} k(x[n], y[n-l]) - \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1} k(x[n], y[m]),$$

where

$$k(x,y) = \frac{1}{\sqrt{2\pi}\sigma}\exp - \left(\frac{||x-y||^2}{2\sigma^2}\right).$$

For large σ, using Taylor's series expansion, we have

$$\begin{aligned}\hat{U}_{XY}[l] &\approx \frac{1}{N-l}\sum_{n=l}^{N-1}\frac{1}{\sqrt{2\pi}\sigma}\left(1-\frac{1}{2\sigma^2}(x[n]-y[n-l])^2\right) - \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\frac{1}{\sqrt{2\pi}\sigma}\left(1-\frac{1}{2\sigma^2}(x[n]-y[m])^2\right)\\ &= -\frac{1}{\sqrt{2\pi}\sigma}\frac{1}{2\sigma^2}\left[\hat{\sigma}_x^2+\hat{\sigma}_y^2-2\hat{\sigma}_{xy}[l]\right] + \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{2\sigma^2}\left[\hat{\sigma}_x^2+\hat{\sigma}_y^2-\sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\frac{2x[n]y[m]}{N^2}\right]\\ &= \frac{1}{\sqrt{2\pi}\sigma}\frac{2\hat{\sigma}_{xy}[l]}{2\sigma^2} - \frac{1}{\sqrt{2\pi}\sigma}\frac{1}{2\sigma^2}\left[\frac{2}{N}\sum_{q=-(N-1)}^{(N-1)}\left(1-\frac{|q|}{N}\right)\hat{\sigma}_{xy}[q]\right]\end{aligned}$$

$$\therefore \hat{U}_{XY}[l] \overset{N\to\infty}{\sim} c\hat{\sigma}_{xy}[l]$$

for large σ and finite order correlations between $\{x[k]\}$ and $\{y[k]\}$ (where c is some constant independent of X and Y). Since the centered correntropy estimate asymptotically reduces to the covariance estimate for large σ, from Eqs. (18) and (19), KPDC reduces to PDC for large σ. ■

An analysis of KPDC for linear processes for arbitrary σ is detailed in Appendix C. We now prove the consistency of KPDC estimates defined by (19).

Theorem 2. The proposed estimator of correntropy Γ is consistent.

Proof. In order to show that KPDC estimates are consistent, we are required to show that $\hat{\Gamma}_{ij}(\omega) \to_p \Gamma_{ij}(\omega)$, where $\hat{e}(x_1,\ldots,x_n) \to_p e$ denotes that $\hat{e}$ converges in probability to e, i.e., $\lim_{N\to\infty} P(|\hat{e}(x_1,\ldots,x_N) - e| \leqslant \epsilon) = 1\ \forall\, \epsilon > 0$.

We make use of the following theorem, which can easily be shown to be true using Chebyshev and triangle inequalities.

Theorem 3. Let $\hat{\theta} \to_p \theta$, $\hat{\eta} \to_p \eta$. Then $g(\hat{\theta},\hat{\eta}) \to_p g(\theta,\eta)$ for any real-valued continuous function g.

Proof. See Sec. 2.1 in [40]. ■

An estimate of KPDC between the ith and jth variables is obtained as

$$\hat{\Gamma}_{ij}(\omega) = \frac{\hat{\tilde{a}}^{\phi}_{ij}(\omega)}{\sum_{i=1}^{m} \left|\hat{\tilde{a}}^{\phi}_{ij}(\omega)\right|^2},$$

$$\hat{\tilde{a}}^{\phi}_{ij}(\omega) = \left(I - \sum_{r=1}^{p} A_r z^{-r}\Big|_{z=e^{-j\omega}} \right)_{ij},$$

$$A_r = \mathcal{D}(\hat{U}_{XY}[l]),$$

$$\hat{U}_{XY}[l] = \frac{1}{N-l}\sum_{n=l}^{N-1} k(x[n],y[n-l]) - \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1} k(x[n],y[m]),$$

where $\mathcal{D}(\hat{U}_{XY}[l])$ represents the estimation of the autoregressive coefficients from the estimates of centered correntropy using the Durbin-Levinson procedure.

Using a combination of theorem 3 and the proof of consistency of the DL estimator (under, of course, the right modeling assumptions) in [41], we conclude that the estimator of KPDC as defined is consistent. ■

B. Null hypothesis and the testing scheme

The test of Granger causality involves testing for the significance of off-diagonal terms (at all frequencies) in the KPDC matrix under the null hypothesis:

$$H_0: \quad \Gamma_{ij}(\omega) = 0, \quad 1 \leqslant i,j \leqslant m, i \neq j,\ \forall \omega \in \Omega,$$

where m denotes the number of variables of interest and Ω denotes the set of frequencies at which KPDC is computed.

Obtaining theoretical significance levels, however, appears elusive at present since the distributional properties of KPDC are hard to derive. Therefore, a standard permutation test is employed against the null hypothesis that there is no causal link between any two variables of interest. A permutation test [42] involves randomly permuting the time series associated with the variables of interest to remove possible causal links and then evaluating the test statistic on the permuted data [43]. Subsequently, a significance level is empirically derived to determine possible deviations from the null hypothesis. A consequence of adopting the permutation test is that a system-specific value derived from the data, rather than a universal critical value, under the null hypothesis is applied to the off-diagonal terms.

Testing all the off-diagonal terms of KPDC is clearly a multiple hypothesis testing approach. Thus the usual requirement that the probability of falsely rejecting the null hypothesis should not exceed α (control of the type I error) is replaced by the requirement that the probability of one or more false rejections should not exceed α. The latter probability is called the family wise error rate (FWER), denoted by $\mathcal{F}$ [44]. Hence we require that $\mathcal{F} \leqslant \alpha$ for all possible constellations of true and false hypotheses. This requirement is referred to as *strong control* of the FWER [44].

To obtain significance levels for the family of $\Gamma_{ij}(\omega)$ from a multiple testing standpoint using a permutation test, we compute a suitable percentile value for each $\Gamma_{ij}(\omega)$ using the (sequential) Bonferroni procedure [44,45] based on the following lemma [44].

Lemma 4. (Bonferroni procedure) If, for $i = 1,\ldots,s$, hypothesis H_i is rejected when $\hat{p}_i \leqslant \frac{\alpha}{s}$, then $\mathcal{F} \leqslant \alpha$ (here $\hat{p}_i$ is the p value for testing each H_i).

Proof. See Theorem 9.1.1 in [44]. ■

A significance level which ensures strong control of the FWER is obtained by computing the $100 \times (1 - \frac{\alpha}{s})$ percentile value for each $\Gamma_{ij}(\omega)$. A slightly modified sequential stepdown method for generating significance levels is obtained using Holm's procedure [45].

A possible drawback of the Bonferroni procedure is its low statistical power [44,46,47]. However, the authors note that application of the sequential Bonferroni procedure for generating significance levels for KPDC has been successfully employed in correctly uncovering all causal links for various case studies (Sec. IV). A possible alternative for generating significance levels, based on the so-called *weak control* of the FWER [44], is a false discovery rate (FDR) based approach [48]. Weak control of the FWER might yield a multiple testing procedure with far greater power than the Bonferroni procedure, and might be suitable when KPDC estimates are highly correlated in the frequency domain. Such a statistical approach could be the focus of future investigations.

IV. SIMULATIONS

In this section, the performance of the proposed KPDC measure for detecting Granger causality in nonlinear processes is illustrated on several complex nonlinear systems. For the first five case studies, a single realization of sample size 1000 (unless specified) of a nonlinear process is generated and the performance of the proposed KPDC measure is illustrated. The sixth case study considered is that of a five-variable nonlinear system for which the KPDC is evaluated for several different process realizations of different sample sizes. In addition, the sensitivity of KPDC to the value of the kernel width (σ) used is analyzed for several different process realizations of the same system.

The data generated in each example is normalized such that the mean and variance of each variable are 0 and 1, respectively, so as to obtain standard σ values using Silverman's rule. It is observed that KPDC is an efficient GC estimator and correctly detects the causal relationships in most of the cases for reasonable sample sizes. The (i,j)th plot ($\hat{\Gamma}_{ij}$) indicates the influence of the jth variable on the ith variable as determined

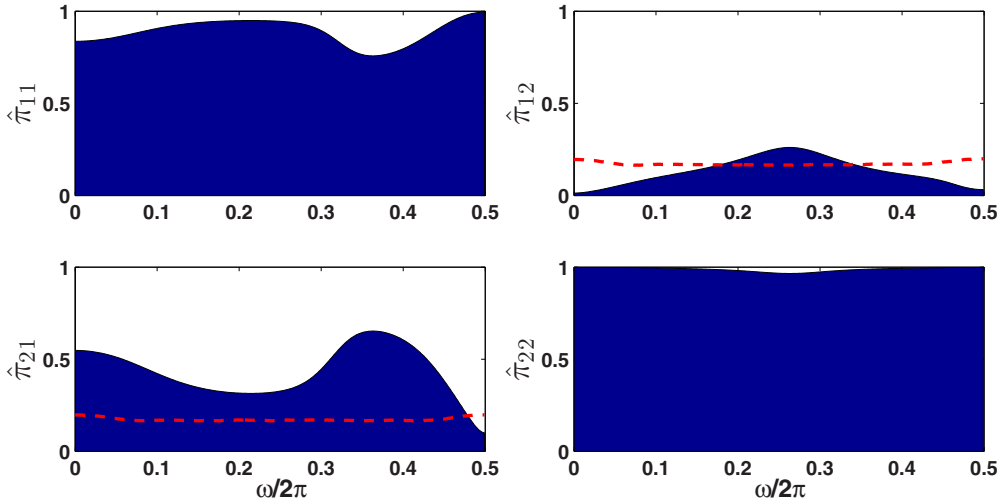

FIG. 3. (Color online) Linear PDC for $a = 1.8$, $s = 0.01$, and $c = 0.2$ for the system defined in Eq. (25).

by KPDC. In each plot, the vertical axis corresponds to the magnitude of PDC/KPDC (bounded between 0 and 1) and the horizontal axis corresponds to the ordinary frequency (ranging from 0 to 0.5). For the sake of visual clarity, the vertical axes labels are omitted and, instead, variable names (corresponding to the matrix of KPDC plots) are included along the rows and columns for systems containing three or more variables. The significance levels correspond to strong control of the FWER at $\alpha = 0.01$, and are indicated for the off-diagonal plots.

Example 1: Coupled noisy logistic maps

To illustrate the effectiveness of KPDC towards detecting GC in nonlinear systems, we consider a bivariate coupled noisy logistic map system [21] defined in Eq. (25). It is observed from the system model that x_1 causes x_2 alone. However, PDC spuriously detects a bidirectional causal relationship between x_1 and x_2, highlighting its inability to deal with nonlinear causal systems, while KPDC correctly uncovers the causal relationship with a kernel width of 0.23 as seen in Figs. 3 and 4:

$$\begin{aligned} x_1[k] &= 1 - a x_1^2[k-1] + s\, e_1[k], \\ x_2[k] &= (1-c)\big(1 - a x_2^2[k-1]\big) \\ &\quad + c\big(1 - a x_1^2[k-1]\big) + s\, e_2[k]. \end{aligned} \tag{25}$$

Example 2: Coupled nonlinear dynamical system

We consider a coupled nonlinear system [29] defined in Eq. (26) to further illustrate the ability of KPDC in detecting GC in nonlinear processes as opposed to PDC. For this system, as observed from the model, x_1 causes x_2 alone. However, PDC fails to detect the causal relationship between x_1 and x_2, while KPDC correctly uncovers the causal relationship with a kernel width of 0.23 as seen in Figs. 5 and 6:

$$\begin{aligned} x_1[k] &= 3.4 x_1[k-1]\big(1 - x_1^2[k-1]\big) e^{-x_1^2[k-1]} + 0.8 x_1[k-2], \\ x_2[k] &= 3.4 x_2[k-1]\big(1 - x_2^2[k-1]\big) e^{-x_2^2[k-1]} \\ &\quad + 0.5 x_2[k-2] + c x_1^2[k-2]. \end{aligned} \tag{26}$$

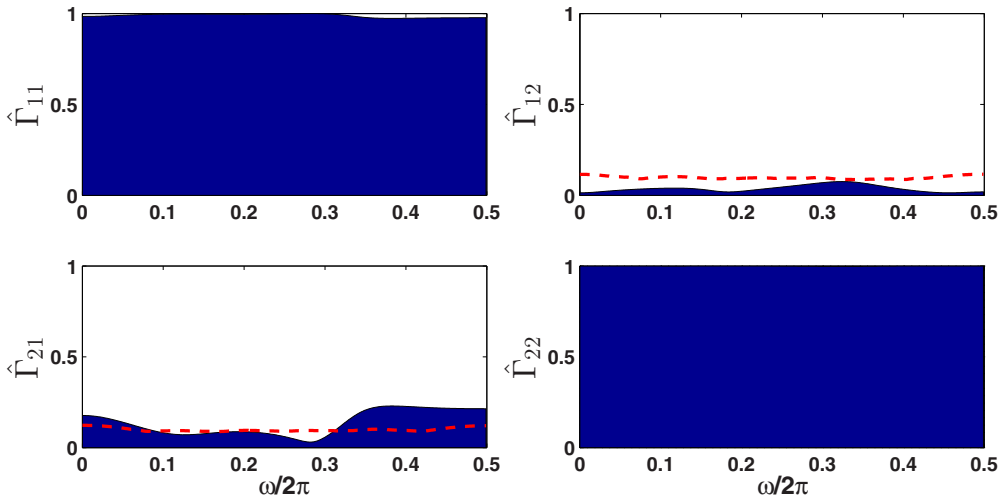

FIG. 4. (Color online) KPDC using $\sigma = 0.23$ for $a = 1.8$, $s = 0.01$, and $c = 0.2$ for the system defined in Eq. (25).

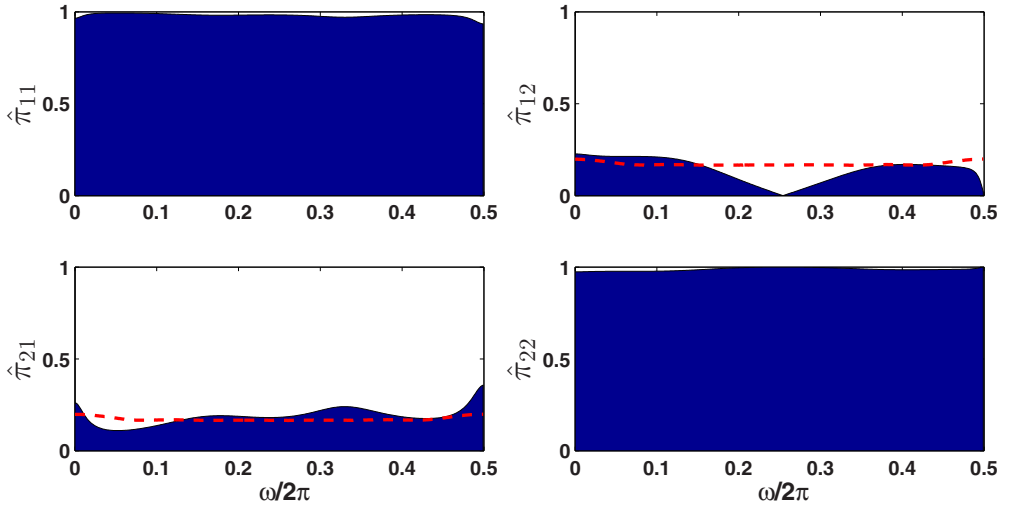

FIG. 5. (Color online) Linear PDC for $c = 0.8$ for the system defined in Eq. (26).

The above two examples show the effectiveness of KPDC over PDC in detecting causal relationships in nonlinear systems, without a great addition of computational effort. In the forthcoming examples, we investigate the performance of KPDC for multivariable nonlinear systems, where confounding has to be addressed in addition to the nonlinearity.

Example 3: Lorenz attractor system

In this example, we consider the popular chaotic Lorenz attractor, defined in Eq. (13), as an example of a nonlinear multivariable system. The parameter values used for the simulation are $R = 28$, $\sigma = 10$, $b = \frac{10}{3}$ and zero-mean, unit-variance Gaussian white noise terms were added to each of the states. For this system, as seen from the model, all variables affect each other except for z not directly influencing x. From the estimates of KPDC in Fig. 7, we observe that all the causal relationships have been correctly identified.

Example 4: Non-Gaussian nonlinear system

We consider the three-variable system defined in Eq. (27) where one of the variables (x_1) has a non-Gaussian underlying distribution [24]. We generate 6000 data points of the system with $x_1[k] \in [4,5]$ falling out of a uniform distribution. To assure stationarity, the first 3000 data points are discarded. The driving forces are generated as $v_1[k], v_2[k] \sim \mathcal{N}(0,0.05)$ with the initial condition $x_3[0] = 0.2$. From the model, it can be seen that x_1 causes both x_2 and x_3 and x_3 causes x_2. KPDC correctly uncovers the three causal links as shown in Fig. 8. This example illustrates that KPDC can uncover nonlinear causal relationships among variables with distributions other than the normal distribution:

$$\begin{aligned} x_2[k] &= 5(x_3[k-1] + 7.2)^2 + 10\sqrt{|x_1[k-1]|} + v_1[k], \\ x_3[k] &= 1 - 2|0.5 - (0.8x_1[k-1] + 0.4\sqrt{x_3[k-1]})| + v_2[k]. \end{aligned} \tag{27}$$

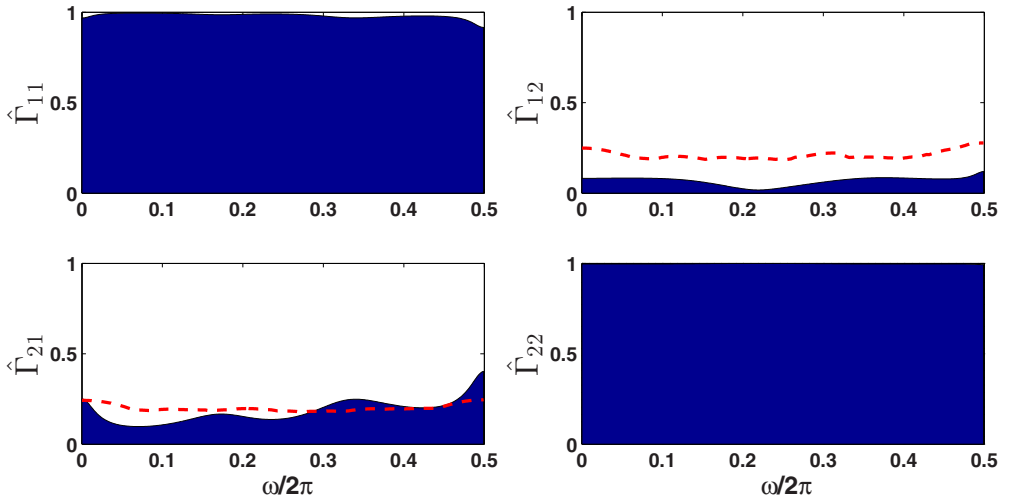

FIG. 6. (Color online) KPDC using $\sigma = 0.23$ for $c = 0.8$ for the system defined in Eq. (26).

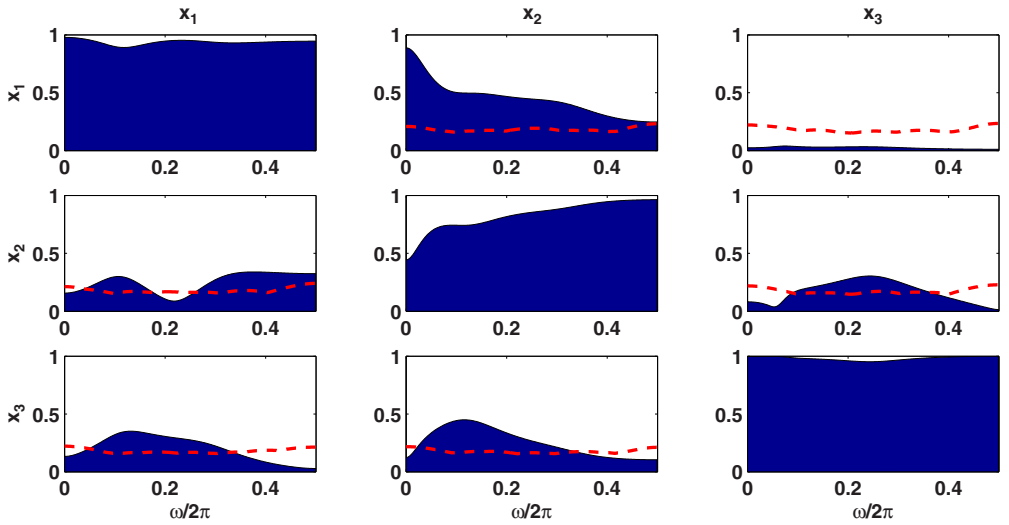

FIG. 7. (Color online) KPDC using $\sigma = 0.11$ for the system defined in Eq. (13) for $R = 28$, $\sigma = 10$, and $b = \frac{10}{3}$.

Example 5: Three-state bioreactor model

In this example, we consider a three-state bioreactor model explored by Lin and Stadtherr [49,50]. The three-state bioreactor model [Eq. (28)] describes the concentration dynamics of the cells (x_1), the substrate (x_2), and the product (x_3),

$$\begin{aligned} \dot{x}_1 &= (\mu - D)x_1, \\ \dot{x}_2 &= D(x_{2f} - x_2) - \frac{\mu x_1}{Y}, \\ \dot{x}_3 &= -Dx_3 + (\alpha\mu + \beta)x_1, \end{aligned} \tag{28}$$

with the growth rate μ being a nonlinear function of both the substrate and the product:

$$\mu = \frac{\mu_{\max}\left[1 - \frac{x_3}{x_{3m}}\right]x_2}{k_s + x_2}. \tag{29}$$

An initial concentration $x_{10} = 6.50$ g/L of the cells was chosen, and the maximum growth rate $\mu_{\max}$ and the saturation parameter k_s were fixed to be 0.46 and 1.1, respectively. The other parameters of the system were fixed as follows: $x_{20} = 5$ g/L, $x_{30} = 15$ g/L, $Y = 0.4$ g/g, $\beta = 0.2$ h^{-1}, $D = 0.202$ h^{-1}, $\alpha = 2.2$ g/g, $x_{3m} = 50$ g/L, and $x_{2f} = 20$ g/L.

The system was simulated using SIMULINK and zero-mean Gaussian white noise terms of variance 0.04 were added to each of the states at each time instant. It is observed that KPDC was able to extract all the causal links in the system using a kernel width of 0.23, as shown in Fig. 9.

Example 6: Coupled map lattice system

We consider a five-variable stochastic coupled map lattice [51], defined in Eq. (30), in which the strength of the unidirectional coupling between pairs of adjacent maps of the lattice is varied from 0.15 to 0.45 in increments of 0.10. We generate 50 realizations of the above system for each of the sample sizes 200, 500, 1000, and 2000 and evaluate the performance of KPDC. The driving forces in (30) are assumed to be normally distributed with zero mean and unit variance.

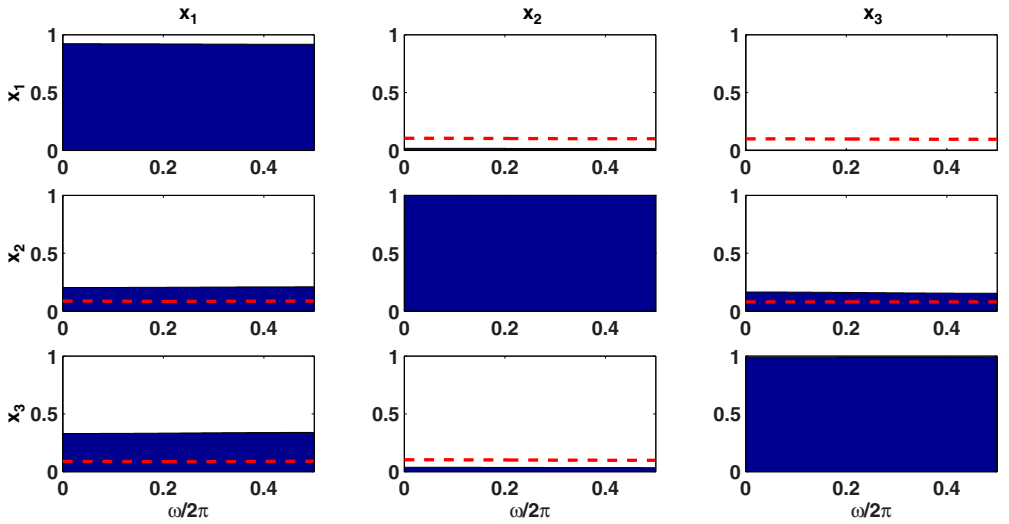

FIG. 8. (Color online) KPDC using $\sigma = 0.23$ for the system defined in Eq. (27).

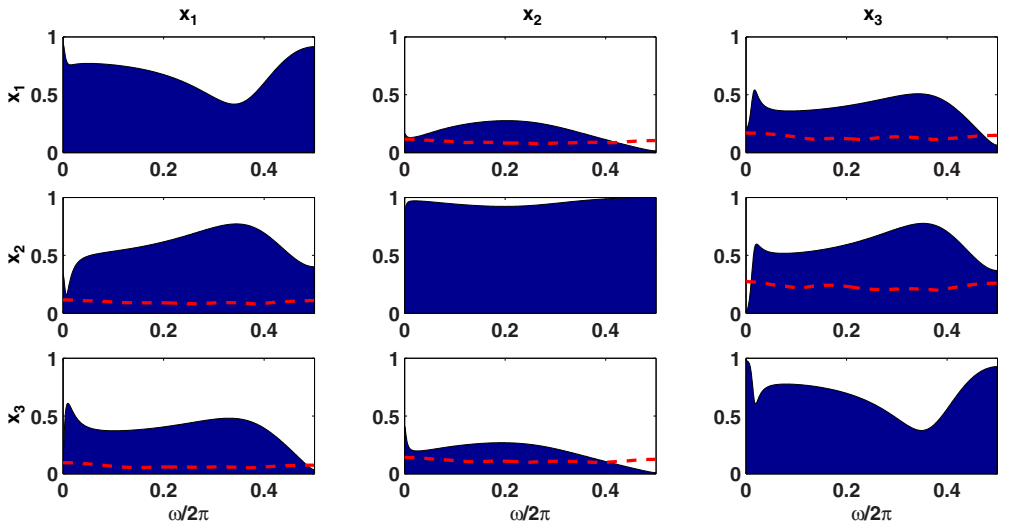

FIG. 9. (Color online) KPDC using $\sigma = 0.23$ for the system defined in Eq. (28).

From the model, it can be seen that x_1 causes x_2, x_2 causes x_3, x_3 causes x_4, and x_4 causes x_5:

$$\begin{aligned}
x_1[k] &= 0.95x_1[k-1] - 0.9025x_1[k-2] + e_1[k],\\
x_2[k] &= 1-2 \mid 0.5 - (0.15x_1[k-1] + 0.35x_2[k-1]) \mid + e_2[k],\\
x_3[k] &= 1-2 \mid 0.5 - (0.25x_2[k-1] + 0.25x_3[k-1]) \mid + e_3[k],\\
x_4[k] &= 1-2 \mid 0.5 - (0.35x_3[k-1] + 0.15x_4[k-1]) \mid + e_4[k],\\
x_5[k] &= 1-2 \mid 0.5 - (0.45x_4[k-1] + 0.05x_5[k-1]) \mid + e_5[k].
\end{aligned} \tag{30}$$

The performance of KPDC for different sample sizes is listed in Table I, and a typical performance of KPDC for a sample size of 1000 is shown in Fig. 10 for comparison with the other case studies. The number of cases in which KPDC was able to correctly detect all the causal relationships for sample sizes 200, 500, 1000, and 2000 were 7, 32, 44, and 49, respectively. It can be seen from Table I that KPDC performs well for reasonably large sample sizes (1000 and 2000) and is fairly robust in correctly establishing the lack of causal relationships between (relevant) variables even for small sample sizes. However, KPDC is unable to consistently detect the relatively weaker causal links (the causal influence of x_1 on x_2) for the smaller sample size cases.

Impact of the kernel width on KPDC performance

The impact of the kernel width (σ) on the performance of KPDC was assessed by evaluating KPDC for 10 realizations of the above system using $0.25\sigma^*$, $0.5\sigma^*$, σ^*, $2\sigma^*$, and $4\sigma^*$ as kernel widths [where σ^* is the kernel width determined using Silverman's rule (20)]. The performance of KPDC for the different kernel widths is listed in Table II. It can be seen from Table II that KPDC performs well for the cases when $0.5\sigma^*$, σ^*, and $2\sigma^*$ were used as kernel widths. However, KPDC is either unable to detect the weaker causal links or detects spurious causal links when the kernel width is either too low or too high. This is consistent with the discussion on correntropy in [34] where the kernel width is likened to a zoom lens.

All the case studies were performed on a computer with a 8.00 GB RAM and a single 2.60 GHz CPU. The computation time for a single realization of the five-variable system in (30) for a sample size of 200 using the kernel width suggested by Silverman's rule was determined to be around 75 min. It was observed that the computation time was an affine function of the sample size, with the 2000 sample size case study using σ^* as the kernel width taking around 1000 minutes per realization, and was an exponential function of the deviation of the kernel width from Silverman's rule, with the 1000 sample size case studies using $0.25\sigma^*$ and $4\sigma^*$ as the kernel widths taking around 5000 min and 210 min, respectively, per realization.

TABLE I. Performance of KPDC for 50 realizations each for sample sizes 200, 500, 1000, and 2000 (in that order, separated by commas) using the kernel width suggested by Silverman's rule. The (i, j)th cell contains the number of instances in which KPDC predicted that the jth variable had a causal influence on the ith variable.

	x_1	x_2	x_3	x_4	x_5
x_1		0,1,0,0	0,0,0,0	0,0,0,0	0,0,0,0
x_2	7,32,45,50		0,0,0,0	0,0,0,0	0,0,0,0
x_3	0,0,0,0	42,50,50,50		0,0,0,0	0,0,0,0
x_4	0,0,0,0	0,0,0,0	50,50,50,50		0,0,0,0
x_5	0,0,0,0	0,0,0,0	0,1,2,1	50,50,50,50	

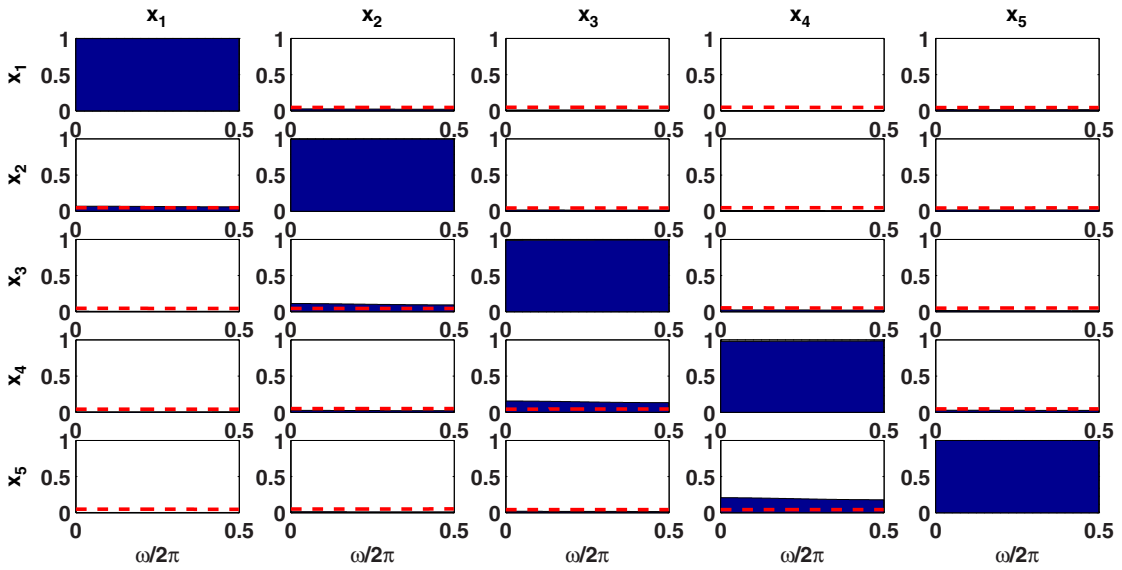

FIG. 10. (Color online) KPDC using $\sigma = 0.23$ for the system defined in Eq. (30) using a sample size of 1000.

The foregoing examples convincingly demonstrate the ability of KPDC in detecting causal links in complex nonlinear systems for reasonable sample sizes. It has been reported in [26] that correntropy is not as sensitive to the value of the kernel width chosen as kernel based density estimation methods [34]. This gives correntropy-based causality detection an advantage over entropy-based methods which use pdf estimates using kernel estimators.

V. CONCLUSIONS

This work presented a measure for detecting Granger causality in a general nonlinear process obtained by extending the idea of partial directed coherence to nonlinear systems using the concept of a generalized correlation function. Issues associated with the existing measures for detecting Granger causality in nonlinear systems were discussed and a motivation for a kernel-based GC measure was provided. The key idea in the proposed method is to handle the nonlinearities through a kernel transformation of the variables while still using PDC in the feature (transformed) space. Theoretical and practical issues concerned with this methodology have been comprehensively studied. An estimator of KPDC has been developed and its consistency has been established. Further, the classical PDC has been shown to be a special case of the proposed method. The main advantage of the proposed method over existing measures is its ability to combine PDC, a measure for linear systems, with an efficient estimator of centered correntropy using the incomplete Cholesky decomposition.

Computationally, KPDC is heavier than the classical version, a naturally expected result. However, in principle, it is significantly lighter than the transfer entropy based methods due to the use of correntropy. An efficient implementation methodology for the proposed method has been outlined in this work. Simulation studies involving complex nonlinear systems showed that KPDC effectively detects connectivity in nonlinear processes in a Granger causal sense. A theoretical basis for the proposed hypothesis testing scheme, however, remains to be proved rigorously. The case studies also highlighted the impact of kernel width on the performance of KPDC and demonstrated that the method has good asymptotic performance.

Directions for future study are along the lines of developing an extension of direct energy transfer to nonlinear systems, obtaining significance levels for KPDC using a false discovery rate-based approach [48], mathematically proving that the permutation significance level converges to the theoretical significance level under the adopted permutation scheme, and extending ideas to processes with heteroskedastic or nonstationary noise sources in addition to the nonlinearity in the functional form.

TABLE II. Performance of KPDC for 10 realizations each for kernel widths $0.25\sigma^*$, $0.5\sigma^*$, σ^*, $2\sigma^*$, and $4\sigma^*$ (in that order, separated by commas) using a sample size of 1000. The (i, j)th cell contains the number of instances in which KPDC predicted that the jth variable had a causal influence on the ith variable.

	x_1	x_2	x_3	x_4	x_5
x_1		0,0,0,0,8	0,0,0,0,0	0,0,0,0,0	0,0,0,0,0
x_2	2,8,10,10,10		0,0,0,0,0	0,0,0,0,0	0,0,0,0,0
x_3	0,0,0,0,0	6,10,10,10,10		0,0,0,0,0	0,0,0,0,0
x_4	0,0,0,0,0	0,0,0,0,0	10,10,10,10,10		0,0,0,0,0
x_5	0,0,0,0,0	0,0,0,0,0	0,0,0,1,0	10,10,10,10,10	

ACKNOWLEDGMENT

The authors would like to specially thank the anonymous referees for their critical comments and suggestions that have resulted in significant improvements in the technical substance of this manuscript.

APPENDIX A: PDC ESTIMATION USING A VAR MODEL

A VAR model [52] is a commonly used representation to quantify linear relationships between variables in a jointly stationary process. Consider an m-dimensional jointly stationary multivariate process, denoted by the vector $\mathbf{x} = [x_1\ x_2\ \cdots\ x_m]^{\mathrm{T}}$. The VAR($p$) model for the process is represented as

$$\mathbf{x}[k] = \sum_{r=1}^{p} \mathbf{A}_r \mathbf{x}[k-r] + \mathbf{e}[k], \tag{A1}$$

where $\mathbf{A}_r$ is the $m \times m$ matrix of autoregressive coefficients at lag r, $\mathbf{e}[k]$ is an m-dimensional vector of white noise sequences, and p is the model order.

It is worthwhile to note two points here: (i) for a scalar (one-dimensional) process, the matrix $\mathbf{A}_r$ reduces to a scalar giving rise to the regular AR representation, and (ii) from a prediction perspective, the quantity $e_j[k]$ due to its uncorrelated nature, represents the unpredictable part or the innovation in $x_j[k]$. The latter point gives rise to the usage of the term innovations for $e_j[k]$. The innovation sequence $\mathbf{e}[k]$ is characterized by its covariance matrix Σ_e. Often for simplicity, Σ_e is assumed to be a diagonal matrix, i.e., the cross correlation between $e_j[k]$ and $e_i[k]$ is assumed to be zero.

An important quantity to be estimated in a VAR model is the model order (p) which is seldom known *a priori*. Selecting a low model order may result in loss of information captured from the process, while choosing a high model order may tune the model coefficients to explain the innovations in the process, resulting in unreliable parameter estimates for the VAR model (overfitting). A widely accepted criterion for selecting the optimum model order is to minimize the Akaike information criterion (AIC), which is given by [30]

$$A_{\mathrm{IC}}(p) = N\ \ln[\det(\Sigma_e)] + 2m^2 p, \tag{A2}$$

where A_{IC} denotes the AIC and N is the number of observations. For the description of nonlinear processes, a high model order is generally chosen to approximate the process using a VAR model. The covariance of innovations, Σ_e, is not generally known and has to be estimated along with the model parameters.

PDC, as detailed in Sec. II A, requires estimates of the inverse transfer function matrix $\bar{\mathbf{A}}(\omega)$, which is efficiently estimated by building a VAR model on the data [Eq. (5)]. An estimate of the transfer function matrix $\mathbf{H}(\omega)$ can obtained from the VAR model by

$$\begin{aligned} \mathbf{H}(\omega) &= \bar{\mathbf{A}}^{-1}(\omega), \\ \bar{\mathbf{A}}(\omega) &= \mathbf{I} - \mathbf{A}(\omega), \\ \mathbf{A}(\omega) &= \sum_{r=1}^{p} A_r z^{-r} \Big|_{z=e^{-j\omega}}, \end{aligned} \tag{A3}$$

where $\mathbf{A}(\omega)$ is the Fourier transform [53] of the VAR coefficients A_r.

A useful interpretation of PDC estimates depends on the reliability of the estimated VAR model. The chosen model order thus plays a key role in the successful estimation of PDC. Underfitting can lead to wrong estimates of the PDC and hence one has to select the model order using a criterion such as the AIC. The AIC is known to fit generally a higher-order model. While it is observed that the PDC estimate is generally invariant above a certain model order, it is noteworthy that too high a model order may lead to oscillations in the estimated PDC and may lead to spurious interpretations.

APPENDIX B: SPECTRAL FACTORIZATION THEOREM

The basis for the quantification of PDC and the direct energy transfer (DET) is the well-known spectral factorization theorem [54], which states that the cross power spectral density of a jointly stationary process $\mathbf{S}(\omega)$ can be factored as

$$\mathbf{S}(\omega) = \mathbf{H}(\omega)\Sigma_e \mathbf{H}^*(\omega), \tag{B1}$$

where Σ_e is the covariance matrix of the innovations (white noise) driving the multivariate process and $\mathbf{H}(\omega)$ is the transfer function matrix in the frequency domain. The superscript $(.)^*$ denotes the Hermitian conjugate of the matrix. The factorization is the key in the separation of direct and indirect effects.

In the general scenario of an input-output multivariate process, the input can be given the representation of a white noise driven VAR process $\mathbf{x}[k] = [x_1[k]\ x_2[k]\ \cdots\ x_m[k]]^{\mathrm{T}}$, i.e., with power spectral density $\mathbf{S}_{xx}(\omega) = \mathbf{H}_x(\omega)\Sigma_e \mathbf{H}_x^*(\omega)$. Subsequently for the output $\mathbf{y}[k]$, $\mathbf{S}_{yy}(\omega) = \mathbf{H}_y(\omega)\mathbf{S}_{xx}\mathbf{H}_y^*(\omega)$. Such a model allows more flexibility in describing a variety of situations (colored or correlated noise) encountered in practice.

APPENDIX C: ANALYSIS OF KPDC FOR LINEAR SYSTEMS

Consider, as an example of two linearly related stationary processes, $\{x[k]\}, \{y[k]\}$ satisfying

$$y[k] = \sum_{d=1}^{p} \alpha_d x[k-d] + v[k],$$

where $\{x[k]\}$ and $\{v[k]\}$ are Gaussian processes with means zero, $\{x[k]\}$ has an autocovariance sequence σ_{xx}, $\{v[k]\}$ is a white noise sequence with variance σ_v^2 and $p > 0$, $|\alpha| < 1$. Assume that $\{x[k]\}$ and $\{v[k]\}$ are uncorrelated.

The centered correntropy estimate between Y and X is obtained as

$$\begin{aligned}\hat{U}_{YX}[l] &= \frac{1}{N-l}\sum_{n=l}^{N-1} k(y[n],x[n-l]) - \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1} k(y[n],x[m]) \\ &= \frac{1}{N-l}\sum_{k=l}^{N-1}\frac{1}{\sqrt{2\pi\sigma}}\sum_{q=0}^{\infty}\frac{(-1)^q}{2^q\sigma^{2q}q!}||y[k]-x[k-l]||^{2q} \\ &\quad - \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\frac{1}{\sqrt{2\pi\sigma}}\sum_{q=0}^{\infty}\frac{(-1)^q}{2^q\sigma^{2q}q!}||y[n]-x[m]||^{2q}. \end{aligned} \tag{C1}$$

Consider the rth term in the expansion: $\sum_{q=0}^{\infty}\frac{(-1)^q}{2^q\sigma^{2q}q!}||y[n]-x[m]||^{2q}$,

$$\begin{aligned}\frac{(-1)^q}{2^q\sigma^{2q}q!}||y[n]-x[m]||^{2q}\Big|_{q=r} &= \frac{(-1)^r}{2^r\sigma^{2r}r!}(y[n]-x[m])^{2r} = \frac{(-1)^r}{2^r\sigma^{2r}r!}\sum_{s=0}^{2r}\binom{2r}{s}(y[n])^s(-x[m])^{2r-s} \\ &= \frac{(-1)^r}{2^r\sigma^{2r}r!}\left[\sum_{s=0}^{2r}\binom{2r}{s}\left(\sum_{d=1}^{p}\alpha_d x[n-d]+v[n]\right)^s(-x[m])^{2r-s}\right] \\ &= \frac{(-1)^r}{2^r\sigma^{2r}r!}\left\{\sum_{s=0}^{2r}\binom{2r}{s}\left[\sum_{t=0}^{s}\binom{s}{t}\left(\sum_{d=1}^{p}\alpha_d x[n-d]\right)^t(v[n])^{s-t}\right](-x[m])^{2r-s}\right\}.\end{aligned}$$

Using the multinomial theorem

$$= \frac{(-1)^r}{2^r\sigma^{2r}r!}\left\{\sum_{s=0}^{2r}(-1)^s\binom{2r}{s}\left[\sum_{t=0}^{s}\binom{s}{t}Q_{t,n,p}(v[n])^{s-t}(x[m])^{2r-s}\right]\right\},$$

where

$$Q_{t,n,p} = \left[\sum_{k_1+\cdots+k_p=t}\binom{n}{k_1,\ldots,k_p}\prod_{1\leqslant f\leqslant p}(\alpha_f\, x[n-f])^{k_f}\right].$$

Because $\{x[k]\}$ and $\{v[k]\}$ are uncorrelated Gaussian processes, they are independent and we have

$$E[(x[k-p_1])^{a_1}(v[k])^{a_2}] = E[(x[k-p_1])^{a_1}]E[(v[k])^{a_2}].$$

The rth term contains the product $(\prod_{1\leqslant f\leqslant p}(x[n-f])^{k_f})(x[m])^{2r-s}(v[n])^{s-t}$, which when summed over m and n can be interpreted as the product of a $(t+2r-s)$th moment of X and a $(s-t)$th moment of V (this separability is valid because X and V are independent Gaussian random variables). To express this more clearly, we interchange the order of summations and first evaluate the summation over m and n for some generic term of the multinomial distribution. Consider the second term (T_{22}) in (C1):

$$\begin{aligned}T_{22} &= \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\frac{1}{\sqrt{2\pi\sigma}}\sum_{r=0}^{\infty}\frac{(-1)^r}{2^r\sigma^{2r}r!}\left\{\sum_{s=0}^{2r}(-1)^s\binom{2r}{s}\left[\sum_{t=0}^{s}\binom{s}{t}Q_{t,n,p}(v[n])^{s-t}(x[m])^{2r-s}\right]\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma}}\sum_{r=0}^{\infty}\frac{(-1)^r}{2^r\sigma^{2r}r!}\sum_{s=0}^{2r}(-1)^s\binom{2r}{s}\left\{\sum_{t=0}^{s}\binom{s}{t}\sum_{\sum_i k_i=t}\binom{n}{k_1,\ldots,k_p}\right. \\ &\quad \left.\times\left[\frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\left(\prod_{1\leqslant f\leqslant p}(x[n-f])^{k_f}\right)(x[m])^{2r-s}(v[n])^{s-t}\right]\right\}.\end{aligned}$$

Evaluating the sums over m and n alone, we have

$$\tau_{m,n} = \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1}\left(\prod_{1\leqslant f\leqslant p}(x[n-f])^{k_f}\right)(x[m])^{2r-s}(v[n])^{s-t}.$$

Using the independence property for the estimates, we obtain

$$= \left[\frac{1}{N^2} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} \left(\prod_{1 \leqslant f \leqslant p} (x[n-f])^{k_f} \right) (x[m])^{2r-s} \right] \left(\frac{1}{N^2} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} (v[n])^{s-t} \right)$$

$$= \left[\frac{1}{N^2} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} \left(\prod_{1 \leqslant f \leqslant p} (x[n-f])^{k_f} \right) (x[m])^{2r-s} \right] \left(\frac{1}{N} \sum_{n=0}^{N-1} (v[n])^{s-t} \right).$$

The latter term in $\tau_{m,n}$ can be easily computed (since $\{v[k]\}$ is a Gaussian white noise process) in terms of σ_v^2 alone.

The former term in $\tau_{m,n}$ is the expected value of a lagged $(t + 2r - s)$th moment of X. Since $\{x[k]\}$ is a Gaussian random process, we can rewrite the former term as the estimate of a higher-order moment of a multivariate normal distribution using Isserlis' theorem [55].

Hence the former term (T_{12}) in $\tau_{m,n}$ can be written as a sum of several autocovariances of $\{x[k]\}$ of orders $(n - f - m)$, where $1 \leqslant n,m \leqslant N$, $1 \leqslant f \leqslant p$,

$$T_{12} = \left[\frac{1}{N^2} \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} \left(\prod_{1 \leqslant f \leqslant p} (x[n-f])^{k_f} \right) (x[m])^{2r-s} \right]$$

$$= \left[\frac{2}{N} \sum_{q=-(N-1)}^{N-1} \left(1 - \frac{|q|}{N} \right) \hat{M}(q; k_f(t), r, s) \right],$$

where $\hat{M}(q; k_f(t), r, s)$ is the estimate of the lagged $(t + 2r - s)$th moment of X (m replaced by $n - q$), i.e.,

$$E\Big[\underbrace{x[n-1] \cdots x[n-1]}_{k_1} \cdots \underbrace{x[n-p] \cdots x[n-p]}_{k_p} \underbrace{x[n-q] \cdots x[n-q]}_{2r-s} \Big].$$

Note that the above defined T_{12} is valid for the second term in (C1), and asymptotically vanishes to zero with N for finite-order autocorrelation sequences for $\{x[k]\}$. For the first term in (C1), we have $m = k - l$ and $n = k$. Hence the first term of $\tau_{m,n}$ reduces to

$$T_{11} = \left[\frac{1}{N-l} \sum_{k=0}^{N-1} \left(\prod_{1 \leqslant f \leqslant p} (x[k-f])^{k_f} \right) (x[k-l])^{2r-s} \right]$$

$$= \hat{M}(l; k_f(t), r, s),$$

where $\hat{M}(l; k_f(t), r, s)$ is the estimate of the following lagged $(t + 2r - s)$th moment of X:

$$E\Big[\underbrace{x[k-1] \cdots x[k-1]}_{k_1} \cdots \underbrace{x[k-p] \cdots x[k-p]}_{k_p} \underbrace{x[k-l] \cdots x[k-l]}_{2r-s} \Big].$$

Thus, from Isserlis' theorem, we see that $\hat{M}(l; k_f(t), r, s)$ can be written solely as a function of autocovariance estimates of $\{x[k]\}$ of orders belonging to the set $\{1, 2, \ldots, p-1, l-1, l-2, \ldots, l-p\}$. These are exactly the various lagged covariances that one would obtain from the cross covariance of $\{y[k]\}$ and $\{x[k]\}$ and autocovariance of $\{y[k]\}$.

Thus the estimate of centered correntropy asymptotically reduces to

$$\hat{U}_{YX}[l] = \frac{1}{\sqrt{2\pi}\sigma} \sum_{r=0}^{\infty} \frac{(-1)^r}{2^r \sigma^{2r} r!} \sum_{s=0}^{2r} (-1)^s \binom{2r}{s} \left[\sum_{t=0}^{s} \binom{s}{t} \sum_{\sum_i k_i = t} \binom{n}{k_1, \ldots, k_p} \hat{M}(l; k_f(t), r, s) \left(\frac{1}{N} \sum_{n=0}^{N-1} (v[n])^{s-t} \right) \right],$$

which can be solely expressed in terms of autocovariances of $\{x[k]\}$ and σ_v^2. Further from Isserlis' theorem, we note that the estimate of correntropy involves products of the estimates of autocovariances at various lags. Hence the estimate of correntropy is not a linear function of the estimate of cross covariance, and thus KPDC estimates aren't equivalent to PDC estimates for linearly related processes (in general) for any σ.

However, we note that for a linear process of the form $y[k] = \alpha x[k-p] + v[k]$, normalized cross-correntropy estimates are a function of the autocorrelation estimate at lag $l - p$ (i.e., $\hat{\rho}[l-p]$) alone, and for sufficiently small $\hat{\rho}[l-p]$, can be approximated as a linear function of $\hat{\rho}[l-p]$, thus reducing to the cross-correlation function. Under such a scenario, KPDC is equivalent to PDC for any kernel width σ.

[1] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, J. Comput. Biol. **7**, 601 (2000).
[2] D. Marinazzo, M. Pellicoro, and S. Stramaglia, Phys. Rev. E **77**, 056215 (2008).
[3] F. Yang, S. L. Shah, and D. Xiao, Int. J. Appl. Math. Comput. Sci. **22**, 41 (2012).
[4] C. W. J. Granger, Econometrica **37**, 424 (1969).
[5] J. Geweke, in *Handbook of Econometrics*, edited by Z. Griliches and M. D. Intriligator (North-Holland, Amsterdam, 1984), Vol. 2, pp. 1101–1144.
[6] M. Kaminski and K. Blinowska, Biol. Cybern. **65**, 203 (1991).
[7] L. Baccala and K. Sameshima, Biol. Cybern. **84**, 463 (2001).
[8] U. Triacca, Theor. Appl. Climatol. **81**, 133 (2005).
[9] M. Eichler, Biol. Cybern. **94**, 469 (2006).
[10] S. Gigi and A. K. Tangirala, Biol. Cybern. **103**, 119 (2010).
[11] B. Schelter, M. Winterhalder, M. Eichler, M. Peifer, B. Hellwig, B. Guschlbauer, C. H. Lucking, R. Dahlhaus, and J. Timmer, J. Neurosci. Methods **152**, 210 (2005).
[12] M. Winterhalder, B. Schelter, W. Hesse, K. Schwab, L. Leistriz, D. Klan, R. Bauer, J. Timmer, and H. Witte, Signal Process. **85**, 2137 (2005).
[13] A. McIntosh and F. Gonzalez-Lima, Hum. Brain Mapp. **2**, 2 (1994).
[14] S. L. Bressler and A. K. Seth, NeuroImage **58**, 323 (2010).
[15] J.-W. Xu, H. Bakardjian, A. Cichocki, and J. C. Principe, in *International Joint Conference on Neural Networks, IJCNN 2007* (IEEE, New York, 2007), pp. 2046–2051.
[16] A. Hegde, D. Erdogmus, Y. N. Rao, J. C. Principe, and J. Gao, in *IEEE 13th Workshop on Neural Networks for Signal Processing, NNSP'03* (IEEE, New York, 2003), pp. 819–828.
[17] M. Jachan, K. Henschel, J. Nawrath, A. Schad, J. Timmer, and B. Schelter, Phys. Rev. E **80**, 011138 (2009).
[18] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya, Phys. Rep. **441**, 1 (2007).
[19] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).
[20] N. Ancona, D. Marinazzo, and S. Stramaglia, Phys. Rev. E **70**, 056221 (2004).
[21] D. Marinazzo, M. Pellicoro, and S. Stramaglia, Phys. Rev. Lett. **100**, 144103 (2008).
[22] D. Marinazzo, W. Liao, H. Chen, and S. Stramaglia, NeuroImage **58**, 330 (2011).
[23] L. Faes, G. Nollo, and K. H. Chon, Ann. Biomed. Eng. **36**, 381 (2008).
[24] P. Duan, F. Yang, T. Chen, and S. L. Shah, in *American Control Conference (ACC), June, 2012* (IEEE, 2012), pp. 3522–3527.
[25] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. (Wiley, New York, 2001).
[26] I. Santamaría, P. Pokharel, and J. C. Príncipe, IEEE Trans. Signal Process. **54**, 2187 (2006).
[27] C. Diks and J. DeGoede, in *Global Analysis of Dynamical Systems*, edited by H. W. Broer, B. Krauskopf, and G. Vegter, (IOP Publishing, Bristol, 2001), pp. 391–403.
[28] M. Paluš, V. Komárek, Z. Hrnčíř, and K. Štěrbová, Phys. Rev. E **63**, 046211 (2001).
[29] I. Park and J. Principe, in *International Conference on Acoustics, Speech and Signal Processing, Nevada, USA* (IEEE, New York, 2008).
[30] M. Priestley, *Spectral Analysis and Time Series* (Academic Press, London, 1981).
[31] Y. Saito and H. Harashima, in *Recent Advances in EEG and MEG Data Processing*, edited by N. Yamaguchi and K. Fujisawa (Elsevier, Amsterdam, 1981), pp. 133–146.
[32] B. W. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, New York, 1986).
[33] J.-W. Xu, Ph.D. dissertation, University of Florida, 2007.
[34] W. Liu, P. P. Pokharel, and J. C. Príncipe, IEEE Trans. Signal Process. **55**, 5286 (2007).
[35] A. Gunduz and J. C. Principe, Signal Process. **89**, 14 (2009).
[36] J. C. Principe, *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives* (Springer, New York, 2010).
[37] R. H. Shumway and D. S. Stoffer, *Time Series Analysis and its Applications* (Springer, New York, 2000).
[38] S. Seth and J. C. Príncipe, in *International Joint Conference Neural Networks, IJCNN 2009* (IEEE, New York, 2009), pp. 2883–2887.
[39] G. Strang, *Linear Algebra and its Applications* (Thomson, Belmont, CA, 2006).
[40] E. Lehmann, *Elements of Large-Sample Theory* (Springer, New York, 1999).
[41] R. S. Tsay and G. C. Tiao, J. Am. Stat. Assoc. **79**, 84 (1984).
[42] E. J. G. Pitman, J. R. Stat. Soc. Suppl. **4**, 119 (1937).
[43] D. S. Moore and G. P. McCabe, *Introduction to the Practice of Statistics* (Freeman, San Francisco, 2005), Chaps. 14–17 .
[44] E. E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses* (Springer Science+Business Media, New York, 2005).
[45] S. Holm, Scand. J. Stat. **1979**, 65 (1979).
[46] Y. Hochberg, Biometrika **75**, 800 (1988).
[47] S. Nakagawa, Behav. Ecol. **15**, 1044 (2004).
[48] Y. Benjamini and Y. Hochberg, J. R. Stat. Soc.: Ser. B (Methodol.) **1995**, 289 (1995).
[49] Y. Lin and M. A. Stadtherr, Ind. Eng. Chem. Res. **46**, 7198 (2007).
[50] J. A. Enszer, Y. Lin, S. Ferson, G. F. Corliss, and M. A. Stadtherr, in *Proceedings of the 3rd International Workshop on Reliable Engineering, Computing* (Georgia Institute of Technology, Savannah, GA, 2008), pp. 89–105.
[51] L. Faes, G. Nollo, and A. Porta, Phys. Rev. E **83**, 051112 (2011).
[52] H. Lutkepohl, *New Introduction to Multiple Time Series Analysis* (Springer, New York, 2005).
[53] S. W. Smith, *Scientist and Engineer's Guide to Digital Signal Processing* (California Technical Publishing, San Diego, CA, 1997).
[54] M. Gevers and B. Anderson, Int. J. Control **33**, 777 (1981).
[55] L. Isserlis, Biometrika **12**, 134 (1918).