LETTER **Communicated by Liangjun Su**

Conditional Association

Sohan Seth
sohan@cnel.ufl.edu
José C. Príncipe
principe@cnel.ufl.edu
Department of Electrical and Computer Engineering, University of Florida,
Gainesville, FL 32608, U.S.A.

Estimating conditional dependence between two random variables given the knowledge of a third random variable is essential in neuroscientific applications to understand the causal architecture of a distributed network. However, existing methods of assessing conditional dependence, such as the conditional mutual information, are computationally expensive, involve free parameters, and are difficult to understand in the context of realizations. In this letter, we discuss a novel approach to this problem and develop a computationally simple and parameter-free estimator. The difference between the proposed approach and the existing ones is that the former expresses conditional dependence in terms of a finite set of realizations, whereas the latter use random variables, which are not available in practice. We call this approach conditional association, since it is based on a generalization of the concept of association to arbitrary metric spaces. We also discuss a novel and computationally efficient approach of generating surrogate data for evaluating the significance of the acquired association value.

1 Introduction

The problem of assessing conditional dependence between two random variables given the knowledge of a third random variable is essential in many scientific problems, for example, in evaluating effective connectivity of a neuronal network (Quinn, Coleman, Kiyavash, & Hatsopoulos, 2011) or assessing causal influence in a biological network (Lozano, Naoki, Liu, & Rosset, 2009). However, the available methods for assessing conditional dependence (Schreiber, 2000; Fukumizu, Gretton, Sun, & Schölkopf, 2008; Su & White, 2008) suffer from several computational drawbacks. First, they are computationally expensive, requiring $\mathcal{O}(n^2)$ to $\mathcal{O}(n^3)$ time complexity where n is the sample size; second, they almost always require selecting several free parameters, such as a kernel, the corresponding kernel size, and, often, a regularization parameter, where selecting the best values of these parameters remains an open problem; third, these measures often assess only conditional independence rather than conditional dependence, that is, they are zero if and only if conditional independence is satisfied, but they do not address when and how conditional dependence increases or decreases or becomes maximum; and fourth, they explore conditional dependence in the context of the random variables, but it remains unclear that given a finite set of realizations, what property of this set of realizations makes them conditionally dependent. These issues prohibit the applicability of these measures to practical problems, where sample size or dimensionality, or both, could be high; where just knowing that the random variables are conditionally independent is not enough; where one has access to only a set of realizations rather than the random variables; and where simpler approaches such as linear Granger causality and partial correlation (PC) remain dominant due to their inherent simplicity (Dhamala, Rangarajan, & Ding, 2008).

Two random variables X and Y are said to be conditionally independent given a third random variable Z (i.e., $X \perp Y|Z$) if and only if

$$\mathrm{P}(X \in A|Y \in B, Z = C) = \mathrm{P}(X \in A|Z = C), \tag{1.1}$$

where A, B, and C are arbitrary subsets of the respective domains where the random variables assume values. In Euclidean space, which is perhaps the most frequently encountered domain in practice, equation 1.1 can be equivalently expressed in several other ways, for example, using cumulative distribution function (CDF), probability density function (PDF), or characteristic function (CHF). A measure of conditional independence is usually derived from these different expressions of conditional independence, for example, using conditional CDF (Linton & Gozalo, 1996), conditional PDF (Diks & Panchenko, 2006), conditional CHF (Su & White, 2007), kernel methods (Fukumizu et al., 2008), copula (Bouezmarni, Rombouts, & Taamouti, 2009), and conditional probability distribution function (Seth & Príncipe, 2012). Such measures have received considerable attention in recent years due to their growing applicability in many practical machine learning problems, such as causal inference (Sun, Janzing, Schölkopf, & Fukumizu, 2007), dimensionality reduction (Fukumizu, Bach, & Jordan, 2004), feature selection (Bontempi & Meyer, 2010), and time series analysis (Su & White, 2008). However, the problem of assessing conditional dependence remains largely unexplored in the contemporary literature. We differentiate between these two often synonymously used ideas in the sense that the latter should provide insight into how two random variables become conditionally dependent given the knowledge of a third random variable, and how this dependence increases or decreases, while the former only needs to ensure that it is zero if and only if conditional independence is satisfied. We argue further that it is perhaps more important to understand how conditional dependence changes in the context of a set of realizations rather a random variable triplet since one does not have access to the latter. We argue that this understanding is missing in the current literature, and it has motivated us to explore alternate methods. At this point, it is perhaps worth emphasizing that the objective of this letter is not to design a test of conditional independence but to explore and quantify conditional dependence from an application perspective.

Assessing conditional dependence is particularly essential in estimating causal flow in the sense of Granger (Quinn et al., 2011). It is well known that assessing Granger noncausality, that is, one time series $\{X_t\}$ not having any causal influence on another time series $\{Y_t\}$, is equivalent to finding whether the present value of $\{Y_t\}$ is conditionally independent of the past values of $\{X_t\}$ given the past values of $\{Y_t\}$ alone (Diks & Panchenko, 2006). However, this approach does not address the issue that if $\{X_t\}$ indeed causes $\{Y_t\}$, then how does one quantify the strength of causation? This issue is usually addressed in terms of the conditional mutual information (CMI), an established measure of conditional dependence (Joe, 1989; Schreiber, 2000). But conditional mutual information is very difficult to estimate, and the properties satisfied by the measure itself are not inherited by its finite sample estimator. To elaborate, consider the random variables (X, Y, Z) with joint probability law $\mathrm{P}(X, Y, Z)$. It is clear that CMI is minimum when equation 1.1 is satisfied, whereas it is maximum when there exists a functional relationship between X and Y for every $Z = z$ (Joe, 1989). Now consider a finite set of realizations $\{(x_i, y_i, z_i)\}_{i=1}^n$ from this probability law. Given these samples, CMI is estimated consistently by estimating the Radon-Nikodym derivative using adaptive binning (Pérez-Cruz, 2008) or kernel smoothing (Joe, 1989). But the resulting estimators neither provide the same understanding of when the estimated value is minimum or maximum, and under what circumstances this value increases or decreases, nor do inherit the desired properties of the measure (e.g., they can be negative and are never invariant to one-to-one transformation). Therefore, there is a clear mismatch between what a measure intends to quantify and what an estimator is able to capture; or from a slightly different aspect, although the meaning of the measure is transparent in the context of the random variables (i.e., the probability law), it remains unclear what attribute makes them conditionally dependent from the perspective of the realizations.

In this letter, we address the following: Given a set of realizations $\{(x_i, y_i, z_i)\}_{i=1}^n$ from a random variable triplet (X, Y, Z), what makes the realizations $\{x_i\}_{i=1}^n$ conditionally dependent on the realizations $\{y_i\}_{i=1}^n$ given the realizations $\{z_i\}_{i=1}^n$? Earlier work in statistics by pioneers such as Galton, Pearson, Spearman, and Kendall shows a similar approach to estimating statistical dependence, where rather than starting with a statistical measure and deriving an appropriate estimator, the authors start with the realizations and provide an intuitive explanation of what dependence should imply (Spearman, 1904; Kendall, 1938). While the existing approaches of assessing conditional dependence follow the former conceptual framework, we follow the latter since, in practice, we have access to only a finite set of realizations rather than the underlying random variables and also since it is rather difficult to materialize the desired properties and understanding of a measure in an estimator. In order to achieve this, we generalize the concept of association to arbitrary metric spaces and then extend it to introduce the concept of conditional association.[1] The proposed approach not only provides an intuitive view of what conditional dependence is and how it changes, but it culminates in a parameter-free and relatively simpler estimator, thus becoming an excellent alternative to the state-of-the-art methods. Another advantage of this approach is that it is defined only in terms of the pairwise distances between the realizations and therefore is applicable to exotic metric spaces such as non-Euclidean space or infinite dimensional space, such as the space of spike trains (Seth et al., 2010).

Although conditional association provides a clear interpretation of what conditional dependence implies, the significance of the acquired value remains to be investigated, especially under small sample size. Since explicitly computing the asymptotic null distribution in this context is computationally intense and the theoretical asymptotic null distribution is often violated for finite samples, we consider generating surrogate data to estimate the null distribution.[2] Unfortunately, generating surrogate data to simulate conditional independence remains an open area of research (Seth & Príncipe, 2012). In this letter, therefore, we also introduce a novel scheme for generating surrogate data for simulating conditional independence. Some interesting properties of the proposed approach are that it involves only one free parameter and it resamples the original data, thus providing a scope for reusing computations involved in estimating the original conditional dependence value, for estimating the surrogate values. However, in its present format, this approach is applicable to Euclidean spaces only where the Lebesgue measure is defined. Therefore, we limit ourselves here to Euclidean spaces.

The rest of the letter is organized as follows. In section 2, we provide a brief overview of the existing literature on measures of conditional dependence and point out their weaknesses. In section 3, we start with a brief overview of the concept of association and then discuss how this concept can be generalized and extended to address the notion of conditional association. In section 4, we propose a novel scheme of generating surrogate data for evaluating the significance of conditional association. In section 5,

[1] Notice that the work presented in this letter is not the same as in Holland and Rosenbaum (1986) where the authors have explored a different concept under the same name.

[2] Our objective is not to design a test for conditional independence, but merely to observe if the acquired conditional association value is significant enough to be considered a sign of conditional dependence. Testing conditional independence requires a measure of conditional independence (Seth & Príncipe, 2012), and we have yet to find a formal derivation that the measure of conditional association satisfies the necessary properties.

we apply the proposed method on synthetic and real data to provide more insight into the proposed method. Finally, in section 6, we conclude with a brief summary of the proposed work and some guidelines for future work.

2 Background

In this section, we briefly go over the available measures of conditional (in)dependence, and discuss their strengths and weaknesses. In terms of CDF, if (X, Y) are conditionally independent given Z, then $F_{XY|Z}(x, y|z) = F_{X|Z}(x|z)\ \forall\,(x, y, z)$, where $F_{U|V}(u, v) = \mathrm{P}(U \leq u|V \leq v)$ is the conditional CDF of the random variable U given the event $V \leq v$.[3] Exploiting the relation $F_{U|V}(u|v) = F_{U|V}(u, v)/F_V(v)$, Linton and Gozalo (1996) proposed the following measure of conditional independence:[4]

$$\mathrm{CM}(X, Y; Z) = \int A^2(x, y, z)\mathrm{d}F_{XYZ}(x, y, z),$$

where

$$A(x, y, z) = F_{XYZ}(x, y, z)F_Z(z) - F_{XZ}(x, z)F_{YZ}(y, z).$$

This measure can be consistently estimated by replacing the CDFs by their empirical estimates, $F_U^n(u) = 1/n\sum_i \mathbb{I}(u \leq u_i)$, where $\{u_i\}_{i=1}^n$ are samples from U and $\mathbb{I}$ is the identity function. The advantage of this approach is that it is parameter free. However, it is vulnerable to small sample size and high dimensionality since in both of these situations, the empirical estimate breaks down and becomes zero.

In practice, therefore, a PDF-based approach is often preferred. In terms of PDF, (X, Y) are conditionally independent given Z if and only if $f_{XY|Z}(x, y|z) = f_{X|Z}(x|z)\ \forall\,(x, y, z)$, where $f_{U|V}(u|v)$ is the conditional PDF of the random variable U given the random variable V. Once again, exploiting the relation $f_{U|V}(u|v) = f_{UV}(u, v)/f_V(v)$, Su and White (2008) proposed the following measure of conditional dependence:[5]

$$\mathrm{HD}(X, Y; Z) = \int \left\{1 - \sqrt{\frac{f_{XZ}(x, z)f_{YZ}(y, z)}{f_{XYZ}(x, y, z)f_Z(z)}}\right\}^2 a(x, y, z)\mathrm{d}F_{XYZ}(x, y, z),$$

[3] The usual definition of conditional CDF is $F_{U|V}(u, v) = \mathrm{P}(U \leq u|V = v)$.

[4] It is not a strict measure of conditional independence since it is only a necessary condition, but not an *if and only if* condition.

[5] Su and White (2008) have explored this measure only in the context of a test of conditional dependence. However, (unweighted) Hellinger distance is a measure of conditional dependence (Joe, 1989).

where $a(x, y, z)$ is an appropriate nonnegative weighting function. This measure can again be consistently estimated by replacing the actual probabilities by their kernel density estimates, $f_U^n(u) = 1/n \sum_{i=1}^n p(u - u_i)$, where $p(u)$ is an appropriate kernel (Su & White, 2008). The advantage of this approach is that it is less susceptible to small sample size and high dimensionality than CM due to the inherent smoothing effect provided by using the kernel. However, it requires choosing an appropriate kernel and associated kernel size, and the performance of the measure heavily relies on the proper selection of these parameters. The best values of these parameters are often chosen by cross-validation, which increases the computational cost of this method. The kernel density estimate, however, still suffers from high dimensionality and small sample size.

The PDF can also be estimated using a *k*th nearest neighbor–based approach. This approach has been adapted by Pérez-Cruz (2008) to consistently estimate CMI,

$$\mathrm{CMI}(X, Y; Z) = \int \log \frac{f_{XYZ}(x, y, z) f_Z(z)}{f_{XZ}(x, z) f_{YZ}(y, z)} \mathrm{d}F_{XYZ}(x, y, z), \tag{2.1}$$

by replacing the PDFs by their empirical estimates,

$$f_U^n(u_i) = k/(n-1)/\mathrm{vol}(\epsilon_k),$$

where $\mathrm{vol}(\epsilon_k)$ denotes the volume of a sphere with radius ϵ_k and ϵ_k is the distance between u_i and its *k*th neighbor. The advantage of this approach is that it does not require selecting a kernel. However, it still has a free parameter, the size of the neighborhood k (Pérez-Cruz, 2008), and to the best of our knowledge, an appropriate criterion for selecting this parameter remains to be found. Notice that both the PDF- and the CDF-based approaches inherently assume that the underlying space where the random variables take values is a finite-dimensional Euclidean space. The CDF-based approach assumes that by considering that the underlying space has an order, whereas the PDF-based approaches assume that either by considering the existence of a density kernel or defining the volume of a sphere with radius ϵ by

$$\mathrm{vol}(\epsilon) = \frac{\pi^{d/2} \epsilon^d}{\Gamma(d/2 + 1)}, \tag{2.2}$$

where d is the dimensionality of the space and Γ is the gamma function. Therefore, these methods cannot be trivially extended to arbitrary metric spaces.

Recently the problem of assessing conditional dependence has also been addressed in the context of kernel-based learning, where Fukumizu et al. (2008) proposed $\mathrm{HSNCIC}(X, Y; Z) = ||\mathcal{V}_{(YZ)(XZ)|Z}||_{HS}^2$ as a measure of conditional dependence, where $\mathcal{V}_{UV|W}$ denotes the normalized conditional cross-covariance operator and $||\cdot||_{HS}$ denotes the Hilbert-Schmidt norm. This measure can be consistently estimated as $\text{HSNCIC}_n = \text{Tr}[\mathbf{R}_{(XZ)}\mathbf{R}_{(YZ)} - 2\mathbf{R}_{(XZ)}\mathbf{R}_{(YZ)}\mathbf{R}_Z + \mathbf{R}_{(XZ)}\mathbf{R}_Z\mathbf{R}_{(YZ)}\mathbf{R}_Z]$, where $\mathbf{R}_U = \mathbf{K}_{UU}(\mathbf{K}_{UU} + n\lambda_n\mathbf{I})^{-1}$, $\mathbf{K}_{UU}(i, j) = \kappa_{\mathcal{U}}(u_i, u_j)$ is a Gram matrix, $\kappa_{\mathcal{U}}(x, y)$ is a strictly positive-definite (spd) kernel, $\mathbf{I}$ is the identity matrix, Tr denotes the trace of a matrix, and λ_n is a regularization parameter that depends on n. The advantage of this approach is that unlike CM, HD, or CMI, the use of an spd kernel allows this measure to be defined on any arbitrary set of random variables that might not take values in $\mathbb{R}^d$. However, there are two major drawbacks to this approach: it requires selecting one more free parameter (i.e., the regularization), and it is computationally more involved than CM and HD, taking $\mathcal{O}(n^3)$ time compared to $\mathcal{O}(n^2)$ by the other methods. The free parameters can be chosen by matching the bootstrap variance of the estimator with the theoretical variance (Fukumizu et al., 2008). However, this approach is computationally involved, especially in conjunction with a permutation test and the high computational complexity of the estimator. Moreover, defining an appropriate strictly positive-definite kernel over an abstract space still remains an active area of research, thus limiting the utility of this approach.

Although these existing approaches (not including CM) are mathematically elegant as measures of conditional (in)dependence: the measures are zero if and only if conditional independence is satisfied among a random variable triplet; they are maximum when one variable is a function of the other given any value of the third variable; and their respective estimators are consistent, that is, they reach their theoretical values when the number of realizations tends to infinity. They do not address what conditional dependence implies in terms of a finite set of realizations—how the estimated value increases or decreases and when it reaches its maximum. In section 3, we address this issue, which leads to an alternate understanding of conditional dependence in the context of a finite set of realizations and provides an opportunity to design simpler estimators that are parameter free, can be applied to any metric space, not just Euclidean, and can effectively quantify conditional dependence, as we will demonstrate with both real and synthetic data.

3 Method

Before proceeding, we briefly discuss the notion of association in statistics. Given realizations $\{(x_i, y_i)\}_{i=1}^n$ from two real valued random variables (X, Y), they are said to be associated if large realizations of X are associated with large realizations of Y. The most widely used measure of association is the correlation, which is defined as $\sum x_i y_i$. However, this measure captures only linear relationships between two random variables since the realizations of X and Y are compared in absolute terms. This idea has been generalized by Spearman (1904), who proposed to use the correlation between the ranks of the realizations as a measure of association. Working with ranks, that is, the relative values rather than the absolute values, allows Spearman's coefficient to capture a monotonic relationship rather than just a linear relationship. Kendall (1938), on the other hand, proposed a measure of association using the ideas of concordance and discordance. Two pairs of realizations, (x_i, y_i) and (x_j, y_j), are said to be concordant if $(x_i - x_j)$ and $(y_i - y_j)$ have the same sign; otherwise, they are said to be discordant. Kendall defined a measure of association as the difference between the number of concordant and discordant pairs normalized by the total number of pairs. It is evident that this idea captures the same attribute of whether relatively large realizations of X are associated with relatively large realizations of Y.

3.1 Generalized Association. The idea of association is defined only on $\mathbb{R}$, where the product (xy) and the ordering $(x < y)$ are well defined, whereas in practice, one often encounters more exotic random variables (e.g., vectors). Therefore, we generalize this idea to a metric space by defining association in the following way: given two random variables (X, Y) in $\mathcal{X} \times \mathcal{Y}$, Y is associated with X if close realization pairs of Y, that is, $\{y_i, y_j\}$, are associated with close realization pairs of X, that is, $\{x_i, x_j\}$, where closeness is defined in terms of the respective metrics of the two spaces where the random variables lie. In other words, if two realizations $\{x_i, x_j\}$ are close in $\mathcal{X}$, then the corresponding realizations $\{y_i, y_j\}$ are close in $\mathcal{Y}$. Notice that by construction, this concept is valid in any metric space, not just Euclidean. However, in this letter, we explore this concept only in the Euclidean space. We call this approach the generalized association.

To quantify this notion of association, we follow this algorithm under the assumption that two realizations do not share the same distance from a third realization. Given realizations $\{(x_i, y_i)\}_{i=1}^{n}$,

1. For all $i \in \{1, \ldots, n\}$, repeat the following.
2. Find x_{j^*} closest to x_i in terms of $\mathfrak{d}_{\mathcal{X}}$, that is, $j^* = \arg\min_{j \neq i} \mathfrak{d}_{\mathcal{X}}(x_i, x_j)$.
3. Find rank r_i of y_{j^*} in terms of $\mathfrak{d}_{\mathcal{Y}}$ that is, $r_i = \#\{j : j \neq i, \mathfrak{d}_{\mathcal{Y}}(y_j, y_i) \leq \mathfrak{d}_{\mathcal{Y}}(y_{j^*}, y_i)$,

where $\mathfrak{d}_{\mathcal{Z}}$ denotes the metric in the space $\mathcal{Z}$. According to the notion of generalized association, the more Y is associated with X, the more the r_i's are skewed toward 1, and to capture that, we simply define the generalized measure of association (GMA) as the normalized area under the empirical distribution of r_i's:

$$\text{GMA} = \frac{1}{n-1} \sum_{r=1}^{n-1} (n-r) \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(r = r_i) \right].$$

It can be shown that when the random variables X and Y are independent, then $\mathbf{E}[\mathrm{GMA}] = 0.5n/(n-1)$. This is because when the random variables are independent, r_i can take any value in the set $\{1, 2, \ldots, n-1\}$ with equal probability. And when the random variables are maximally dependent, $X = Y$, then all the r_i's are 1, and therefore GMA $= 1$. In an intermediate situation, the r_i's should be skewed to a lower value, closer to 1, and therefore GMA considers values between $0.5n/(n-1)$ and 1. So once again, when the random variables are independent, then GMA is close to 0.5, whereas when the random variables are well associated, then GMA is close to 1.

Notice that the concept of generalized association is asymmetric: $\mathrm{GMA}(X, Y) \neq \mathrm{GMA}(Y, X)$. The intuition behind an asymmetric measure of dependence fits the regression scenario very well where the regression function $Y = f(X)$ could be noninvertible: although Y can be completely determined by X, it is not true the other way around. Also, the value of GMA depends on the choice of the metrics $\mathfrak{d}_{\mathcal{X}}$ and $\mathfrak{d}_{\mathcal{Y}}$. This is an undesirable but not surprising fact since any estimator of dependence such as an estimator of mutual information (MI) usually requires choosing suitable metrics (Kraskov, Stögbauer, & Grassberger, 2004; Pérez-Cruz, 2008). However, this selection is often overlooked in the context of a Euclidean space, and the default Euclidean norm, the l_2-norm, is used in the estimators.

3.2 Conditional Association. The notion of generalized association can be exploited to capture conditional association as follows. Given three random variables (X, Y, Z) in $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$, Y is conditionally associated with X given Z, if close realization pairs of Y, that is, $\{y_i, y_j\}$, are more associated with close realization pairs of (X, Z), that is, $\{(x_i, z_i), (x_j, z_j)\}$ than with close realization pairs of Z, that is, $\{z_i, z_j\}$, alone. In other words, given that two realizations $\{z_i, z_j\}$ are close in $\mathcal{Z}$, knowing that the realizations $\{x_i, x_j\}$ are also close in $\mathcal{X}$ brings the realizations $\{y_i, y_j\}$ closer in $\mathcal{Y}$. Notice that like generalized association, this concept is also defined over any metric space where the closeness of two realizations can be assessed by an appropriate metric. Following the definition, we capture conditional association simply by the difference between the generalized associations as follows,

$$\mathrm{MCA}(X, Y; Z) = \mathrm{GMA}((X, Z), Y) - \mathrm{GMA}(Z, Y),$$

where we call MCA the *measure of conditional association*. Then $\mathrm{MCA}(X, Y; Z)$ is greater than zero when Y is conditionally associated with X given Z, whereas $\mathrm{MCA}(X, Y; Z)$ is less than or equal to zero if Y is not conditionally associated with X given Z.

This approach, however, requires designing the metric $\mathfrak{d}_{\mathcal{X}\times\mathcal{Z}}$ by utilizing the metrics $\mathfrak{d}_{\mathcal{X}}$ and $\mathfrak{d}_{\mathcal{Z}}$, where it is not clear, how to make these metrics compatible. For example, they can be combined as $\mathfrak{d}^2_{\mathcal{X}\times\mathcal{Z}} = \mathfrak{d}^2_{\mathcal{X}} + \mathfrak{d}^2_{\mathcal{Z}}$ as in the Euclidean space. But a mere scaling of one of these two metrics can suppress the contribution of the other in the combined metric. This is undesirable, and it restricts the final estimator from being invariant to simple scaling of the individual domains. This, however, is not an issue of conditional association in particular, but any estimator of conditional dependence such as CMI and HSNCIC. Therefore, to evaluate $\mathrm{GMA}((X, Z), Y)$, we consider a slightly different approach. Instead of finding the closest point j^* in terms of metric $\mathfrak{d}_{\mathcal{X}\times\mathcal{Z}}$, we do it in terms of the relative positions of the realizations in the individual domains. To elaborate, we first compute the ranks $\{r_{x_j}\}_{j=1,j\neq i}^{n}$ of realizations $\{x_j\}_{j=1,j\neq i}^{n}$ from x_i using $\mathfrak{d}_{\mathcal{X}}$, and the ranks $\{r_{z_j}\}_{j=1,j\neq i}^{n}$ of realizations $\{z_j\}_{j=1,j\neq i}^{n}$ from z_i using $\mathfrak{d}_{\mathcal{Z}}$, and then find the closest point (x_{j^*}, z_{j^*}) as $j^* = \arg\min_{j\neq i}(r_{x_j} + r_{z_j})$. The intuition behind this approach originates from the understanding that a point in the joint domain would be close if it is close in both individual domains. However, if there are ties in the combined ranks, we resolve them by the ranks of X since we are interested in knowing how important this variable is in the presence of Z.

It is evident that this approach does not suffer from the compatibility issue, and it is invariant to a class of transformations of the individual domains such as scaling on the real line. Also, it preserves our intuition of conditional dependence. To elaborate, let us consider two simple examples. First, consider $X_1 = V_1, Y_1 = U_1, Z_1 = U_1$; where U_1 and V_1 are independent. Then $\mathrm{GMA}(Z_1, Y_1) = 1$, whereas $\mathrm{GMA}((X_1, Z_1), Y_1) < \mathrm{GMA}(Z_1, Y_1)$ since the arbitrary ranks of X_1 affect the combined ranking and disrupt the perfect alignment (closest point closest) between Z_1 and Y_1. Therefore, $\mathrm{MCA}(X_1, Y_1; Z_1) \leq 0$. Notice that as a special case, $\mathrm{MCA}(X, X; X) = 0$ (Dawid, 1998). Next, consider $X_2 = U_2, Y_2 = U_2, Z_2 = V_2$, where U_2 and V_2 are independent random variables. Then $\mathrm{GMA}(Z_2, Y_2) \approx 0.5$, whereas $1 \approx \mathrm{GMA}((X_2, Z_2), Y_2) > \mathrm{GMA}(Z_2, Y_2)$ since the perfect alignment (closest point closest) of X_2 helps bring the combined ranks closer to aligning with Y_2, and therefore, $\mathrm{MCA}(X_2, Y_2; Z_2) > 0$.

Notice that the expression of conditional association is very similar to the expression of CMI, since it can be expressed as

$$\mathrm{CMI}(X, Y; Z) = \mathrm{MI}(X, (Y, Z)) - \mathrm{MI}(X, Z),$$

where mutual information (MI) is defined as

$$\mathrm{MI}(X, Y) = \int f_{XY}(x, y) \log(f_{XY}(x, y)/f_X(x) f_Y(y))\mathrm{d}x\mathrm{d}y.$$

The similarity between these two expressions provides more intuition on how conditional association works. However, unlike CMI, we do not claim that conditional association is a necessary and sufficient condition for conditional independence. A formal proof or validity of this statement remains to be found. Also, the computational complexity of MCA is similar to that of CMI as described in section 2, since both require computing the pairwise distances between the realizations (a $\mathcal{O}(n^2)$ step) and then sorting them to get their relative positions (a $\mathcal{O}(n \log n)$ step on an average). Although we have assumed that the realizations $\{x_i\}_{i=1}^n$, $\{y_i\}_{i=1}^n$, and $\{z_i\}_{i=1}^n$ are distinct, this restriction can be easily lifted by employing a method for breaking ties between ranks.

4 Surrogate Data

Although conditional association, or any other measure of conditional dependence, returns a value, say, q, the significance of this value remains obscure, since a large value can result due to the presence of conditional dependence or simply due to a lack of evidence, that is, a sufficient number of realizations. Therefore, to remove the effect of small sample size, it is essential to judge the significance of this value in the context where conditional dependence is absent. This can be achieved by generating surrogate data sets simulating conditional independence and observing the values of the measure on these surrogate data sets. Let q_n be the original estimated value of the realizations $\{(x_i, y_i, z_i)\}_{i=1}^n$ and $\{q_n^{(t)}\}_{t=1}^T$ be the surrogate values estimated from the surrogate data sets $\{(x_i^{(t)}, y_i^{(t)}, z_i^{(t)})\}_{i=1}^n$ for $t = 1, \ldots, T$. Then we can consider q_n to be significant if $\operatorname{card}\{t : q_n^{(t)} > q_n\} < \alpha T$ where α is sufficiently close to zero.

Generating surrogate data simulating conditional independence, however, is not trivial. Two popular approaches for generating surrogate data have been proposed by Paparoditis and Politis (2000) and Diks and DeGoede (2001). Given realizations $\{x_i, y_i, z_i\}_{i=1}^n$ from (X, Y, Z), the first approach generates realizations $\{x_i', y_i', z_i'\}_{i=1}^n$, representing (X', Y', Z'), such that $X' \perp Y'|Z'$ and $(X, Z) \sim (X', Z')$, $(Y, Z) \sim (Y', Z')$.[6] This is done by sampling from the conditional distributions $\hat{f}_{X|Z}(x|z)$ and $\hat{f}_{Y|Z}(y|z)$, respectively. However, these distributions are estimated using Parzen's approach and require selecting an appropriate resampling width, which becomes difficult in higher dimensions. On the other hand, the second approach generates realizations $\{x_i', y_i', z_i'\}_{i=1}^n$ such that $(X', Z') \perp Y'$. This is done by simply permuting the realizations of Y with respect to the realizations of (X, Z). Although this approach is simple and does not involve any free parameter, $(X', Z') \perp Y'$ is only a sufficient condition for $X' \perp Y'|Z'$ but not necessary.

Here, we discuss a different approach for generating surrogate data by modifying the first approach. We follow the suggestion in Paparoditis and Politis (2000) in the sense that we first generate samples from $\hat{f}_Z(z)$ and then from $\hat{f}_{X|Z}(x|z)$ and $\hat{f}_{Y|Z}(y|z)$, respectively. However, we perform the following modifications in order to make this approach computationally

[6] The expression $X \sim Y$ implies that the random variables X and Y follow the same distribution.

more attractive in the context of the permutation test. First, we assume that the estimated densities $\hat{f}$ exist only on the sample locations; second, we use a nearest neighbor–based approach to estimate the conditional density functions at these locations as described in section 2 in the context of estimating CMI; and third, we reuse the realizations $\{z_i\}_{i=1}^n$, $\{x_i\}_{i=1}^n$, and $\{y_i\}_{i=1}^n$ as realizations from $f_Z(z)$, $f_{X|Z}(x|z)$, and $f_{Y|Z}(y|z)$, respectively. Reusing the original data is computationally advantageous in the context of a permutation test since the computation involved in computing the true conditional dependence value can be reused to compute the surrogate values, which includes reusing the distance matrix or the kernel Gram matrix. Before discussing the other aspects, we present the algorithm in detail.

Consider that we have realizations $\{(u_i, v_i)\}_{i=1}^n$ from a joint distribution $f_{UV}(u, v)$. Then, using the definition of conditional density function, we get $f_{U|V}(u|v = v_j) \propto f_{UV}(u, v_j)$, since $f_V(v_j)$ is a constant normalizing factor. Therefore, we can estimate $f_{U|V}(u_i|v_j) \propto f_{UV}(u_i, v_j)$ following equation 2.2, where we need to specify a neighborhood parameter k. Given the estimate $\hat{f}_{U|V}(u_i|v_j)$, we can then sample from this density function (after normalizing) assuming the density function exists over only the realization values $\{u_i\}_{i=1}^n$. Based on this approach, we follow three simple steps to generate surrogate data: for each i, (1), assign $z'_i = z_i$, (2) sample x'_i from the ensemble $\{x_j\}_{i=1}^n$ with probability $\hat{f}_{X|Z}(x_j|z_i)$, and (3) sample y'_i from the ensemble $\{y_j\}_{i=1}^n$ with probability $\hat{f}_{Y|Z}(y_j|z_i)$.

Although this process is simple, it involves a free parameter k, which, in some sense controls the smoothness of the estimated density function. We set $k = \lceil \sqrt{n} \rceil$, a popular choice in nearest neighbor–based density estimation (Pérez-Cruz, 2008). It should be noted that the choice of this parameter may not be optimal. However, we empirically show that it works well in practice. Also, notice that in some sense, this approach can be understood as the first approach with a variable kernel size and the computational complexity of this approach is higher than the second approach.

4.1 Sanity Check. To demonstrate the validity of the proposed measure and the surrogate data generation technique, we consider the following two examples.

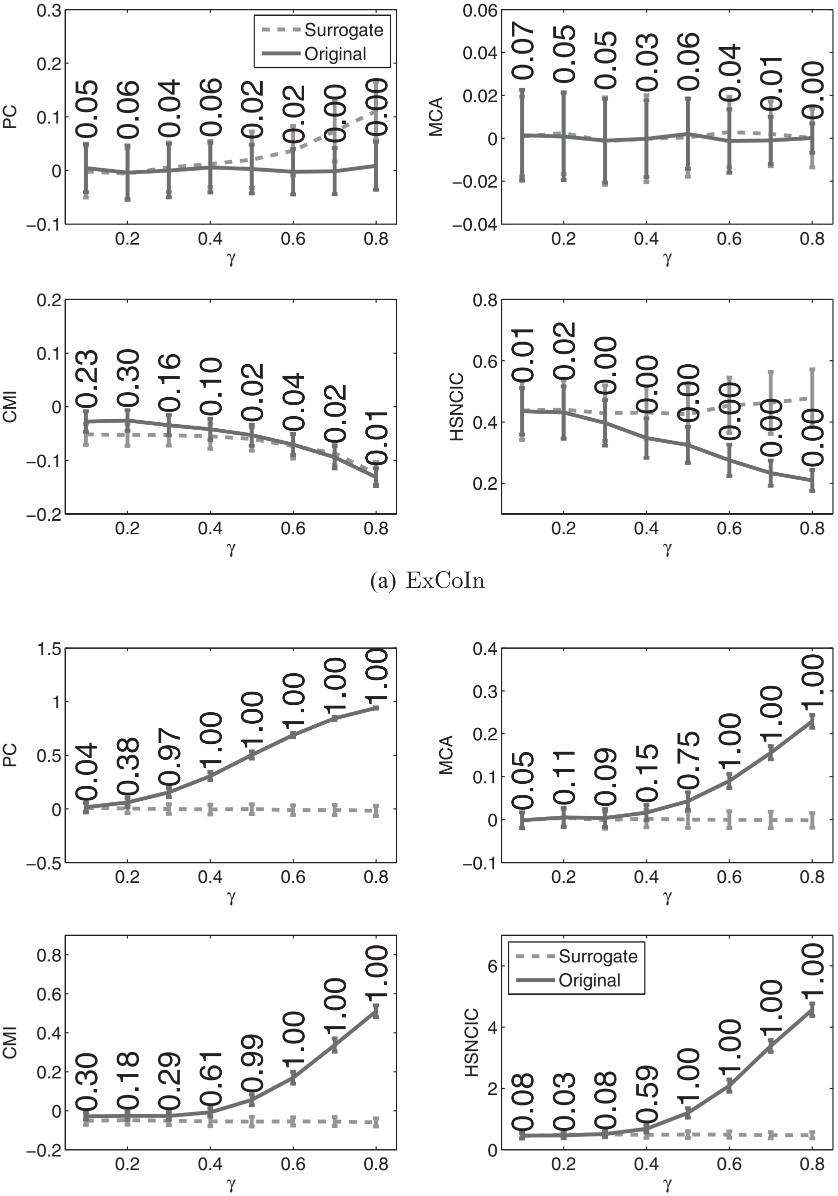
- *Conditionally dependent but independent variables*. Consider three independent normally distributed random variables X, Y, and ϵ, $X, Y, \epsilon \sim \mathcal{N}(0, 1)$, and a third random variable $Z = \gamma(X - Y) + (1 - \gamma)\epsilon$ where $0 < \gamma < 1$. Here Y and X are independent, but they are conditionally dependent given the event $Z = z$ since then $Y = X - (z - (1 - \gamma)\epsilon)/\gamma$, that is, the value of Y can be partially determined from the value of X and vice versa. Therefore, we should expect that $\mathrm{GMA}(X, Y) \approx 0.5$ and $\mathrm{MCA}(X, Y; Z) > 0$. We call this example ExCoDe.

- *Conditionally independent but dependent variables*. Consider three independent normally distributed random variables $Z, \epsilon_1, \epsilon_2$, that is, $Z, \epsilon_1, \epsilon_2 \sim \mathcal{N}(0, 1)$, and construct two random variables X and Y where $X = \gamma Z + (1-\gamma)\epsilon_1$ and $Y = \gamma Z + (1-\gamma)\epsilon_2$ where $0 < \gamma < 1$. Here X and Y are dependent, since both originate from Z and are corrupted by two independent noises. However, these two random variables are conditionally independent given the event $Z = z$ since then $X = \gamma z + (1-\gamma)\epsilon_1$ and $Y = \gamma z + (1-\gamma)\epsilon_2$, which are independent by construction. Therefore, we should expect that $\text{GMA}(X, Y) > 0.5$ and $\text{MCA}(X, Y; Z) \leq 0$. We call this example, ExCoIn.

Notice that the quality of the surrogate data should be judged by a measure of conditional dependence. However, the significance of the measured value itself is judged by the surrogate data. Therefore, to evaluate the quality of the surrogate data, we rely on a simpler and more established measure of conditional dependence, the partial correlation (PC), along with the methods described in this letter: MCA, CMI, and HSNCIC. PC can be reliably applied in these two examples since the joint probability law for both the examples is gaussian. For CMI, we use $k = \lceil \sqrt{n} \rceil$, whereas for HSNCIC, we use a gaussian kernel with the kernel size set to the median of the intersample distances and set the regularization value to $1/n$. We use 1000 sets of 500 realizations to estimate the true values, $T = 100$ surrogates to judge the significance of these values, and set $\alpha = 0.05$. In Figure 1, we show the performance of these methods.

We observe that for both examples, the surrogate values exist around zero for PC and MCA, which is promising. However, the distributions of surrogate values for CMI and HSNCIC are biased, which is an usual observation for finite sample estimation. For ExCoDe, the true values are much higher than the surrogate values, which indicates the presence of conditional dependence, whereas for ExCoIn, the true values are almost identically distributed as the surrogate values, which is expected. Also, for ExCoDe, we observe a monotonic increase in the estimated values, a desired characteristic of a measure of conditional dependence.

For ExCoDe, γ controls the difficulty of the problem in terms of signal-to-noise ratio, that is, a higher γ injects more noise (ϵ) relative to the signal (X), making it difficult to observe the contribution of X on Y given Z. The proportion of significant values out of all trials is a sign of how accurately the methods have assessed conditional dependence. We observe that PC achieves the best performance in precisely identifying conditional dependence at $\gamma = 0.4$. The performance of PC is justified since the original realizations are gaussian in nature. The performances of CMI and HSNCIC are better than that of MCA. However, this performance is achieved by proper selection of the parameter values and at the expense of more computational cost. Also, we observe that CMI tends to wrongfully establish conditional dependence in both examples.



(a) ExCoIn

(b) ExCoDe

Figure 1: The performance of the proposed surrogate data generation procedure and measures of conditional dependence for two examples. The vertical values in the plots are the fraction of significant values. See section 4.1.

Although the true and the surrogate values of the measures for ExCoDe follow the desired pattern, that is, the true values monotonically increase for increasing γ and the surrogate values maintain a steady low, we do not observe the same effect for ExCoIn. Notice that the surrogate values become larger for PC for large γ. A possible reason for this is that for large γ, the joint distributions of (X, Z) and (Y, Z) are almost singular and thus difficult to estimate. Therefore, it is possible that the surrogate data generated from these distributions are not accurate and therefore not gaussian, thus manipulating the values returned by PC. HSNCIC demonstrates a different behavior. Although it maintains a similar surrogate value distribution over different γ, the true value estimated by this measure monotonically decreases with increasing γ, thus forcing it to miss detecting conditional dependence. A probable cause of this observation is, again, the choice of free parameters. Since for large γ, the probability law becomes narrower, HSNCIC perhaps require a smaller kernel size for proper estimation than the selected kernel size.

In summary, this experiment shows that, first, the surrogate data generated by the proposed method may not be accurate with respect to all measures of conditional dependence, but they are still sufficiently reliable for assessing significance in the context of MCA, and second, the existing measures of conditional dependence rely on the accurate choice of parameters to make a proper decision (i.e., to avoid misleading assessment of significance).

5 Simulation

In this section, we apply the proposed approach to several real and synthetic data sets over varying sample sizes and dimensionality for assessing causal strength in the sense of Granger (1980) and addressing its pros and cons. The other available methods of conditional dependence are usually applied to small-scale problems due to their computational load and inherent difficulty in choosing appropriate values of the free parameters (Fukumizu et al., 2008; Sun, 2008; Seth & Príncipe, 2012).

5.1 Conditional Granger Causality.

Given stationary stochastic processes $\{X_t\}$ and $\{Y_t\}$, $\{X_t\}$ is said to cause $\{Y_t\}$, that is, $\{X_t\} \rightarrow \{Y_t\}$, in the sense of Granger, if the past values $[X_{t-1}, X_{t-2}, \ldots]$ of $\{X_t\}$ contain additional information about the present value Y_t of $\{Y_t\}$ that is not contained in past values $[Y_{t-1}, Y_{t-2}, \ldots]$ of $\{Y_t\}$ alone (Granger, 1980), that is,

$$\{X_t\} \nrightarrow \{Y_t\} \iff Y_t \perp (X_{t-1}, X_{t-2}, \ldots) | (Y_{t-1}, Y_{t-2}, \ldots).$$

We apply MCA as a measure of causal influence to assess how strongly one process affects the other. Notice that the use of MCA is justified in the context of time series since we have not made any assumption that the samples $\{x_i, y_i, z_i\}_{i=1}^n$ are independent and the intuition behind assessing causal strength is very similar to the intuition behind conditional association: a strong causal influence implies that knowing the past of one time series brings the current observations of the other time series closer than just knowing the past of the latter time series alone. Notice, however, that the surrogate data generation technique that we explored assumes that $\{(x_i, y_i, z_i)\}_{i=1}^n$ are independent realizations and the resulting surrogate samples $\{(x_i^{(t)}, y_i^{(t)}, z_i^{(t)})\}_{i=1}^n$ are independent of each other. This is not desirable since we fail to capture the dependence structure in the surrogate data; but it is a standard practice (Su & White, 2008).[7] A better method for generating surrogate data in the context of assessing Granger causal influence remains an open area of research. Also, we assume that the time series under evaluation is stationary, that is, the samples $\{(x_i^{(t)}, y_i^{(t)}, z_i^{(t)})\}_{i=1}^n$ originate from the same probability law.

The issue of causal influence between two time series can be trivially extended to multivariate time series involving three or more time series where it is often desired to separate a direct cause from an indirect one, that is, to judge if the time series $\{X_t\}$ and $\{Y_t\}$ are causally connected or not, given a third time series $\{Z_t\}$. $\{X_t\}$ is said to cause $\{Y_t\}$ given $\{Z_t\}$, that is, $\{X_t\} \to \{Y_t\}|\{Z_t\}$, if the past values $[X_{t-1}, X_{t-2}, \ldots]$ of $\{X_t\}$ contain additional information about the present value Y_t of $\{Y_t\}$ that is not contained in past values $[(Y_{t-1}, Z_{t-1}), (Y_{t-2}, Z_{t-2}), \ldots]$ of $\{Y_t, Z_t\}$. In terms of conditional independence, $\{X_t\} \not\to \{Y_t\}|\{Z_t\}$ implies that the present value Y_t of $\{Y_t\}$ is conditionally independent of the past values $[X_{t-1}, X_{t-2}, \ldots]$ of $\{X_t\}$ given past values $[(Y_{t-1}, Z_{t-1}), (Y_{t-2}, Z_{t-1}), \ldots]$ of $\{Y_t, Z_t\}$.

For our experiments, we generate a multivariate time series $\{W\}$ with two to five elements. To separate a direct cause from an indirect cause, we quantify the conditional causal influence, that is, we quantify the causal influence of $\{W_i\}$ on $\{W_j\}$ by the conditional association of $Y = W_j(t)$ with $X = [W_i(t-1), \ldots, W_i(t-L)]$ given $Z = [W_j(t-1), \ldots, W_j(t-L), W_{\sim i,j}(t-1), \ldots, W_{\sim i,j}(t-L)]$ where $W = W_i \cup W_j \cup W_{\sim i,j}$, and L is the number of past values that we condition on. Since we are working in the Euclidean space, we use the l_2 distance as metric for all three spaces $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{Z}$. For each experiment, we use $T = 100$ trials to evaluate the significance of the acquired values and present our results over 128 repetitions of the same experiment. For all the simulation results, we observe a common feature that the performance of MCA in terms of providing significant values improves over increasing sample sizes n and it degrades over

[7] It has been established that the asymptotic null distribution of (smoothed) measures of conditional independence is usually independent of certain dependence structure. However, a similar proof for finite samples and arbitrary dimension is usually unavailable in the literature.

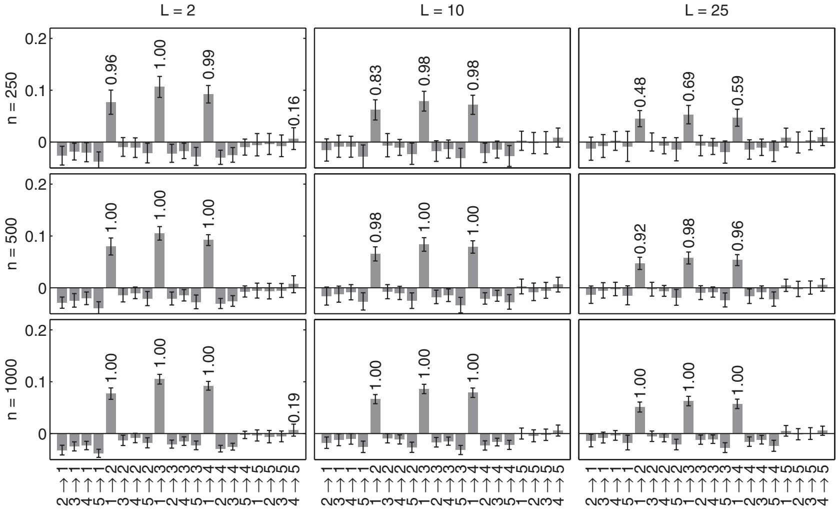


Figure 2: The MCA values obtained while evaluating the conditional causal flow in the network described in section 5.2. The bars show the means of the estimated MCA values, and the lines around them show the variations in terms of plus or minus unit standard deviation. The vertical numbers denote the fraction of significant values and are shown only when they are larger than 0.1.

increasing embedding dimensions L. This is expected since L controls the dimensionality, and the higher the value of L, the more difficult and sparse the problem is.

5.2 Linear System. First, we consider the following linear system (Zou & Feng, 2009),

$$\begin{aligned}
W_1(t) &= 0.95\sqrt{2}W_1(t-1) - 0.9025W_1(t-2) + \eta_1,\\
W_2(t) &= 0.5W_1(t-2) + \eta_2,\\
W_3(t) &= -0.4W_1(t-2) + \eta_3,\\
W_4(t) &= -0.5W_1(t-1) + 0.25\sqrt{2}(W_4(t-1)\\
&\qquad + W_5(t-1)) + \eta_4,\\
W_5(t) &= -0.25\sqrt{2}(W_4(t-1) - W_5(t-2)) + \eta_5,
\end{aligned}$$

where $\eta_1, \eta_5 \sim \mathcal{N}(0, 0.6^2)$, $\eta_2 \sim \mathcal{N}(0, 0.5^2)$, $\eta_3 \sim \mathcal{N}(0, 0.3^2)$, and $\eta_4 \sim \mathcal{N}(0, 0.3^2)$. Here $\{W_1\}$ causes $\{W_2\}$, $\{W_3\}$, and $\{W_4\}$, whereas $\{W_4\}$ and $\{W_5\}$ cause each other. We present the result of this experiment in Figure 2 with $L = 2, 10, 25$, and $n = 250, 500, 1000$. We observe that MCA has been able to recover (positive MCA) the causal flow between $\{W_1\} \to \{W_2\}, \{W_3\}, \{W_4\}\}$ with high efficiency while suppressing (negative MCA) the others where there are no causal flows. However, it has failed to establish significant causal flow between $\{W_4\} \leftrightarrow \{W_5\}$ for low sample sizes and high embedding dimensions. A possible explanation of this observation is that the connection strengths between $\{W_4\}$ and $\{W_5\}$ are rather weak compared to the other existing connections. A similar characteristic is also observed for other measures of conditional dependence (Seth & Príncipe, 2012).

5.3 Nonlinear System. Next, we consider the following nonlinear system (Narendra, 1997),

$$W_1(t+1) = \left(1 + \frac{W_1(t)}{1+W_1^2(t)}\right)\sin W_2(t) + \eta_1,$$

$$W_2(t+1) = W_1(t)\exp\left(-\frac{W_1^2(t)+W_2^2(t)}{8}\right) + W_2(t)\cos W_2(t)$$
$$+ \frac{W_4^3(t)}{(1+W_4(t))^2 + 0.5\cos(W_1(t)+W_2(t))} + \eta_2,$$

$$W_3(t+1) = \frac{W_1(t)}{1+0.5\sin W_2(t)} + \frac{W_2(t)}{1+0.5\sin W_1(t)} + \eta_3,$$

where $W_4(t) \sim \mathcal{N}(0,1)$ is an external input, and $\eta_{1,2,3} \sim \mathcal{N}(0, 0.1^2)$. Here $\{W_1\}$ and $\{W_2\}$ cause each other, and they both cause $\{W_3\}$, whereas $\{W_4\}$ causes $\{W_2\}$. We present the result of this experiment in Figure 3 for $L = 2, 10, 25$ and $n = 250, 500, 1000$. We observe that although MCA has been able to recover the rest of the causal flow well, it has failed to establish the causal flow between $\{W_1\} \to \{W_2\}, \{W_3\}$ for small sample sizes and higher embedding dimensions. This observation can be attributed to the nonlinear nature of the signal, where for high-embedding dimensions, the contribution of the nonlinearity is overshadowed by the other components (Seth & Príncipe, 2012). We have observed that like other measures of conditional dependence (Seth & Príncipe, 2012), MCA has been able to recover the true causal flow with very high efficiency for $L = 1$. This observation signifies the contribution of proper embedding dimensions and that it might be crucial for detecting causal influence in the presence of nonlinearity.

5.4 Varying Coupling Strength. Next, we consider the following time series (Chen, Rangarajan, Feng, & Ding, 2004),

$$W_1(t) = 3.4W_1(t-1)\big(1 - W_1^2(t-1)\big)e^{-W_1^2(t-1)} + 0.8W_1(t-2) + \eta_1,$$

$$W_2(t) = 3.4W_2(t-1)\big(1 - W_2^2(t-1)\big)e^{-W_2^2(t-1)}$$
$$+ 0.5W_2(t-2) + \gamma W_1^2(t-2) + \eta_2,$$

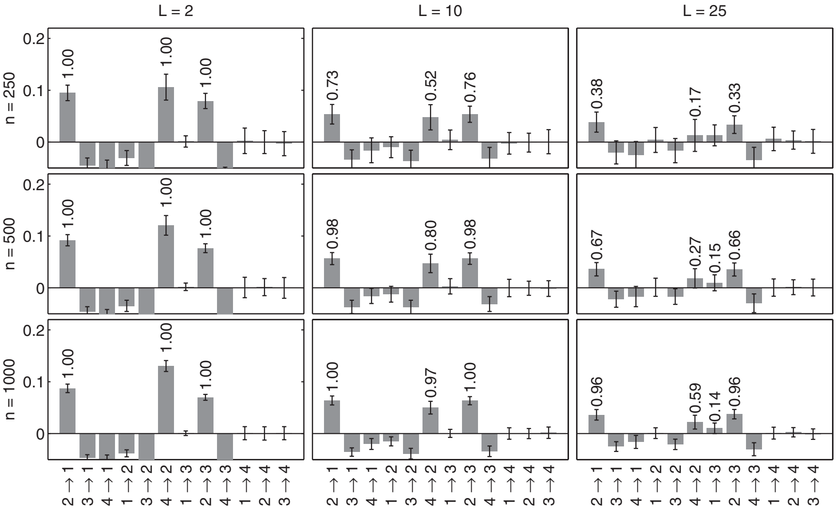


Figure 3: The MCA values obtained while evaluating the conditional causal flow in the network described in section 5.3. The bars show the means of the estimated MCA values, and the lines around them show the variations in terms of plus-or-minus unit standard deviation. The vertical numbers denote the fraction of significant values and are shown only when they are larger than 0.1.

where $\eta_1, \eta_2 \sim \mathcal{N}(0, 0.1^2)$ and $0 \leq \gamma \leq 1$ is the coupling strength. Here $\{W_1\}$ causes $\{W_2\}$ for $\gamma > 0$, and the causal flow increases with increasing γ. Due to the nonlinear nature of this time series, Granger causality fails to detect the true causal direction. Therefore, Chen et al. (2004) proposed a nonlinear version of Granger causal inference by piece-wise linear approximation that successfully captures the true causal direction. However, this approach requires a relatively high number of realizations to sample the time series well. From Figure 4, we observe that MCA has been able to produce the expected result, it clearly demonstrates the increase in causal flow in the MCA values over different sample sizes and embedding dimensions. Also, the proportion of significant values increases with γ as the problem becomes easier.

5.5 Multivariate Time Series. Next, we consider the following bivariate time series where each element, $\{W_1\}$ or $\{W_2\}$, is four-dimensional (Fukumizu et al., 2008),

$$
\begin{aligned}
W_1^{(1)}(t+1) &= 1.4 - W_1^{(1)}(t)^2 + 0.3W_1^{(2)}(t),\\
W_1^{(2)}(t+1) &= W_1^{(1)}(t),\\
W_1^{(2)}(t+1) &= 1.4 - \left\{cW_1^{(1)}(t)W_2^{(1)}(t) + \ (1-\gamma)\, W_2^{(2)}(t)^2\right\} + 0.1W_2^{(2)}(t),\\
W_2^{(2)}(t+1) &= W_2^{(1)}(t),
\end{aligned}
$$

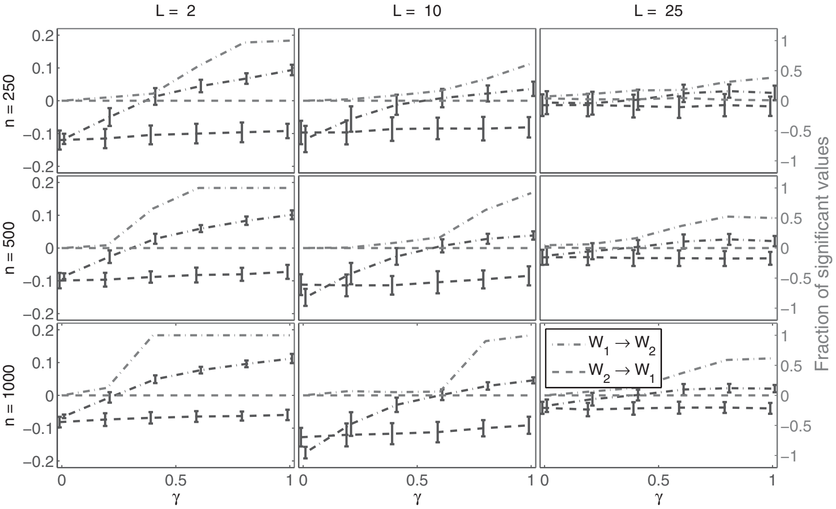


Figure 4: The MCA values obtained while evaluating the conditional causal flow in the network described in section 5.4. The dark error bars show the means and standard deviations of the estimated MCA values. The lighter lines show the fraction of significant values, and they follow the axis on the right.

where $W_1^{(3)}, W_1^{(4)}, W_2^{(3)}, W_2^{(4)} \sim \mathcal{N}(0, 0.5^2)$, and $0 \leq \gamma \leq 0.6$ is the coupling strength, and the causal flow increases with increasing γ. This problem is relatively difficult since the dimensionality of the joint space is much higher. However, from Figure 5, we observe that although worse than the previous example, MCA still successfully indicates the increase in causal flow for increasing γ with increasing significant values.

5.6 Heart Rate and Respiration Force. Next, we apply MCA on the bivariate time series of heart rate (H) and respiration force (R) of a patient with sleep apnea. This data have been acquired from the Santa Fe Institute time series competition.[8] Normally, respiration force has a causal influence on heart rate. However, for a patient with sleep apnea, this causal direction is reversed. Therefore, for this data set, a strong causal influence is observed from heart rate to respiration force, whereas a weak influence is observed in the other direction (Schreiber, 2000). We randomly select $n = 480, 720, 960$ long segments from the time series, and use $L = 32, 40, 48, 56, 64, 72$. Since, the time series is sampled at 2 Hz, these measurements are equivalent to $n = 4, 6, 8$ minutes and $L = 16, 20, 24, 28, 32, 36$ seconds. We observe from

[8] http://www-psych.stanford.edu/∼andreas/Time-Series/SantaFe.html.

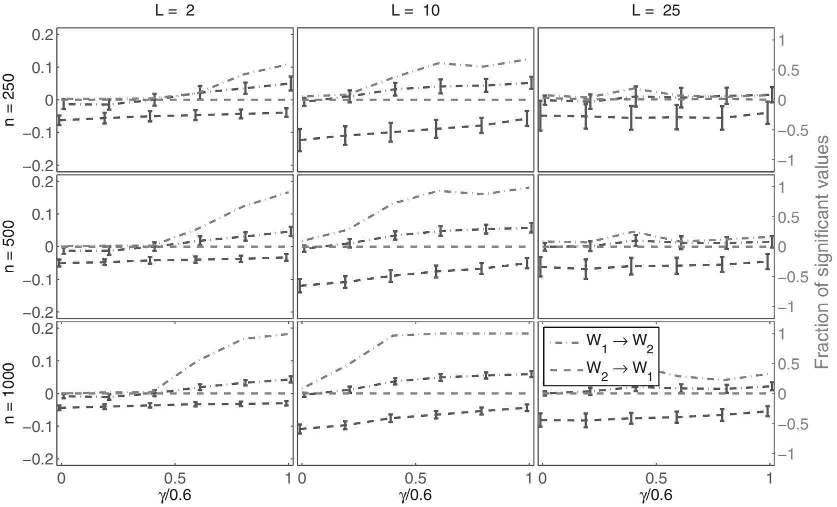


Figure 5: The MCA values obtained while evaluating the conditional causal flow in the network described in section 5.5. The dark error bars show the means and standard deviations of the estimated MCA values. The lighter lines show the fraction of significant values, and they follow the axis on the right.

Figure 6 that MCA strongly supports the causal influence of heart rate over respiration force.

6 Discussion

In this letter, we introduced the concept of conditional association as a substitute for conditional dependence. The major difference between the proposed approach and existing ones is that it explores the concept of conditional dependence in the context of the realizations rather than random variables or the probability law. Unlike available measures of association, the proposed approach is parameter free and relatively easy to compute, thus making it an excellent tool for inferring causal influence among random variables and stochastic processes. We have also introduced a novel scheme for generating surrogate data to evaluate the significance of the acquired value, which is attractive since it resamples the original data, allowing the computations involved in computing the conditional measure to be reused to compute surrogate values.

Although the initial results of these two approaches are very promising, a few aspects require further investigation:

- We have observed that the proposed approach usually provides less power compared to PC in the context of gaussian probability law.

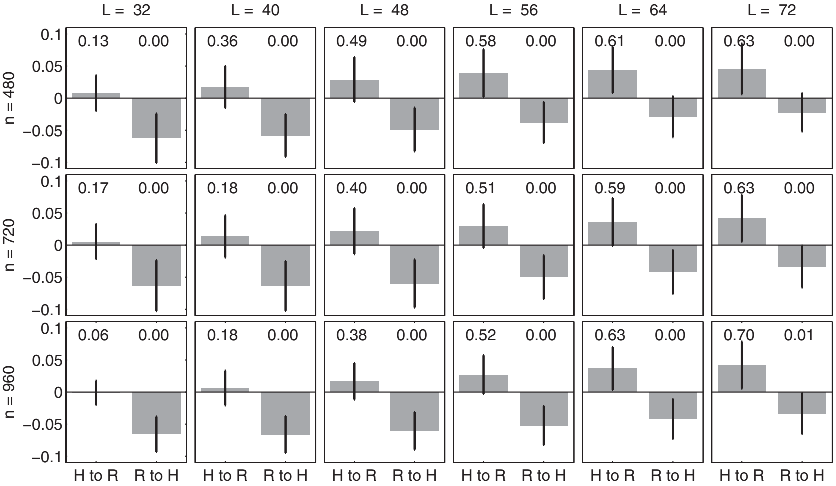


Figure 6: The performance of MCA in evaluating the causal flow in the network presented in section 5.6. The bars show the means of the estimated MCA values, whereas the lines around them show the variation in terms of plus-or-minus unit standard deviation. The horizontal numbers indicate the fraction of significant values.

Although this is not a drawback of this method, it is certainly an undesirable property. Therefore, more sophisticated measures of association should be investigated in order to improve the performance of the proposed approach in terms of statistical power.

- We have observed that the quality of the surrogate values generated by the proposed method is somewhat poor when the probability law is close to degenerate. This issue should be explored in more detail. Also, the full extent of the effect of the free neighborhood parameter remains to be explored.
- We have explored the measure of conditional association only the context of the realizations. However, a corresponding population version of this approach would be interesting to investigate in order to establish if conditional association is a necessary and sufficient condition for conditional independence.

Finally, we have noted that the proposed approach can be applied to any metric spaces. However, in this letter, we have restricted ourselves to Euclidean space, partly due to simplicity in understanding and also due to the unavailability of surrogate data generation technique. It would be interesting to apply this approach to more abstract spaces to fully understand its capabilities and limitations.

Acknowledgments

This project is partially supported by the NSF grant IIS-096419. We thank Austin Brockmeier for many insightful discussions.

References

Bontempi, G., & Meyer, P. E. (2010). Causal filter selection in microarray data. In *Proceedings of the 27th International Conference on Machine Learning*. Madison, WI: Omnipress.

Bouezmarni, T., Rombouts, J. V. K., & Taamouti, A. (2009). *A nonparametric copula based test for conditional independence with applications to Granger causality* (CIRANO working papers). Montreal: CIRANO, 2009.

Chen, Y., Rangarajan, G., Feng, J., & Ding, M. (2004). Analyzing multiple nonlinear time series with extended Granger causality. *Physics Letters A*, *324*, 26.

Dawid, A. P. (1998). *Conditional independence*. New York: Wiley-Interscience.

Dhamala, M., Rangarajan, G., & Ding, M. (2008). Analyzing information flow in brain networks with nonparametric Granger causality. *NeuroImage*, *41*, 354–362.

Diks, C., & DeGoede, J. (2001). *Global analysis of dynamical systems*. London: IOP Publishing.

Diks, C., & Panchenko, V. (2006). A new statistic and practical guidelines for nonparametric Granger causality testing. *Journal of Economic Dynamics and Control*, *30*, 1647–1669.

Fukumizu, K., Bach, F. R., & Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.*, *5*, 73–99.

Fukumizu, K., Gretton, A., Sun, X., & Schölkopf, B. (2008). Kernel measures of conditional dependence. In J. C. Platt, D. Koller, Y. Singer, & S. Rowels (Eds.), *Advances in neural information processing systems, 20* (pp. 489–496). Cambridge, MA: MIT Press.

Granger, C. (1980). Testing for causality: A personal viewpoint. *J. Economic Dynamics and Control*, *2*, 329–352.

Holland, P. W., & Rosenbaum, P. R. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics*, *14*, 1523–1543.

Joe, H. (1989). Relative entropy measures of multivariate dependence. *Journal of the American Statistical Assocation*, *84*(405), 157–164.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1), 81–93.

Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review, E: Statistical, Nonlinear, and Soft Matter Physics, 69* (6 Pt. 2).

Linton, O., & Gozalo, P. (1996). *Conditional independence restrictions: Testing and estimation*. (Cowles Foundation Discussion Papers 1140). New Haven, CT: Cowles Foundation, Yale University.

Lozano, C. A., Naoki, N., Liu, Y., & Rosset, S. (2009). Grouped graphical Granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, *25*(12), 110–118.

Narendra, K. S. (1997). Neural networks for intelligent control. *American Control Conference Workshop*, 1997.

Paparoditis, E., & Politis, D. (2000). The local bootstrap for kernel estimators under general dependence conditions. *Annals of the Institute of Statistical Mathematics*, *52*(1), 139–159.

Pérez-Cruz, F. (2008). Estimation of information theoretic measures for continuous random variables. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems, 21* (pp. 1257–1264). Cambridge, MA: MIT Press.

Quinn, C. J., Coleman, T. P., Kiyavash, N., & Hatsopoulos, N. G. (2011). Estimating the directed information to infer causal relationships in ensemble neural spike train recordings. *J. Comput. Neurosci.*, *30*, 17–44.

Schreiber, T. (2000). Measuring information transfer. *Physical Review Letters*, *85*(2), 461–464.

Seth, S., Park, I., Brockmeier, A., Semework, M., Choi, J., Francis, J., et al. (2010). A novel family of non-parametric cumulative based divergences for point processes. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, & A. Culotta (Eds.), *Advances in neural information processing systems, 23* (pp. 2119–2127). Red Hook, NY: Curran Associates.

Seth, S., & Príncipe, J. C. (2012). A test of Granger non-causality based on nonparametric measure of conditional independence. *IEEE Transactions on Neural Network*, *23*, 47–59.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*(1), 72–101.

Su, L., & White, H. (2007). A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, *141*, 807–834.

Su, L., & White, H. (2008). A nonparametric Hellinger metric test for conditional independence. *Econometric Theory*, *24*, 829–864.

Sun, X. (2008). Assessing nonlinear Granger causality from multivariate time series. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases—Part II* (pp. 440–455). New York: Springer.

Sun, X., Janzing, D., Schölkopf, B., & Fukumizu, K. (2007). A kernel-based causal learning algorithm. In *Proceedings of the 24th International Conference on Machine Learning* (pp. 855–862). New York: ACM.

Zou, C., & Feng, J. (2009). Granger causality vs. dynamic Bayesian network inference: A comparative study. *BMC Bioinformatics*, *10*(1), 122.

Received October 21, 2011; accepted December 28, 2011.