Behaviormetrika (2022) 49:343–362
https://doi.org/10.1007/s41237-022-00170-2

INVITED PAPER

A Wasserstein-based measure of conditional dependence

Jalal Etesami[1] · **Kun Zhang**[2] · **Negar Kiyavash**[1]

Received: 31 December 2021 / Accepted: 10 May 2022 / Published online: 25 June 2022

Abstract
Measuring conditional dependencies among the variables of a network is of great interest to many disciplines. This paper studies some shortcomings of the existing dependency measures in detecting direct causal influences or their lack of ability for group selection to capture strong dependencies and accordingly introduces a new statistical dependency measure to overcome them. This measure is inspired by Dobrushin's coefficients and based on the fact that there is no dependency between X and Y given another variable Z, if and only if the conditional distribution of Y given $X = x$ and $Z = z$ does not change when X takes another realization x' while Z takes the same realization z. We show the advantages of this measure over the related measures in the literature. Moreover, we establish the connection between our measure and the integral probability metric (IPM) that helps to develop estimators of the measure with lower complexity compared to other relevant information theoretic-based measures. Finally, we show the performance of this measure through numerical simulations.

Keywords Conditional dependence measure · Causality · Wasserstein

Communicated by Shohei Shimizu.

✉ Jalal Etesami
seyed.etesami@epfl.ch

Kun Zhang
kunz1@cmu.edu

Negar Kiyavash
negar.kiyavash@epfl.ch

[1] Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland

[2] Carnegie Mellon University, Pittsburgh, USA

1 Introduction

Identifying the conditional independencies (CIs) among the variables or processes in a systems is a fundamental problem in scientific investigations in different fields such as biology, econometric, social sciences, and many others.

In probability theory, two events X and Y are conditionally independent given a third event Z, if the occurrence or non-occurrence of X and Y are "independent" events in their conditional probability distribution given Z (Gorodetskii 1978). There are several CI measures in literature that have been developed for different applications to capture such independency. For instance, the most commonly used one is conditional mutual information (CMI) (Gorodetskii 1978) that is an information theoretical quantity. This measure has been used in different fields such as communication engineering, channel coding (Cover and Thomas 2012), and causal discovery (Spirtes et al. 2000b). CMI between X and Y given Z is defined by comparing two conditional distributions: $P(X|Y, Z)$ and $P(X|Z)$ using KL-divergence and then taking average over the conditioning variable Z. Hence, it is limited to those realizations with positive probability (see Sect. 4.1). One shortcoming of such measure is that it cannot capture CIs that occur rarely or even over zero measure sets. Another shortcoming of this measure is that it is symmetric and thus it fails to encode asymmetric dependencies such as causal directions in a network.

Most of the conditional dependency or independency measures are defined similar to the CMI in a sense that they take average over the conditioning variables. Kernel-based method in Zhang et al. (2011) is another example. Consequently, such measures may fail to distinguish the range of the conditioning variable Z in which the dependency between the variables of interest X and Y is more clearer. For example, consider a treatment that has different effects on a special disease for different genders. There are scenarios in which the previous CI measures (e.g., CMI) fail to identify for which gender the effect of the treatment on the disease is maximized (see Sect. 4.3).

Discovering the causal relationships in a network is one of the main applications for CI measures (Spirtes et al. 2000b). In this area, it is important to capture the direct causal influence between two variables in a network independent of the other causal indirect influences between them. As we will show in Sect. 4.2, previous CI measures (e.g., CMI) cannot capture the direct causal influences between two variables (cause and effect) in a network when some variables in the indirect causal path depend on the cause almost deterministically.

The main contribution of this paper is the introduction of a statistical metric inspired by Dobrushin's coefficient (Dobrushin 1970) to measure the dependency or independency between X and Y given Z in a network from their realizations. Our metric has been developed based on the paradigm that if Y has no dependency on X given Z, then the conditional distribution of Y given $X = x$ and $Z = z$ will not change if x varies and Z takes the same realization z. We will show that this dependency measure overcomes the aforementioned limitations. Moreover, we will establish the connection between our measure and the IPM to develop estimators for our metric with lower complexity compared to other relevant information-theoretic based

measures such as CMI. This is because the proposed estimators depend on the sample points only through the metric of the space, and thus its complexity is independent of the dimension of the samples.

Perhaps the best known paradigm for visualizing the CIs among the variables of a network is Bayesian networks (Pearl 2003). They are directed acyclic graphs (DAGs) in which nodes represent random variables and directed edges denote the direction of causal influences. Analogously, using the dependency measure in this work, we can represent the causal structure of a network via a DAG that possesses the same properties as the Bayesian networks.

It is also worth mentioning that there exist several measures to capture CIs and the causal influences among time series, for instance, transfer entropy (Schreiber 2000) and directed information (Massey 1990). Measuring the reduction of uncertainty in one variable after knowing another variable is the key idea in such measures. Because these measures are defined based on CMI, they also suffer the aforementioned limitations. Note that the proposed measure can easily be modified to capture such influences in time series as well.

2 Definitions

In this Section, we review some basic definitions and our notation. Throughout this paper, we use capital letters to represent random variables, lowercase letters to denote a realization of a random variable, and bold capital letters to denote matrices. We denote a subset of random variables with index set $\mathcal{K} \subseteq [m]$, where $[m] := \{1, ..., m\}$ by $\underline{X}_{\mathcal{K}}$ and $[m] \setminus \{j\}$ by $-\{j\}$.

In a directed graph $\overrightarrow{G} = (V, \overrightarrow{E})$, we denote the parent set of a node $i \in V$ by $Pa_i := \{j : (j, i) \in \overrightarrow{E}\}$, and denote the set of its non-descendant[1] by Nd_i. We use $X \perp\!\!\!\perp Y | Z$ to denote X and Y are independent given Z.

Bayesian Network: A Bayesian network is a graphical model that represents the conditional independencies among a set of random variables via a directed acyclic graph (DAG) (Spirtes et al. 2000b). A set of random variables $\underline{X}$ is Bayesian with respect to a DAG $\overrightarrow{G}$, if

$$P(\underline{X}) = \prod_{i=1}^{m} P(X_i | \underline{X}_{Pa_i}). \tag{1}$$

Up to some technical conditions (Lauritzen 1996), this factorization is equivalent to the *causal Markov* condition. Causal Markov condition states that a DAG is only acceptable as a possible causal hypothesis if every node is conditionally independent of its non-descendant given its parents.

[1] A node v is a non-descendant of another node u, if there is no direct path from u to v.

Corresponding DAG of a joint distribution possesses *Global Markov* condition if for any disjoint set of nodes $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ for which $\mathcal{A}$ and $\mathcal{B}$ are d-separated[2] by $\mathcal{C}$, then $\underline{X}_{\mathcal{A}} \perp\!\!\!\perp \underline{X}_{\mathcal{B}} | \underline{X}_{\mathcal{C}}$. It is shown in Lauritzen (1996) that causal Markov condition and Global Markov condition are equivalent.

Faithfulness: A joint distribution is called *faithful* with respect to a DAG if all the conditional independence (CI) relationships implied by the distribution can also be found from its corresponding DAG using d-separation and vice versa[3]Judea (2014). It is possible that several DAGs encode the same set of CI relationships. In this case, they are called Markov equivalence.

3 New dependency measure

As we mentioned earlier, we use the following paradigm to define our measure of independency: if Y has no dependency on X given Z, then the conditional distribution of Y given $X = x$ and $Z = z$ should not change when X takes different realization x' while Z takes the same realization z. This paradigm is similar in nature to Pearl's paradigm of causal influence (Pearl 2003). He proposed that the influence of a variable (potential cause) on another variable (effect) in a network is assessed by assigning different values to the potential cause, while other variables' effects are removed, and observing the behavior of the effect variable. Below, we formally introduce our dependency measure.

Consider $\underline{X}$ a collection of m random variables. To identify the dependency of X_i on X_j, we select a set of indices $\mathcal{K}$, where $\mathcal{K} \subseteq -\{i,j\}$ and consider the following two probability measures:

$$\begin{aligned} \mu_i(\underline{x}_{\mathcal{K}\cup\{j\}}) &:= P\Big(X_i \Big| \underline{X}_{\mathcal{K}\cup\{j\}} = \underline{x}_{\mathcal{K}\cup\{j\}}\Big), \\ \mu_i(\underline{y}_{\mathcal{K}\cup\{j\}}) &:= P\Big(X_i \Big| \underline{X}_{\mathcal{K}\cup\{j\}} = \underline{y}_{\mathcal{K}\cup\{j\}}\Big), \end{aligned} \tag{2}$$

where $\underline{x}_{\mathcal{K}\cup\{j\}}$ and $\underline{y}_{\mathcal{K}\cup\{j\}} \in E^{|\mathcal{K}|+1}$ are two realizations for $\underline{X}_{\mathcal{K}\cup\{j\}}$ that are the same every where except at X_j. Further, assume $\underline{x}_{\mathcal{K}\cup\{j\}}$ at position X_j equals x and $\underline{y}_{\mathcal{K}\cup\{j\}}$ equals y ($y \neq x$) at this position. If there exists a subset $\mathcal{K} \subseteq -\{i,j\}$ such that for all such realizations, $\mu_i(\underline{x}_{\mathcal{K}\cup\{j\}})$ and $\mu_i(\underline{y}_{\mathcal{K}\cup\{j\}})$ are the same, then we say X_i has no dependency on X_j. This is analogous to the conditional independence that states if X_j and X_i are independent given some $\underline{X}_{\mathcal{K}}$, then there is no causal influence between them. Note that using mere observational data, comparing the two conditional probabilities in (2) reveals the dependency between X_i and X_j. However, when interventional data are available, we can identify whether X_j causes X_i, i.e., the direction of influence.

To compare the two probability measure in (2), a metric on the space of probability measures is required. There are several metrics that can be used such as

[2] It is d-separated by Z if it contains a collider $\rightarrow \cdot \leftarrow$ whose descendants are not in Z or a non-collider in Z.

[3] The set of distributions that do not satisfy this assumption has measure zero (Meek 1995).

KL-divergence, total variation, etc (Gibbs and Edward 2002). For instance, using the KL-divergence will lead to develop CI test-based approaches (Singh and Valtorta 1995). In this work, we use Wasserstein distance and discuss the advantage of using such metric in Sect. 5.1.

Definition 1 Let (E, d) be a metrical complete and separable space equipped with the Borel field $\mathcal{B}$, and let $\mathcal{M}$ be the space of all probability measures on $(E, \mathcal{B})$. Given $\nu_1, \nu_2 \in \mathcal{M}$, the Wasserstein metric between ν_1, ν_2 is given by $W_d(\nu_1, \nu_2) := \inf_\pi \left(\mathbb{E}_\pi[d(x, y)]\right)$, where the infimum is taken over all probability measures π on $E \times E$ such that its marginal distributions are ν_1 and ν_2, respectively.

Using the above distance, we define the dependency of X_i on X_j given $\mathcal{K} \subseteq -\{i, j\}$ as follows:

$$c_{i,j}^{\mathcal{K}} = \sup_{\underline{x}_{\mathcal{K}\cup\{j\}} = \underline{y}_{\mathcal{K}\cup\{j\}},\ \text{off } j} \frac{W_d(\mu_i(\underline{x}_{\mathcal{K}\cup\{j\}}), \mu_i(\underline{y}_{\mathcal{K}\cup\{j\}}))}{d(x, y)}. \tag{3}$$

The supremum is over all realizations $\underline{x}_{\mathcal{K}\cup\{j\}}$ and $\underline{y}_{\mathcal{K}\cup\{j\}}$ that only differ at the jth variable. Moreover, we assume $\underline{x}_{\mathcal{K}\cup\{j\}}$ at jth position equals x and $\underline{y}_{\mathcal{K}\cup\{j\}}$ equals y $(y \neq x)$ at this position. When $\mathcal{K} = -\{i, j\}$, $c_{i,j}^{\mathcal{K}}$ is called Dobrushin's coefficient (Dobrushin 1970). Similarly, we define the dependency of a set of nodes $\mathcal{B}$ on a disjoint set $\mathcal{A}$ given $\mathcal{K}$, where $\mathcal{K} \cap (\mathcal{A} \cup \mathcal{B}) = \emptyset$, as follows,

$$c_{\mathcal{B},\mathcal{A}}^{\mathcal{K}} = \sup_{\underline{x}_{\mathcal{K}\cup\mathcal{A}} = \underline{y}_{\mathcal{K}\cup\mathcal{A}},\ \text{off } \mathcal{A}} \frac{W_d(\mu_{\mathcal{B}}(\underline{x}_{\mathcal{K}\cup\mathcal{A}}), \mu_{\mathcal{B}}(\underline{y}_{\mathcal{K}\cup\mathcal{A}}))}{d(\underline{x}_{\mathcal{A}}, \underline{y}_{\mathcal{A}})}. \tag{4}$$

Remark 1 The dependency measure of i on j given $\mathcal{K}$ in (3) is defined by taking supremum over all realizations. Alternatively we could have taken an average over all realizations. More precisely, we can introduce an alternative measure as follows

$$\int_E \prod_{k\in\mathcal{K}} P(\underline{X}_k = \underline{x}_k) P(X_j = y) P(X_j = x) \frac{W_d\left(\mu_i(\underline{x}_{\mathcal{K}\cup\{j\}}), \mu_i(\underline{y}_{\mathcal{K}\cup\{j\}})\right)}{d(x, y)} d\underline{x}_k dx dy. \tag{5}$$

Clearly, this expression is bounded above by (3). One caveat of taking the expectation versus the supremum is that similar to the conditional mutual information $I(X_i; X_j | \underline{X}_{\mathcal{K}})$ which is also defined via taking an expectation, the measure in (5) cannot capture dependencies that occur over zero measures sets.

3.1 Maximum mean discrepancy

Using a special case of the duality theorem of Kantorovich and Rubinstein (Villani 2003), we obtain an alternative approach for computing the Wasserstein metric as follows:

$$W_d(\nu_1, \nu_2) = \sup_{f \in \mathcal{F}_L} \left| \int_E f d\nu_1 - \int_E f d\nu_2 \right|, \tag{6}$$

where $\mathcal{F}_L$ is the set of all continuous functions satisfying the Lipschitz condition: $||f||_{\text{Lip}} := \sup_{x \neq y} |f(x) - f(y)|/d(x, y) \leq 1$. This representation of the Wasserstein metric is a special form of integral probability metric (IPM) (Müller 1997) that has been studied extensively in probability theory (Dudley 2002) with applications in empirical process theory (Der Vaart and Wellner 1996), transportation problem (Villani 2003), etc. IPM is defined similar to (6) but instead of $\mathcal{F}_L$, the supremum is taken over a class of real-valued bounded measurable functions on E.

One particular instance of IPM is maximum mean discrepancy (MMD) in which the supremum is taken over $\mathcal{F}_{\mathcal{H}} := \{f : ||f||_{\mathcal{H}} \leq 1\}$. More precisely, MMD is defined as

$$\text{MMD}(\nu_1, \nu_2) := \sup_{f \in \mathcal{F}_{\mathcal{H}}} \left| \int_E f d\nu_1 - \int_E f d\nu_2 \right|. \tag{7}$$

Here, $\mathcal{H}$ represents a reproducing kernel Hilbert space (RKHS) (Aronszajn 1950) with reproducing kernel $k(\cdot, \cdot)$. MMD has been used in statistical applications such as independence testing and testing for conditional independence (Gretton et al. 2006; Fukumizu et al. 2007; Sun et al. 2007).

It is shown in Gretton et al. (2006) that when $\mathcal{H}$ is a universal RKHS (Micchelli et al. 2006), defined on the compact metric space E, then $\text{MMD}(\nu_1, \nu_2) = 0$ if and only if $\nu_1 = \nu_2$. In this case, MMD can also be used to compare the two conditional distributions in (2). This is because, $\text{MMD}(\mu_i(\underset{\sim}{x}_{\mathcal{K} \cup \{j\}}), \mu_i(\underset{\sim}{y}_{\mathcal{K} \cup \{j\}})) = 0$ implies that the two conditional distributions are the same. This allows us to define a new dependency measure which we denoted it by $\tilde{c}_{i,j}^{\mathcal{K}}$ similar to (3) that uses MMD instead of Wasserstein distance. It is straight forward to show that this measure has similar properties as the one in (3). The main difference between these two measures is their estimation method that we discuss in Sect. 5.1.

4 Advantages of the dependency measure

Herein, we discuss the advantages of our measure over other dependency measures in the literature.

4.1 Mutual information and information flow

Conditional mutual information is an information theoretic measure that has been used in the literature to identify the conditional independence structure of a network. This measure compares two probability measures $P(X_i|X_j, \underline{X}_{\mathcal{K}})$ and $P(X_i|\underline{X}_{\mathcal{K}})$ using the KL-divergence as follows,

$$I(X_i;X_j|\underline{X}_{\mathcal{K}}) := \sum_{x_i,x_j,\underline{x}_{\mathcal{K}}} P(x_i, x_j, \underline{x}_{\mathcal{K}}) \log \frac{P(x_i|x_j, \underline{x}_{\mathcal{K}})}{P(x_i|\underline{x}_{\mathcal{K}})}. \tag{8}$$

This measure is symmetric and hence it cannot capture the direction of influence. Moreover, it only compares the probability measures over all pairs (X_i, X_j) that have positive probability. Note that any other measures in the literature that is based on conditional independence test such as the kernel-based methods in Sun et al. (2007); Zhang et al. (2011) have the similar limitation.

Example 1 Consider a network of two variables X and Y, in which $X \sim \mathcal{N}(0, 1)$ is a zero mean Gaussian variable and Y is $\mathcal{N}(0, 1)$ whenever X is a rational number and $\mathcal{N}(1, 2)$ otherwise. In this network, Y is dependent on X but it cannot be captured using CI. This is because $I(X;Y) = 0$. On the other hand, we have $c_{y,x} > 0$ and $c_{x,y} = 0$.

Another quantity that has been introduced in the literature to quantify causal influences in a network is information flow (Ay and Polani 2008). This quantity is defined using Pearl's do-calculus (Pearl 2003). Intuitively, operating $do(x_i)$ removes the dependencies of X_i on its parents, and replaces $P(X_i|\underline{X}_{Pa_i})$ with the delta function. Herein, to give an interpretation on how (3) can be used to identify causal relationships that are defined in terms of intervention, we compare our measure with information flow.

Below, we introduce the formal definition of information flow from $\underline{X}_A$ to $\underline{X}_B$ imposing $\underline{X}_{\mathcal{K}}$, $I(\underline{X}_A \to \underline{X}_B|do(\underline{X}_{\mathcal{K}}))$, where A, B, and $\mathcal{K}$ are three disjoint subsets of V.

$$\sum_{\underline{X}_{A\cup B\cup\mathcal{K}}} P(\underline{x}_{\mathcal{K}})P(\underline{x}_A|do(\underline{x}_{\mathcal{K}}))P(\underline{x}_B|do(\underline{x}_{A\cup\mathcal{K}})) \log \frac{P(\underline{x}_B|do(\underline{x}_{A\cup\mathcal{K}}))}{\sum_{\underline{x}'_A} P(\underline{x}'_A|do(\underline{x}_{\mathcal{K}}))P(\underline{x}_B|do(\underline{x}'_A, \underline{x}_{\mathcal{K}}))}. \tag{9}$$

This is defined analogous to the conditional mutual information in (8). But unlike the conditional mutual information, the information flow is defined for all pairs $(\underline{x}_A;\underline{x}_C)$ rather than being limited to those with positive probability (similar to our measure). Similar measures are introduced in Janzing et al. (2013); Nihat and Krakauer (2007) which are also based on do-calculation. Analogously, we can define our measure based on do-operation to capture the direction of causal influences in a network by substituting the conditional distributions in (2) with their *do* versions. More precisely, we use the following measures in (3),

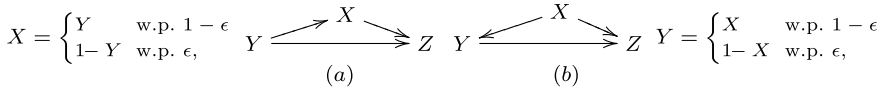

Fig. 1 DAGs for which information flow fails to capture the influence

$$\mu_i^{do}(\underline{x}_{\mathcal{K}\cup\{j\}}) := P\Big(X_i \Big| do(\underline{X}_{\mathcal{K}\cup\{j\}} = \underline{x}_{\mathcal{K}\cup\{j\}})\Big),$$
$$\mu_i^{do}(\underline{y}_{\mathcal{K}\cup\{j\}}) := P\Big(X_i \Big| do(\underline{X}_{\mathcal{K}\cup\{j\}} = \underline{y}_{\mathcal{K}\cup\{j\}})\Big).$$

Because the Wasserstein metric can be estimated using a linear programming (see Sect. 5.1), our measure has computational advantages over the information flow or other similar measures that uses KL-divergence. Another advantage of (3) over the information flow is that it requires less number of interventions in case of using interventional data. More precisely, calculating (9) requires at least two do-operations ($do(\underline{x}_{A\cup\mathcal{K}})$ and $do(\underline{x}_{\mathcal{K}})$) but (3) requires only one ($do(\underline{x}_{\mathcal{K}\cup\{j\}})$). Moreover, as the next example shows, unlike our measure, the information flow depends on the underlying DAG.

Example 2 Consider a network of three binary random variables $\{X, Y, Z\}$ with $Z = X \oplus Y$ an XOR. Suppose the underlying DAG of this network is given by Fig. 1b, in which X takes zero with probability b. In this case, $I(X \to Z|do(Y)) = H(b)$, where H denotes the entropy[4]. However, if the underlying DAG is given by Fig. 1a, we have $I(X \to Z|do(Y)) = H(\epsilon)$. Now, consider a scenario in which ϵ tends to zero. In this scenario, both DAGs describe a system in which $X = Y$ and $Z = 0$. However, in (b), we have $I(X \to Z|do(Y)) = H(b) > 0$, while in (a), $I(X \to Z|do(Y)) \to 0$. But $c^y_{z,x}$ in both DAGs is independent of ϵ and it is positive.

4.2 A better measure for direct causal influences

Consider a network comprises of three random variables $\{X, Y, Z\}$, in which $Y = f(X, W_1)$ and $Z = g(X, Y, W_2)$, such that the transformations from (X, W_1) to (X, Y) and from (X, Y, W_2) to (X, Y, Z) are invertible and W_1 and W_2 are independent exogenous noises. In other words, there exist functions ϕ and φ such that $W_1 = \phi(X, Y)$ and $W_2 = \varphi(X, Y, Z)$. Furthermore, f is an injective function in its first argument, i.e., if $f(x_1, w) = f(x_2, w)$ for some w, then $x_1 = x_2$.

To measure the direct influence from X to Z, one may compute the conditional mutual information between X and Z given Y, i.e., $I(X; Z|Y)$. However, this is not a good measure because as the dependency of Y on X grows, i.e., $H(Y|X) \to 0$, then $I(X;Z|Y) \to 0$. This can be explained by the fact that as $H(Y|X)$ goes to zero, in other words, as P_{W_1} tends to $\delta_{w_0}(W_1)$ for some fixed value w_0, then by specifying the value

[4] More precisely, $H(b) = -b \log b - (1-b)\log(1-b)$.

of X, the ambiguity about the value of Y will go to zero. Thus, using the injective property of f, it is straightforward to show $I(X;Z|Y) \to 0$. Note that if f is not injective, for fixed w_1 and y, there are several x such that $y = f(x, w_1)$. Thus, specifying the value of Y does not determine X uniquely and $I(X; Z|Y)$ will not go to zero.

This analysis shows that $I(X; Z|Y)$ fails to capture the direct influence between X and Z when Y depends on X almost in a deterministic manner. However, looking at $c^y_{z,x}$, we have

$$c^y_{z,x} = \sup_{y,x,x'} \frac{W_d\big(P_{x,y}(Z), P_{x',y}(Z)\big)}{d(x, x')},$$

where $P_{x,y}(Z) := P_{W_2}(\varphi(x, y, Z))|\frac{\partial g}{\partial W_2}(x, y, \varphi(x, y, Z))|^{-1}$. This distribution depends only on realizations of (X, Y) and it is independent of $P_{X,Y}$. Hence, changing the dependency between X and Y will not affect $c^y_{z,x}$, which makes it a better candidate to measure the direct influences between variables of a network. As an illustration, we present a simple example. But first, we need the following result. All proofs are presented in the Appendix.

Theorem 1 *Consider* $\overline{X} = \mathbf{A}\overline{X} + \overline{W}$, *where* $\mathbf{A}$ *has zero diagonals and its support represents a DAG.* $\overline{W}$ *is a vector of zero mean independent random variables. Then,* $c^{Pa_i\setminus\{j\}}_{i,j} = |A_{i,j}|$.

Example 3 Consider a network of three variables $\{X, Y, Z\}$ in which $Y = aX + W_1$ and $Z = bX + cY + W_2$ for some non-zero coefficients $\{a, b, c\}$ and exogenous noises $\{W_1, W_2\}$. Hence,

$$\begin{aligned} I(X;Z|Y) &= H(Z|Y) - H(Z|X, Y) \\ &= H(bX + W_2|aX + W_1) - H(W_2). \end{aligned} \tag{10}$$

As we mentioned earlier, by reducing the variance of W_1, the first term in (10) tends to $H(bX + W_2|X) = H(W_2)$. Hence, (10) goes to zero. But, using the result of Theorem 1, we have $c^y_{z,x} = |b|$, which is independent of the variance of W_1.

4.3 Group selection for effective intervention

Consider a network of three variables $\{X, Y, C\}$ in which C is a common cause for X and Y, and X influences Y. In this network, to measure the influence of X on Y, one may consider $P(Y|do(X))$ that is given by $\sum_c P(Y|X, c)P(c) = \mathbb{E}_c[P(Y|X, c)]$. See, e.g., the back-door criterion in Pearl (2003). This conditional distribution is an average over all possible realizations of the common cause C.

Consider an experiment that is been conducted on a group of people with different ages C in which the goal is to identify the effect of a treatment X on a special disease Y. Suppose that this treatment has clearer effect on that disease for elderly people and less obvious effect for younger ones. In this case, averaging the effect of

the treatment on the disease for all people with different ages, i.e., $P(Y|do(X))$ might not reveal the true effect of the treatment. Hence, it is important to identify a regime (in this example age range) of C in which the influence of X on Y is maximized. As a consequence, we can identify the group of subjects on which the intervention is effective.

Note that this problem cannot be formalized using do-operation or other measures that take average over all possible realizations of C. However, using the measure in (3), we can formulate this problem as follows: given $X = x$ and two different realizations for C, say c and c', we obtain two conditional probabilities $P(Y|x, c)$ and $P(Y|x, c')$. Then, we say in group $C = c$, the causal influence between X and Y is more obvious compare to the group $C = c'$, if given $C = c$, changing the assignments of X leads to larger variation of the conditional probabilities compared to changing the assignment of X given $C = c'$. More precisely, if $c_{y,x}^{C=c} \geq c_{y,x}^{C=c'}$, where

$$c_{y,x}^{C=c} := \sup_{x \neq x'} \frac{W_d\Big(P(Y|x, c), P(Y|x', c)\Big)}{d(x, x')}. \tag{11}$$

Note that $c_{y,x}^{c} = \sup_c c_{y,x}^{C=c}$, where $c_{y,x}^{c}$ is given in (3). Using this new formulation, we define the range of C in which the influence from X to Y is maximized as $\arg\max_c c_{y,x}^{C=c}$.

Example 4 Suppose that $Y = CX + W_2$ and $X = W_1/C$, where C takes value from $\{1, ..., M\}$ w.p. $\{p_1, ..., p_M\}$ and $W_i \sim \mathcal{N}(0, 1)$. In this case, we have $c_{y,x}^{C=c} = |c|$. Thus, $C = M$ will show the influence of X on Y more clearer. On the other hand, such property cannot be detected using other measures. For example, we have

$I(X;Y|C = c) = 0.5 \log(2)$, for all c.

5 Properties of the measure

Lemma 1 *The measure defined in* (3) *possesses the following properties*: (1) *Asymmetry*: *In general* $c_{i,j}^{\mathcal{K}} \neq c_{j,i}^{\mathcal{K}}$. (2) $c_{i,j}^{\mathcal{K}} \geq 0$ *and when it is zero, we have* $X_i \perp\!\!\!\perp X_j | \underline{X}_{\mathcal{K}}$. (3) *Decomposition*: $c_{i,\{j,k\}}^{\mathcal{K}} = 0$ *implies* $c_{i,j}^{\mathcal{K}} = c_{i,k}^{\mathcal{K}} = 0$. (4) *Weak union*: *If* $c_{i,\{j,k\}}^{\mathcal{K}} = 0$, *then* $c_{i,j}^{\mathcal{K}\cup\{k\}} = c_{i,k}^{\mathcal{K}\cup\{j\}} = 0$. (5) *Contraction*: *If* $c_{i,j}^{\mathcal{K}} = c_{i,\mathcal{K}} = 0$, *then* $c_{i,\mathcal{K}\cup\{j\}} = 0$. (6) *Intersection*: *If* $c_{i,j}^{\mathcal{K}\cup\{k\}} = c_{i,k}^{\mathcal{K}\cup\{j\}} = 0$, *then* $c_{i,\{j,k\}}^{\mathcal{K}} = 0$.

Note that unlike the intersection property of the conditional independence, which does not always hold, the intersection property of the dependency measure in (3) always holds. This is due to the fact that (3) is defined for all realizations $(x_j, \underline{x}_{\mathcal{K}})$ not only those with positive measure. See Example 1 for the asymmetric property of $c_{i,j}^{\mathcal{K}}$.

We say a DAG possesses global Markov property with respect to (3) if for any node i and disjoint sets $\mathcal{B}$, and $\mathcal{C}$ for which i is d-separated from $\mathcal{B}$ by $\mathcal{C}$, we have $c^{\mathcal{C}}_{i,\mathcal{B}} = c^{\mathcal{C}}_{\mathcal{B},i} = 0$. Using the above Lemma and the results of Theorem 3.27 in Lauritzen (1996), it is straightforward to show that a faithful network of m random variables whose causal structure is a DAG possesses the global Markov property[5]. This property can be used to develop reconstruction algorithms (e.g., PC algorithm (Spirtes et al. 2000b)) for the causal structure of a network.

5.1 Estimation

The measure introduced in (3) can be computed explicitly for special probability measures. For instance, if the joint distribution of $\underline{X}$ is Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix Σ, then using the results of Givens and Michael (1984), we obtain $c^{\mathcal{K}}_{i,j} = |\Sigma_{i,\{j,\mathcal{K}\}}\left(\Sigma_{\{j,\mathcal{K}\},\{j,\mathcal{K}\}}\right)^{-1}\mathbf{e}_1|$, where $\Sigma_{i,\{j,\mathcal{K}\}}$ denotes the sub-matrix of Σ comprising row i and columns $\{j,\mathcal{K}\}$, and $\mathbf{e}_1 = (1,0,...,0)^T$. Hence, in such systems, one can estimate the dependency measure by estimating the covariance matrix. However, this is not the case in general. Therefore, we introduce a non-parametric method for estimating our dependency measure using kernel method.

Given $\{x^{(1)},...,x^{(N_1)}\}$ and $\{x^{(N_1+1)},...,x^{(N_1+N_2)}\}$ that are i.i.d. samples drawn randomly from ν_1 and ν_2, respectively, the estimator of (6) is given by Sriperumbudur et al. (2010),

$$\widehat{W}_d(\hat{\nu}_1,\hat{\nu}_2) := \max_{\{\alpha_i\}} \frac{1}{N_1}\sum_{i=1}^{N_1}\alpha_i - \frac{1}{N_2}\sum_{j=1}^{N_2}\alpha_{j+N_1}, \tag{12}$$

such that $|\alpha_i - \alpha_j| \le d(x^{(i)},x^{(j)})$, $\forall i,j$. In this equation, $\hat{\nu}_1$ and $\hat{\nu}_2$ are empirical estimator of ν_1 and ν_2, respectively. The estimator of MMD is given by

$$(\widehat{\mathrm{MMD}}(\hat{\nu}_1,\hat{\nu}_2))^2 := \sum_{i,j=1}^{N_1+N_2} y_i y_j k(x^{(i)},x^{(j)}), \tag{13}$$

where $y_i := 1/N_1$ for $i \le N_1$ and $y_i := -1/N_2$, elsewhere. $k(\cdot,\cdot)$ represents the kernel of $\mathcal{H}$. It is shown in Sriperumbudur et al. (2010) that (12) converges to (6) as $N_1, N_2 \to \infty$ almost surely as long as the underlying metric space is totally bounded. It is important to mention that the estimator in (12) depends on $\{x^{(j)}\}$s only through the metric $d(\cdot,\cdot)$, and thus its complexity is independent of the dimension of $x^{(i)}$, unlike the KL-divergence estimator (Qing et al. 2005). The estimator in (13) also converges to (7) almost surely with the rate of order $\mathcal{O}(1/\sqrt{N_1} + 1/\sqrt{N_2})$, when $k(\cdot,\cdot)$ is measurable and bounded.

Consider a network of m random variables $\underline{X}$. Given N i.i.d. realizations of $\underline{X}$, $\{\underline{z}^{(1)},...,\underline{z}^{(N)}\}$, where $\underline{z}^{(l)} \in E^m$, we use (12) and define

[5] See Appendix for more details.

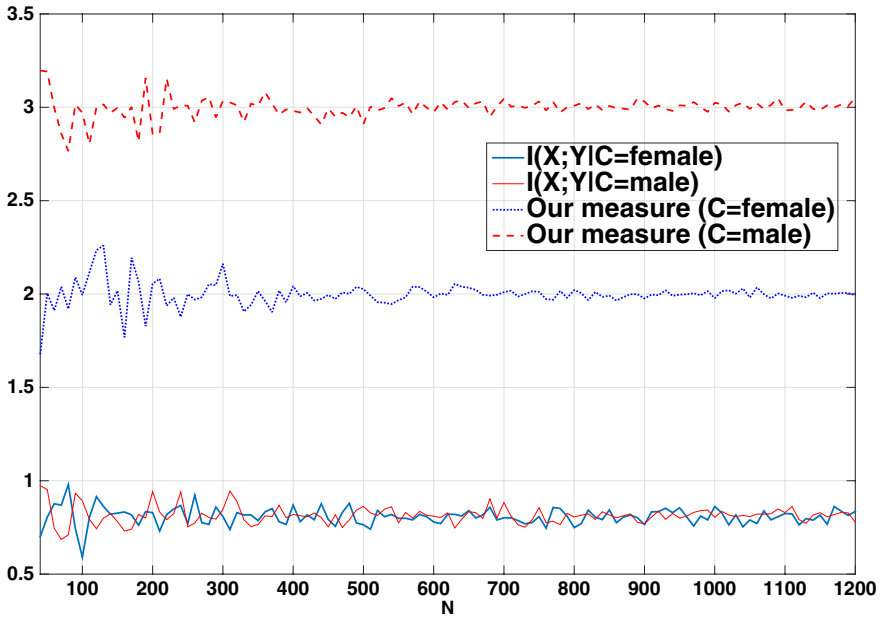

Fig. 2 Estimated measures for different N

$$\widehat{c}_{i,j}^{\mathcal{K}} := \max_{1\leq l,k\leq N} \frac{\widehat{W}_d\left(\hat{\mu}_i\left(\underline{z}_{\mathcal{K}\cup\{j\}}^{(l)}\right), \hat{\mu}_i\left(\underline{z}_{\mathcal{K}\cup\{j\}}^{(k)}\right)\right)}{d(z_j^{(l)}, z_j^{(k)})}, \tag{14}$$

such that $\underline{z}_{\mathcal{K}\cup\{j\}}^{(l)} = \underline{z}_{\mathcal{K}\cup\{j\}}^{(k)}$ off j. Similarly, one can introduce an estimator for $\tilde{c}_{i,j}^{\mathcal{K}}$ using (13). By applying the result of Corollary 5 in Spirtes et al. (2000a), we obtain the following result.

Corollary 1 *Let (E, d) be a totally bounded metric space and a network of random variables with positive probabilities, then $\widehat{c}_{i,j}^{\mathcal{K}}$ converges to $c_{i,j}^{\mathcal{K}}$ almost surely as N goes to infinity.*

6 Experimental results

Herein, we present two simulations in order to verify the theoretical results. In particular, the first experiment verifies the group selection advantages and the second one shows an application of the measure for capturing rare dependencies.

Group selection for: In this simulation, we considered a group of individuals ($C \in$ {male,female}) to study the effect of an special treatment X on their health condition Y. For instance, X can denote sleep aids and Y can represent the individual's

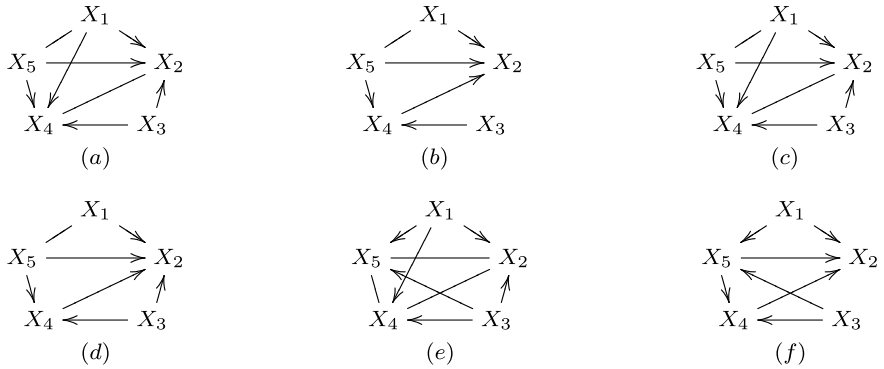

Fig. 3 Recovered DAGs of the system given in (15) for different sample sizes. **a**, **b** Use the measure in (3) and pure observation. **c**, **d** Use kernel-based method and pure observation. **e**, **f** Use the measure in (3) and interventional data. **f** Shows the true structure.

awareness level in the next morning. Most psychotropic drugs are metabolized in the liver. Because the male body breaks down Ambien and other sleep aids faster, women typically have more of the drug in their system the next morning. For this simulation, we considered a mathematical model between X, Y, and C as follows: $X = \mathcal{N}(1.5, 1)$ and $Y = 2X + \mathcal{N}(0, 1)$, when $C =$ female and $X = \mathcal{N}(1, 4)$ and $Y = 3X + \mathcal{N}(0, 9)$, otherwise.

Accordingly, we generated different sample sizes $N \in \{40, ..., 1200\}$ and estimated $I(X; Y|c)$ and $\hat{c}^c_{y,x}$. Figure 2 depicts the results. Since for given c, (X, Y) is jointly Gaussian, we estimated $I(X; Y|c)$ by estimating the covariance matrix Cover and Thomas (2012), and estimated our measure using (13) with Gaussian kernels. As Fig. 2 shows, although the treatment has different effects on different genders, $I(X; Y|C)$ cannot capture that.

Capturing rare dependencies: We simulated the following non-linear system with $W_i \sim U[-1, 1]$ and learned its corresponding structure.

$$\begin{aligned} &X_1 = W_1, \;\; X_2 = X_1^2 + 2X_4 - |X_5| + W_2, \\ &X_4 = X_3 - X_5 + W_4, \;\; X_3 = W_3, \\ &X_5 = W_5, \text{ if } X_3 \text{ is a natural number, } X_5 = 2\sqrt{|X_1|} + W_5, \text{ o.t.} \end{aligned} \tag{15}$$

In this example, the event that X_3 is a natural number occurs rarely since the measure of natural numbers is zero in $[-1, 1]$. We used the estimator of MMD given in (13) with Gaussian kernels and estimated the dependency measures. We obtained the corresponding DAG of this network given a set of observation of size $N \in \{900, 2500\}$. Using the results on the convergence rate of the MMD estimator, we used a threshold of order $\mathcal{O}(1/\sqrt{N})$ to distinguish positive and zero measure. Fig. 3 depicts the resulting DAGs. We also compared the performance of our measure with the kernel-based method proposed in Zhang et al. (2011). Note that in this example, since the influence of X_3 on X_5 is not detectable by mere observation, the best we can learn from mere observation is the DAG presented in Fig. 3b. This is due to the fact that the probability of X_3 being a natural number is zero and therefore, in

the observational data, we have $X_5 = 2\sqrt{|X_1|} + W_5$, almost surely. However, with the same number of observations, the kernel-based method identifies an extra edge, Fig. 3d.

Next, we fixed the value of X_3 to be natural number and irrational, separately and observed the outcome of the other variables for different sample sizes. Figure 3e, f depict the outcomes of the learning algorithm that uses our measure. In this case, $X_3 \to X_5$ was identified and then the Meek rules helped to detect all the directions even the direction of $X_1 - X_5$ as it is shown in Fig. 3f.

7 Conclusion

We studied several shortcomings of the existing dependency measures in detecting direct causal influences in the literature and introduced a new statistical dependency measure to overcome them. This measure is inspired by Dobrushin's coefficients and is based on the fact that there is no dependency between two variables if and only if the conditional distribution between them remains unchanged after assigning different realizations to the conditioning variable. We presented the advantages of this measure over the related measures. By establishing the connections between our measure and the integral probability metric (IPM), we developed low complexity estimators for our measure compared to other state-of-the-art relevant information-theoretic-based measures such as conditional mutual information.

Appendix

Preliminaries

Herein, we present additional information about Wasserstein (or Kantorovich) metric and IPM.

Definition 2 Let $(\mathbb{R}, d)$ be a metrical complete and separable space, and let $\mathcal{M}$ be the space of all probability measures on $\mathbb{R}$. If P and Q are the distribution functions of probability measures μ and $\nu \in \mathcal{M}$, respectively, the Kantorovich metric is defined by

$$d_K(\mu, \nu) := \int_{-\infty}^{\infty} |P(x) - Q(x)| dx = \int_0^1 |P^{-1}(t) - Q^{-1}(t)| dt.$$

For any separable metric space, this is equivalent to

$$d_K(\mu, \nu) := \sup \left\{ \left| \int h d\mu - \int h d\nu \right| : \ \forall h(x), \ s.t. \ |h(x) - h(y)| \leq d(x, y) \right\}.$$

By the Kantorovich–Rubinstein theorem, the Kantorovich metric is equal to the Wasserstein metric defined in 1. For an overview, see (Gibbs and Edward 2002).

Definition 3 Let (E, d) be a metrical complete and separable space. The integral probability metrics (IPM) between two measure μ and ν is defined by

$$\text{IPM}(\mu, \nu) := \sup_{f \in \mathcal{F}} \left| \int_E f d\nu_1 - \int_E f d\nu_2 \right|,$$

where $\mathcal{F}$ is a class of real-valued bounded measurable functions on E.

The choice of function class $\mathcal{F}$ determines the type of IPM metric. For instance, if $\mathcal{F}$ is set to be the class of all continuous Lipschitz functions, then it becomes the Wasserstein metric. For MMD, $E = \mathcal{H}$ is a RKHS and $\mathcal{F} = \mathcal{F}_{\mathcal{H}} := \{f : ||f||_{\mathcal{H}} \leq 1\}$. When $\mathcal{F}$ is set to be $\{f : ||f||_{\infty} \leq 1\}$, we obtain the total variation metric. For further details, see (Sriperumbudur et al. 2010) and the references within.

Proof of Lemma 1

• $c_{i,j}^{\mathcal{K}} \geq 0$ since Wasserstein is a metric. If $c_{i,j}^{\mathcal{K}} = 0$, we have

$$W_d\big(P(X_i|x_j, \underline{x}_{\mathcal{K}}), P(X_i|y_j, \underline{x}_{\mathcal{K}})\big) = 0,$$

for all realizations x_j, y_j and $\underline{x}_{\mathcal{K}}$. Using the fact that Wasserstein is a metric on the space of probability measures, the above equality, and total probability law, we obtain

$$\begin{aligned} P(X_i|\underline{x}_{\mathcal{K}}) &= \sum_{x_j} P(X_i|x_j, \underline{x}_{\mathcal{K}}) P(x_j|\underline{x}_{\mathcal{K}}) \\ &= P(X_i|y_j, \underline{x}_{\mathcal{K}}) \sum_{x_j} P(x_j|\underline{x}_{\mathcal{K}}) = P(X_i|y_j, \underline{x}_{\mathcal{K}}). \end{aligned}$$

The above equality holds for all y_j and $\underline{x}_{\mathcal{K}}$. This implies $X_i \perp\!\!\!\perp X_j | \underline{X}_{\mathcal{K}}$.

• We show this by an example. Let $X = U_{[0,1]}$ to be uniformly distributed between zero and one, and

$$Y = \begin{cases} V_{[0,1]} & \text{if } X \in \mathcal{A}, \\ U_{[0,1]} & \text{otherwise}, \end{cases}$$

where $\mathcal{A} = \{\frac{i}{i+1} : i \in \mathbb{N}\}$, and $V_{[0,1]}$ is a random variable independent of U that is distributed non-uniformly over [0, 1]. In this case, we have

$$c_{y,x} \geq \frac{W_d(P(Y|X=1/2), P(Y|X=\sqrt{2}))}{d(1/2, \sqrt{2})} > 0.$$

On the other hand, it is easy to see that Y has a uniform distribution over [0, 1] almost surely. Furthermore, for two measurable sets C and B in the σ-algebra, we have

$$P(X \in C|Y \in B) = \frac{P(Y \in B|X \in C)P(X \in C)}{P(Y \in B)} =$$
$$\frac{P(Y \in B|X \in C \cap \mathcal{A})P(X \in C \cap \mathcal{A}) + P(Y \in B|X \in C \setminus \mathcal{A})P(X \in C \setminus \mathcal{A})}{P(Y \in B)}$$
$$= \frac{P(Y \in B|X \in C \setminus \mathcal{A})P(X \in C \setminus \mathcal{A})}{P(Y \in B)} = P(X \in C \setminus \mathcal{A}).$$

The last equality uses the fact that $P(Y \in B) = P(Y \in B|X \notin \mathcal{A}) = P(Y \in B|X \in C \setminus \mathcal{A})$. Thus, changing the value of Y will not affect the conditional distribution of X given Y, i.e., $c_{x,y} = 0$.

• If $c^{\mathcal{K}}_{i,\{j,k\}} = 0$, $W_d(P(X_i|x_j, x_k, \underline{x}_{\mathcal{K}}), P(X_i|y_j, y_k, \underline{x}_{\mathcal{K}})) = 0$, for all realization $x_j, y_j, x_k, y_k, \underline{x}_{\mathcal{K}}$. By the total probability law, we obtain

$$P(X_i|x_k, \underline{x}_{\mathcal{K}}) = \sum_{x_j} P(X_i|x_j, x_k, \underline{x}_{\mathcal{K}})P(x_j|x_k, \underline{x}_{\mathcal{K}})$$
$$= P(X_i|y_j, y_k, \underline{x}_{\mathcal{K}}) \sum_{x_j} P(x_j|x_k, \underline{x}_{\mathcal{K}}) = P(X_i|y_j, y_k, \underline{x}_{\mathcal{K}}).$$

This implies that $P(X_i|x_k, \underline{x}_{\mathcal{K}}) = P(X_i|y_j, y_k, \underline{x}_{\mathcal{K}}) = P(X_i|y_k, \underline{x}_{\mathcal{K}})$. Hence, $c^{\mathcal{K}}_{i,k} = 0$. Similarly, we can prove that $c^{\mathcal{K}}_{i,j} = 0$.

• Suppose $c^{\mathcal{K}}_{i,\{j,k\}} = 0$, then from the previous proof, we have $P(X_i|x_k, \underline{x}_{\mathcal{K}}) = P(X_i|y_k, y_j, \underline{x}_{\mathcal{K}})$, for all realizations $y_j, x_k, y_k, \underline{x}_{\mathcal{K}}$. Thus,

$$P(X_i|x_k, \underline{x}_{\mathcal{K}}) = P(X_i|y_k, x_j, \underline{x}_{\mathcal{K}})$$

This is equivalent to say $c^{\mathcal{K}\cup\{j\}}_{i,k} = 0$. The other part can be shown similarly.

• If $c^{\mathcal{K}}_{i,j} = c_{i,\mathcal{K}} = 0$, then from $c^{\mathcal{K}}_{i,j} = 0$ and total probability law, we obtain that

$$W_d(P(X_i|x_j, \underline{x}_{\mathcal{K}}), P(X_i|\underline{x}_{\mathcal{K}})) = 0. \tag{16}$$

On the other hand, using the triangle inequality of the Wasserstein metric, we have

$$W_d(P(X_i|x_j, \underline{x}_{\mathcal{K}}), P(X_i|y_j, \underline{y}_{\mathcal{K}})) \leq$$
$$W_d(P(X_i|x_j, \underline{x}_{\mathcal{K}}), P(X_i|\underline{x}_{\mathcal{K}})) + W_d(P(X_i|\underline{x}_{\mathcal{K}}), P(X_i|\underline{y}_{\mathcal{K}}))$$
$$+ W_d(P(X_i|\underline{y}_{\mathcal{K}}), P(X_i|y_j, \underline{y}_{\mathcal{K}})).$$

The first and third expressions on the right-hand side are zero due to (16) and the second expression is zero due to $c_{i,\mathcal{K}} = 0$.

- If $c_{i,j}^{\mathcal{K}\cup\{k\}} = 0$, $W_d(P(X_i|x_j, x_k, \underline{x}_{\mathcal{K}}), P(X_i|y_j, x_k, \underline{x}_{\mathcal{K}})) = 0$. This implies that $P(X_i|x_j, x_k, \underline{x}_{\mathcal{K}}) = P(X_i|x_k, \underline{x}_{\mathcal{K}})$ for all realizations x_j, x_k, and $\underline{x}_{\mathcal{K}}$. Similarly, because of $c_{i,k}^{\mathcal{K}\cup\{j\}} = 0$, we have $P(X_i|x_j, x_k, \underline{x}_{\mathcal{K}}) = P(X_i|x_j, \underline{x}_{\mathcal{K}})$ for all realizations x_j, x_k, and $\underline{x}_{\mathcal{K}}$. Hence, for all realizations, we have $P(X_i|x_j, \underline{x}_{\mathcal{K}}) = P(X_i|x_k, \underline{x}_{\mathcal{K}})$. This result and the total probability law will establish the result.

The global Markov property

Since the influence structure of this network is a DAG, there exists an ordering of the variables such that for every node i, all its parents have indices less that i. Without loss of generality suppose that $\{X_1, ..., X_m\}$ is that ordering. Furthermore, using the chain rule, we have

$$P(\underline{X}) = \prod_{i=1}^{m} P(X_i|\underline{X}_{\{<i\}}), \tag{17}$$

where $\underline{X}_{\{<i\}}$ denotes all the variables with indices less than i. Due to the nature of this ordering, all the nodes in $\{<i\}$ that do not belong to Pa_i are non-descendants of node i. Hence, by the definition of ID, they have zero influence on X_i given the parents of i and because of the first property in Lemma 1, they can be dropped from the conditioning in (17).

The global Markov property is a direct consequence of Lemma 1 and Theorem 3.27 in Lauritzen (1996).

Proof of theorem 1

To complete the proof, we need the following technical lemmas. When $d(\cdot, \cdot)$ is the Euclidean distance, we denote the Wasserstein metric by $W_E(\cdot, \cdot)$.

Lemma 2 *For real-valued random variables, we have*

$$\begin{aligned} \left|\mathbb{E}_{\nu_1}[x] - \mathbb{E}_{\nu_2}[y]\right| \quad &\leq \qquad W_E(\nu_1, \nu_2) \\ &\leq \sqrt{\mathbb{E}_{\nu_1}[x^2] + \mathbb{E}_{\nu_2}[y^2] - 2\mathbb{E}_{\pi}[xy]}, \end{aligned} \tag{18}$$

where π is any joint distribution of x and y such that its marginals are ν_1 and ν_2.

Proof The lower bound is due to the dual representation of the Wasserstein metric and the fact that $f(x) = x$ is Lipschitz.

For the upper bound, we use the Jensen's inequality, that is

$$W_d(\nu_1, \nu_2) \leq \inf_{\pi} \left(\mathbb{E}_{\pi}[d^p(x, y)]\right)^{1/p}, \tag{19}$$

for $p \geq 1$. For $p = 2$, we use the monotonicity of $\sqrt{x}$, and the fact that the space of probability measures is complete and obtain the result. □

Consider a network of variables in which every variable X_i functionally depends on a subset of other variables $\underline{X}_{Fp_i}$ (the parent set of node i) as follows,

$$X_i = F_i(\underline{X}_{Fp_i}) + G_i(\underline{X}_{Fp_i})W_i, \quad \forall i, \tag{20}$$

where F_i, G_i are arbitrary functions such that $G_i \neq 0$. $\{W_i\}$s denote exogenous noises with mean zero.

Lemma 3 *For a system described by* (20)*, the influence of node j on its child i given the rest of i's parents* $Fp_i \setminus \{j\}$ *under Euclidean metric, is bounded as follows*

$$\begin{aligned} \sup_{\substack{\overline{x}_{Fp_i} = \overline{y}_{Fp_i} \\ \text{off } j}} \left| \frac{F_i(\overline{x}_{Fp_i}) - F_i(\overline{y}_{Fp_i})}{x - y} \right| \leq c_{i,j}^{Fp_i \setminus \{j\}} \leq \sup_{\substack{\overline{x}_{Fp_i} = \overline{y}_{Fp_i} \\ \text{off } j}} \\ \left[\left(\frac{F_i(\overline{x}_{Fp_i}) - F_i(\overline{y}_{Fp_i})}{x - y} \right)^2 + \left(\frac{G_i(\overline{x}_{Fp_i}) - G_i(\overline{y}_{Fp_i})}{x - y} \sigma_i \right)^2 \right]^{1/2}, \end{aligned} \tag{21}$$

where the supremum is taking over all realizations of $\underline{X}_{-\{i\}}$ *that are only different at* X_j*.*

Proof Using the lower bound in Lemma 2 and the fact that W_is have zero mean, we obtain the lower bound in (21).

To obtain the upper bound, we again use the result of Lemma 2, with the following joint distribution $\pi(X_i, Y_i)$,

$$\frac{1}{|G_i(\overline{x}_{Fp_i})|} f_{W_i}\left(\Theta_{\overline{x}_{Fp_i}}(X_i)\right) \mathbb{I}_{\{\Theta_{\overline{x}_{Fp_i}}(X_i) = \Theta_{\overline{y}_{Fp_i}}(Y_i)\}},$$

where

$$\Theta_{\overline{x}_{Fp_i}}(X_i) := \frac{X_i - F_i(\overline{x}_{Fp_i})}{G_i(\overline{x}_{Fp_i})},$$

and f_{W_i} denotes the probability density function of W_i and $\mathbb{I}$ denotes the indicator function. Using this joint distribution, we obtain the upper bound in (21). □

Applying the above result to a linear system in which $F_i(\overline{y}_{Fp_i}) = (\mathbf{A}\overline{x})_i$ and $G_i(\overline{x}_{Fp_i}) = 1$, we obtain that $c_{i,j}^{Fp_i \setminus \{j\}} = |A_{i,j}|$.

Funding Open access funding provided by EPFL Lausanne. This work was supported by ONR grant W911NF-15-1-0479 and SNSF grant 200021-20435.

Declarations

Conflict of interest Kun Zhang and Negar Kiyavash were two invited speakers at International Symposium on Causal Inference and Machine Learning, September 10-11, 2021. The authors have no other affliations.

References

Alfred Müller (1997) Integral probability metrics and their generating classes of functions. Advances in applied probability. Springer, Berlin, pp 429–443

Aronszajn N (1950) Theory of reproducing kernels. Trans Am Math Soc 68(3):337–404

Arthur G, Karsten MB, Malte R, Bernhard S, Smola AJ (2006) A kernel method for the two-sample-problem. Advances in neural information processing systems. Springer, Berlin, pp 513–520

Arthur G, Kenji F, Choon HT, Le S, Bernhard S, Smola AJ (2007) A kernel statistical test of independence. Advances in neural information processing systems. Springer, Berlin, pp 585–592

Ay N, Polani D (2008) Information flows in causal networks. Adv Complex Syst 11(01):17–41

Cover MT, Thomas AJ (2012) Elements of information theory. John Wiley & Sons, New Jersey

Der Vaart AW, Van Wellner JA (1996) Weak convergence. Springer, Berlin

Dobrushin LR (1970) Prescribing a system of random variables by conditional distributions. Theory Probab Appl 15(3):458–486

Dudley MR (2002) Real analysis and probability, 74th edn. Cambridge University Press, Cambridge

Fukumizu K, Gretton A, Sun X, Schölkopf B (2007) Kernel measures of conditional dependence. NIPS 20:489–496

Gibbs LA, Edward SF (2002) On choosing and bounding probability metrics. Int Stat Rev 70(3):419–435

Givens RC, Michael SR et al (1984) A class of Wasserstein metrics for probability distributions. Michigan Math J 31(2):231–240

Gorodetskii VV (1978) On the strong mixing property for linear sequences. Theory Probab Appl 22(2):411–413

Janzing D, Balduzzi D, Grosse-Wentrup M, Schölkopf B et al (2013) Quantifying causal influences. Ann Stat 41(5):2324–2358

Judea P (2014) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann, Burlington

Lauritzen LS (1996) Graphical models. Oxford University Press, Oxford

Massey J (1990) Causality, feedback and directed information. In: Proceedings of International Symposium Information Theory Applications(ISITA-90), Citeseer, pp 303–305

Meek C (1995) Strong completeness and faithfulness in bayesian networks. Rroceedings of the Eleventh conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc, Burlington, pp 411–418

Micchelli AC, Yuesheng X, Haizhang Z (2006) Universal kernels. J Mach Learn Res 7:2651–2667

Nihat A, Krakauer CD (2007) Geometric robustness theory and biological networks. Theory Biosci 125(2):93–121

Pearl J (2003) Causality: models, reasoning, and inference. Econom Theor 19:675–685

Qing W, Kulkarni Sanjeev R, Sergio V (2005) Divergence estimation of continuous distributions based on data-dependent partitions. IEEE Trans Inf Theory 51(9):3064–3074

Schreiber T (2000) Measuring information transfer. Phys Rev Lett 85(2):461

Singh M, Valtorta M (1995) Construction of Bayesian network structures from data: a brief survey and an efficient algorithm. Int J Approx Reason 12(2):111–131

Spirtes P, Glymour C, Scheines R, Kauffman S, Aimale V, Wimberly F (2000a) Constructing Bayesian network models of gene expression networks from microarray data

Spirtes P, Glymour CN, Scheines R (2000b) Causation, prediction, and search, 81st edn. MIT Press, London

Sriperumbudur BK, Fukumizu K, Gretton A, Schölkopf B, Lanckriet G (2010) Non-parametric estimation of integral probability metrics. In: Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on IEEE, pp 1428–1432

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, ACM, pp 855–862

Villani C (2021) Topics in optimal transportation (American Mathematical Sc., vol. 58)

Zhang K, Peters J, Janzing D, Schölkopf B (2011) Kernel-based conditional independence test and application in causal discovery. AUAI Press, Corvallis, pp 804–813