Network Inference using Directed Information: The Deterministic Limit

Arman Rahimzamani and Sreeram Kannan
Department of Electrical Engineering, University of Washington, Seattle, WA
Email: {armanrz, ksreeram}@uw.edu

***Abstract*—Consider a network comprised of many variables, which interact with each other following a causal stochastic dynamical system model. Given time-series measurements of these variables, an important task is the inference of the structure of the causal dynamical system. Recently, directed information has been proposed as a generalization of Granger causality to accurately infer the structure of the network. We observe a curious phenomenon: in the limit that the relationships become purely deterministic, this measure loses power completely. In this paper, we study this phenomenon, explore its connection to Taken's delay embedding theorem, propose a remedy called restricted directed information (RDI); and finally, demonstrate its efficacy in simulations. In particular we show that, RDI recovers the graph correctly in all instances, deterministic or stochastic, where there is enough information to recover the graph uniquely.**

I. Introduction

Determining the network of interactions between random variables from observations has a long history in graphical models [1]. With the emergence of pervasive measurement techniques in different domains, including in finance, social networks, computational biology, and smart cities, it has become increasingly commonplace to observe time-series of measurements. These time-series are usually well modeled by random processes that interact and evolve; and a common problem of interest in many domains is the inference of the underlying network structure of interactions from the observed time-series. It is for performing this particular task in financial time series that Granger proposed a notion called Granger causality [2], which can infer the network under an assumption of linear relationships among the time-series. In a series of recent work [3], [4], methods utilizing directed information, originally proposed in communications theory literature [5], have been proposed as the appropriate extension to non-linear problems in order to infer the network correctly. The directed information between a pair of stochastic processes is given as,

$$\mathrm{DI}(X \to Y|Z) := \frac{1}{T}\sum_{t=1}^{T}\{H(Y_t|Y_{t-1},...,Y_1,Z_{t-1},...,Z_1) \\ -H(Y_t|Y_{t-1},...,Y_1,X_{t-1},...,X_1,Z_{t-1},...,Z_1)\}.$$

In [4], it is shown that there is a causal link from X_i to X_j if and only if $\mathrm{DI}(X_i \to X_j|X_{\{i,j\}^c}) > 0$, provided all transition probabilities are non-zero and there are no latent variables.

A particularly simple case of such problems occurs when the evolution of the time-series is purely deterministic; and this is indeed the scope of study in the present paper. For example, consider a popular non-linear dynamical system, referred to as the Lorenz system [6]. It is a system with 3 time series, x, y, z which evolves according to the following ordinary differential equation with parameters σ, ρ, β.

$$\begin{aligned} \dot{x}(t) &= \sigma(y(t) - x(t)) && (1) \\ \dot{y}(t) &= x(t)(\rho - z(t)) - y(t) && (2) \\ \dot{x}(t) &= x(t)y(t) - \beta z(t) && (3) \end{aligned}$$

This equation system is particularly interesting for some parameter regimes, for example $\sigma = 10, \beta = \frac{8}{3}, \rho = 28$, in the sense that it has no fixed points or orbits of fixed period; rather the system exhibits a "strange" attractor inside which the system remains. One may wish to infer the system's connectivity, shown in Figure 1, from the observations of the time series. It is pointed out in a recent work on inferring causality in ecological systems [7] that for inferring the connectivity of such systems, Granger causality and its extensions cannot be satisfactorily employed. This is, because, in the case of dynamical systems, Taken's theorem [8] guarantees that the information about the state of a system $(x(t), y(t), z(t))$ is contained already in the lags of any one of its variables, for example, $x(t-1), x(t-2), x(t-3)$. Indeed in that case,

$$\begin{aligned} \mathrm{DI}(Y \to X) \quad := \quad & \frac{1}{T}\sum_{t=1}^{T} H(X_t|X_{t-1},...,X_1) \\ & -H(X_t|Y_{t-1},...,Y_1,X_{t-1},...,X_1) = 0. \end{aligned}$$

Therefore all directed information terms may be zero, and indeed the system is not inferable using directed-information graphs. In this regime, [7] proposed some algorithms which can infer the underlying graph using lagged embeddings. These algorithms are guaranteed to return the correct answer when the system has only two interacting variables and when the system is purely deterministic. In this paper, we try to bridge the gap between the two extremes: in the purely stochastic extreme, directed-information graphs seem to work well and in the purely deterministic setting, alternative methods seem to be needed. This problem is also interesting because as the SNR increases, i.e., the noise in the system decreases, one may intuitively expect better performance but directed-information based methods deteriorate (see Figure 3). The question that we study is the following: is there a universal algorithm that can recover the graph correctly in both the deterministic as well as stochastic settings?

Consider a system that is undergoing a deterministic evolution; in such a case directed information is not even well defined. In order to make this quantity well-defined, let us assume that the initial conditions are random so that there is some randomness in the system. As already discussed, directed information does not infer this graph correctly. We propose a natural metric called the restricted directed information, which is the same as the directed information but only takes into account the past time-step.

$$\begin{aligned} RDI(X \to Y) &= H(Y(t)|Y(t-1)) \\ &\quad -H(Y(t)|Y(t-1), X(t-1)). \end{aligned} \quad (4)$$

Clearly, such a metric is tailored to first-order Markov relationships; and one cannot hope to infer higher order Markov relationships using such a metric (unlike directed information). However, under the first-order Markov assumption, this quantity can be estimated with far fewer samples, lending practical viability to such an estimator. In this paper, we do not dwell on the estimators, except to point out that estimation is an interesting task and needs some thought. In our implementations, we have adapted estimators based on the KSG estimator [9], whose properties were analyzed and generalized in some recent work [10], [11]. The proposed method of using RDI needs to estimate a conditional mutual information depending on a m-dimensional joint distribution (where m is the number of nodes). There are methods that can utilize additional assumptions so that the sample complexity can be reduced further [12].

The question that we would like to answer is when can RDI infer the causal graph of a system correctly. It turns out that if thribue system indeed had non-trivial noise at all the nodes, it is easy to show that under the first order Markov assumption the metric infers the graph correctly. In this paper, therefore, the focus is on the deterministic or semi-deterministic cases, where only some nodes may have noise added to them. Our main result in the paper is that in the deterministic case, RDI infers the graph correctly in cases where there is enough information to do so (see Theorem III.1, Theorem III.3 and Theorem IV.1). We also characterize the minimum number of nodes in which noise is present for a linear system to be inferred correctly; and show that RDI is optimal in this case as well (see Theorem III.10).

II. Assumptions, Definitions and Notations

A dynamical system is a system of m random processes $\{X_i(t)\}_{1 \leq i \leq m}$ for $t = 1, 2, .., T$. At each time t, $X_i(t)$ is a random variable taking values in $\mathcal{X}_i$ and we denote them altogether by the $m \times 1$ vector $\underline{X}(t)$. A specific realization of each random variable at time t is denoted by $x_i(t)$ for $i \in [m]$ and the whole *state* of the system at time t is shown by $\underline{x}(t)$. We define $[m] := \{1 \leq i \leq m\}$.

The dynamical system starts evolving from an initial state $\underline{X}(0)$ which we assume is chosen randomly according to a distribution $P_{\underline{X}}(0)$. The probability distribution at time t is then denoted as $P_{\underline{X}}(t)$. The mechanism of the system's evolution is defined by a *vector transition function* at each time moment, i.e. for each $t > 0$, $\underline{X}(t) = g\left(\underline{X}(t-1), \underline{N}(t)\right)$ in which g is the transition function and $\underline{N}$ is an $m \times 1$ vector of mutually-independent random variables called *noise vector at time t* which is independent of the past states of the system. Therefore the system is first-order Markov. Note that g does not depend on the time index, hence the Markov chain is time-homogeneous.

Furthermore we assume that each variable $X_i(t)$ is a function of only its corresponding noise element. In other words, if $g = [g_1, g_2, \ldots, g_m]^T$, then $X_i(t) = g_i\left(\underline{X}(t-1), N_i(t)\right)$. Note that both the transition functions g_i and the initial state $P_{\underline{X}}(0)$ are required to fully specify the distribution of the random process. We will assume that the system is stationary, i.e., $P_{\underline{X}}(t) = P_{\underline{X}}(0)$,

Causality relationship and causality graph of a dynamical system:

We expressed the state of each variable X_j as $X_j(t) = g_j\left(\underline{X}(t-1), N_j(t)\right)$. But in general, g_j might be not a function of all the variables $X_i(t-1)$. In particular, it may depend only on a subset of variables, denoted by a set $Pa(X_j)$. So for each variable X_j we introduce a set of parent nodes denoted by $Pa(X_j) \subset [m]$ which influence X_j, in other words $X_j(t)$ is a function of $X_i(t-1)$ for $\forall X_i \in Pa(X_j)$. We say there is a causal relationship from X_i to X_j or "X_i causes X_j" if $X_i \in Pa(X_j)$. We can also encapsulate these causal relationships on a graph, with nodes being random processes and edges depending on the causal relationship. The directed graph $G_C = (V, E)$ is called the causality graph of a dynamical system, in which $V = [m]$ and $(i, j) \in E$ iff $X_i \in Pa(X_j)$.

The goal of network inference is to infer the casual graph of the system from observations of the time series $\underline{x}(0), ..., \underline{x}(t)$. In some contexts, one may observe multiple different runs of the random process, which may provide additional information.

Restricted Directed Information:

The restricted directed information (RDI) from the random process X to Y is defined as follows.

$$RDI\left(X \to Y\right) = I\left(X(t-1); Y(t) | Y(t-1)\right) \quad (5)$$

We study only stationary dynamical systems and the associated random processes X_i so that for all $t > 0$, $I\left(X(t-1), Y(t)|Y(t-1)\right)$ is unique and hence $RDI\left(X \to Y\right)$ is meaningful.

We can also define the conditional RDI similarly. The restricted directed information from X to Y given Z is defined as:

$$RDI\left(X \to Y|Z\right) = I\left(X(t-1); Y(t)|Y(t-1), Z(t-1)\right) \quad (6)$$

We also define an RDI graph. The directed graph $G_{RDI} = (V, E)$ is called the RDI graph of a dynamical system, in which $V = [m]$ and $(i, j) \in E$ iff $RDI\left(X_i \to X_j | \{X_k\}_{k \in [m]-\{i,j\}}\right) > 0$.

We overload the notation $H(X)$ to denote the entropy of a discrete variable X or to denote the differential entropy of an absolutely continuous real-valued random variable.

If the system is ergodic, then it is possible to estimate RDI and therefore the RDI graph from a single run of the time-series. In cases where the system is stationary but not ergodic, then multiple runs of the dynamical system with independent initializations are needed in order to estimate the RDI.

Throughout the rest of the paper, we are interested in finding out the casual relationships between the variables by using RDI. In other words, for different scenarios we try to find conditions under which $G_C = G_{RDI}$. The rest of the paper is organized as follows: we study linear systems with and without noise in Section III. We study deterministic non-linear systems with discrete alphabet in Section IV. Finally we demonstrate the performance of our proposed algorithms in simulations in Section V.

III. Linear Systems

In general, a linear dynamical system is a dynamical system in which all the alphabets $\mathcal{X}_i$ are real lines, and the functions g_i are linear. Equivalently, a linear system can be described as $\underline{X}(t) = A\underline{X}(t-1) + \underline{N}(t)$ in which A is an $m \times m$ matrix and $\underline{X}(t)$ and $\underline{N}(t)$ are $m \times 1$ vectors, expressing the state of the system and the additive noise respectively. We assume that $\underline{N}(t)$ follows a Gaussian distribution $\mathcal{N}(0, \Sigma_N)$. Note that $A_{ji} \neq 0 \iff (i,j) \in E$, where E is the edge-set of G_C, the casual graph of the dynamical system.

A. Deterministic linear systems

A linear deterministic dynamical system is described as $\underline{X}(t) = A\underline{X}(t-1)$, i.e, the noise variance is zero. Thus one can accurately predict the system states for all $t > 0$ if the initial system state $\underline{X}(0)$ and the matrix A are known. We think of such deterministic systems initialized with a random state $P_{\underline{X}}(0) = \mathcal{N}(0, \Sigma_{X(0)})$. Then the system evolves as a Gaussian random process with $P_{\underline{X}}(t) = \mathcal{N}(0, \Sigma_{X(t)})$.

As discussed before, the RDI is meaningful only when the system is stationary. For a general system to have a stationary distribution, it is necessary that the equation $\Sigma_X(t) = \Sigma_X(t-1)$ has a non-trivial answer. If the system is linear and deterministic with the transition matrix A, this condition is reduced to $\Sigma_X = A\Sigma_X A^T$. Furthermore, Σ_X will be identity if and only if A is orthonormal. Note that, even when A is orthonormal, there are other solutions to the equation as well, for example $\Sigma_X = 0$ is a valid solution; which stationary distribution the system will converge to depends on the initial distribution.

Now for the linear determinsitic system described above, we're looking for the conditions under which the RDI method will return the correct causality graph.

Definition III.1. *A graph G_C and stationary covariance Σ_X satisfy* Property A *if for all edges (i,j) in G_C we have*

$$\forall \underline{u} \in \mathbb{R}^m, \quad u_i \neq 0 : \quad \underline{u}^T \Sigma_X \underline{u} \neq 0. \tag{7}$$

We now present an examples where Property A is satisfied: if G_C is arbitrary, but Σ_X is positive definite, then Property A is satisfied. We note that if G_C is such that each node has an outgoing edge; then Property A is equivalent to Σ_X being positive definite.

Theorem III.1. (a) *If for a linear deterministic system property A is satisfied, then $G_{RDI} = G_C$.*

(b) *If Property A is not satisfied, then there will be an ambiguity in the correct system and no method will be able to return the causality graph correctly.*

We will prove this theorem in the rest of this section. Lemma III.2 suggests a simpler condition under which the RDI graph will be the same as causality graph. The proof of Part (a) of Theorem III.1 is identical to proof of that lemma; and follows from observing that Property A is all that is needed for the proof to go through.

Before that, the Prop. III.1 is stated without proof, which will be used through the proof of III.2.

Proposition III.1. *Let Σ_X be a covariance matrix of m zero-mean jointly Gaussian random variables denoted by $\underline{X}$. Then $\exists \quad \underline{u} \neq \underline{0} \quad \forall \underline{x} : \underline{u}^T(\underline{x}) = 0$ if and only if Σ_X is not positive definite.*

Lemma III.2. *If the covariance matrix of the linear deterministic system Σ_x is positive definite, then $G_{RDI} = G_C$.*

Proof. In a linear deterministic system, every variable X_j at each time t can be described as $X_j(t) = \sum_u A_{j,u} X_u(t-1)$. Let us assume $G_C = (V, E)$ is the causality graph of the system. it can be observed that $(i,j) \in E$ if $A_{j,i} \neq 0$, and $(i,j) \notin E$ otherwise.

We will show that given Σ_X is positive definite, for each pair (i,j) if $A_{j,i} = 0$ then $RDI\left(X_i \to X_j | \{X_u\}_{u \in [m]-\{i,j\}}\right) = 0$. Similarly, $RDI(X_i \to X_j | \{X_u\}_{u \in [m]-\{i,j\}}) > 0$ if $A_{j,i} \neq 0$.

We can write:

$$\begin{aligned} &RDI\left(X_i \to X_j | \{X_u\}_{u \in [m]-\{i,j\}}\right) \\ &= I(X_i(t-1); X_j(t) | \{X_u(t-1)\}_{u \in [m]-\{i\}}) \\ &= I(X_i(t-1); A_{j,i} X_i(t-1) | \{X_u(t-1)\}_{u \in [m]-\{i\}}) \end{aligned} \tag{8}$$

If $A_{j,i} = 0$, then $RDI\left(X_i \to X_j | \{X_u\}_{u \in [m]-\{i,j\}}\right) = 0$ yielding the lemma.

If $A_{j,i} \neq 0$, then

$$I(X_i(t-1); X_i(t-1) | \{X_u(t-1)\}_{u \in [m]-\{i\}}) = 0, \tag{9}$$

implies that $X_i(t-1)$ is a function of $\{X_u(t-1)\}_{u \in [m]-\{i\}}$, i.e. there exists a linear function l_i such that $\forall t : X_i(t) = l_i\left(\{X_u(t)\}_{u \in [m]-\{i\}}\right)$ which in turn implies $\exists u \neq 0 : u^T x(t-1) = 0$, so from Prop.III.1 we conclude that Σ_x is not positive definite which contradicts our assumption. Thus $RDI\left(X_i \to X_j | \{X_u\}_{u \in [m]-\{i,j\}}\right) > 0$ which yields the lemma. □

So we see for a deterministic linear system, if the Σ_x is positive definite, then RDI can return the correct causality graph. As pointed out earlier, only Property A is needed for the proof of this Lemma and that concludes the proof of part (a) of Theorem III.1.

We now need to prove (b) of Theorem III.1 which shows that if property A is not true, then there is no method that can infer the correct causality graph. $\exists \underline{u}, u_{i_0} \neq 0 : \underline{u}^T \underline{X} = 0$ can be written as $\forall t : X_{i_0}(t) = \sum_{k \in [m]-\{i_0\}} (-u_k) X_k(t)$. From the dynamics of the system and knowing that the edge $(i_0, j_0) \in E$, i.e. $A_{j_0,i_0} \neq 0$, we can write:

$$\begin{aligned} X_j(t) &= \sum_{k \in [m]} A_{j,k} X_k(t-1) \\ &= \sum_{k \in [m]-\{i_0\}} A_{j,k} X_k(t-1) + A_{j,i_0} X_{i_0}(t-1) \\ &= \sum_{k \in [m]-\{i_0\}} \left(A_{j,k} - u_k\right) X_k(t-1) \\ &= \sum_{k \in [m]-\{i_0\}} \tilde{A}_{j,k} X_k(t-1) \end{aligned} \tag{10}$$

So there is an alternative system $\tilde{A} \neq A$ which gives us the same dynamics for the system. So there is an ambiguity in the system and no method will be able to return the correct graph. This completes the proof of Theorem III.1.

In fact, it is known that when the eigenvalues are strictly smallers, the Gaussian process is ergodic as well [13]. We note that in linear deterministic systems Σ_X is equal to identity only when A is orthonormal, in which case, the system is not ergodic. Thus in linear deterministic systems, there is no example of a system that is both ergodic and has an invertible covariance matrix.

B. Additive Noise

In the previous subsection, we saw that for a deterministic linear system, Σ_X being P.D is sufficient and nearly necessary to have $G_C = G_{RDI}$. Now we study the effect of noise present only in the evolution of certain variables. So, the modified system model can be described as $\underline{X}(t) = A\underline{X}(t-1) + \underline{N}(t)$ in which $\underline{N}(t)$ is an $m \times 1$ additive noise vector. We assume the noise to be a vector of mutually-independent zero-mean Gaussian variables with arbitrary variance. Coupled with a Gaussian initial state $N(0, \Sigma_X(0))$, such a linear dynamical system then gives rise to a Gaussian process. For a specific node $i \in [m]$, $var\,(N_i) = 0$ means there is no noise injected to the node. For such a linear dynamical system with additive noise, the stationary condition $\Sigma_X(t) = \Sigma_X(t-1)$ can be expressed as $\Sigma_X = A \Sigma_X A^T + \Sigma_N$. The answer to this equation, if it exists, is:

$$\Sigma_X = \sum_{t=0}^{\infty} A^t \Sigma_N (A^t)^T \tag{11}$$

Theorem III.3. (a) *If for a linear system with independent noise, property A is satisfied, then $G_{RDI} = G_C$.*

(b) *If Property A is not satisfied, then there will be an ambiguity in the correct system and no method will be able to return the causailty graph correctly.*

We note that the proof of part (b) follows similar to Theorem III.1. For part (a), we first prove a version that states that Σ_X is positive definite implies that $G_{RDI} = G_C$ in Lemma III.4. We observe that only Property A is needed for the proof of Lemma III.4, which concludes the proof of this theorem.

Lemma III.4. *Let Σ_X be the covariance matrix of the stationary distribution for a linear system with additive Gaussian noise. If Σ_X is positive definite, then $G_{RDI} = G_C$.*

Proof. We take the steps similar to the ones in the Theorem III.1. In the linear system described above, every variable X_j at each time t can be described as $X_j(t) = \sum_u A_{j,u} X_u(t-1) + N_j(t)$. We will show that given Σ_X is positive definite, for each pair (i,j) if $A_{j,i} = 0$ then $RDI\left(X_i \to X_j | \{X_u\}_{u \in [m]-\{i,j\}}\right) = 0$. Similarly, $RDI\left(X_i \to X_j | \{X_u\}_{u \in [m]-\{i,j\}}\right) > 0$ if $A_{j,i} \neq 0$.

We can write:

$$\begin{aligned} &RDI\left(X_i \to X_j | \{X_u\}_{u \in [m]-\{i,j\}}\right) \\ =&H\left(X_j(t) | \{X_u(t-1)\}_{u \in [m]-\{i\}}\right) \\ &- H\left(X_j(t) | \{X_u(t-1)\}_{u \in [m]}\right) \\ =&H\left(A_{j,i} X_i(t-1) + N_j(t) | \{X_u(t-1)\}_{u \in [m]-\{i\}}\right) \\ &- H\left(N_j(t)\right) \end{aligned} \tag{12}$$

If $A_{j,i} = 0$, then the RHS of (12) is zero, which yields the theorem. Consider the case $A_{j,i} \neq 0$. Since the process is Gaussian, and $N_j(t)$ is independent of the process at time $t-1$, $H\left(A_{j,i} X_i(t-1) + N_j(t) | \{X_u(t-1)\}_{u \in [m]-\{i\}}\right)$ is equal to $\frac{1}{2}\log(2\pi e(A_{j,i}^2 \sigma_e^2 + \sigma_N^2))$ where σ_e^2 is the mean-square error in estimating $X_i(t-1)$ from $X_u(t-1)_{u \in [m]-\{i\}}$. Note that $H\left(N_j(t)\right) = \frac{1}{2}\log(2\pi e \sigma_N^2)$. Now, (12) becomes 0 if and only if $\sigma_e = 0$ which implies that the covariance matrix is singular (indeed, there exists a $\underline{u}$ with $u_i \neq 0$ such that $\underline{u}^T \Sigma_X \underline{u} = 0$). □

Next we show that stability of the dynamical system implies that $\Sigma_X = \sum_{t=0}^{\infty} A^t \Sigma_n (A^t)^T$ is well defined.

Lemma III.5. *Let A and Σ_n be $m \times m$ matrices. Then if the eigenvalues of A are strictly smaller than 1, the limit $\sum_{t=0}^{\infty} A^t \Sigma_n (A^t)^T$ exists, i.e. the series $\sum_{t=0}^{\tau} A^t \Sigma_n (A^t)^T$ is absolutely convergent.*

We omit the proof of this elementary lemma for brevity.

We study a simple case where we can utilize Theorem III.3 to prove that Σ_X is non-singular and hence, RDI can infer the correct graph.

Theorem III.6. *Let A , Σ_X and Σ_N be $m \times m$ matrices, in such a way that $\Sigma_X = \sum_{t=0}^{\infty} A^t \Sigma_N (A^t)^T$ exists. If $\Sigma_N = I$, then Σ_X will be positive definite.*

Proof.

$$\begin{aligned} \forall u \neq 0 : u^T \Sigma_X u &= \sum_{t=0}^{\infty} u^T A^t (A^t)^T u \\ &= \|u\|^2 + \sum_{t=1}^{\infty} \|u^T A^t\|^2 > 0 \\ &\Rightarrow \Sigma_X > 0 \end{aligned} \tag{13}$$

□

C. Semi-deterministic setting

Suppose there is independent noise injected at a group of nodes S_m, a subset of $[m]$. We want to understand when the matrix Σ_X will be positive-definite (P.D.) and RDI method will work. The next theorem states the condition under which a single-node noise injection will result in Σ_X being P.D. Then an immediate result will be stated as a generalization, for the situations when multi-node noise injection will result in a P.D Σ_X.

Theorem III.7. *In III.6, let us assume $\Sigma_N = e_j e_j^T$ where e_j is an arbitrary standard unit vector. So Σ_X will be positive definite if and only if the rank of the matrix $[\dots|A^{k-1}e_j|\dots|Ae_j|e_j]$ is m.*

Proof.

$$\begin{aligned}\Sigma_x > 0 \iff & \forall u \neq 0 : u^T \Sigma_X u = \sum_{t=0}^{k-1} u^T A^t e_j e_j^T (A^t)^T u \\ & = \sum_{t=0}^{k-1} \|u^T A^t e_j\|^2 > 0 \\ \iff & \exists t \geq 0 : u^T A^t e_j \neq 0 \\ \iff & u^T [\dots|A^{k-1}e_j|\dots|Ae_j|e_j] \neq 0 \\ \iff & rank\left([\dots|A^{k-1}e_j|\dots|Ae_j|e_j]\right) = m \end{aligned} \tag{14}$$

□

Corollary III.7.1. *Assume that Σ_N is the covariance matrix of noise in the system, i.e. let $\Sigma_N = \sum_{j \in S_m} r_j e_j e_j^T$ where S_m is a subset of $[m]$. Let us define $B_j = [\dots|A^{k-1}e_j|\dots|Ae_j|e_j]$. Then $\Sigma_x > 0$ if and only if the rank of the matrix $[B_{j_1}|\dots|B_{j_{|S_m|}}]$ in which $(j_1, \dots, j_{|S_m|} \in S_m)$ is m.*

Proof.

$$\begin{aligned}\Sigma_x > 0 \iff & \forall u \neq 0 : u^T \Sigma_x u = \sum_{t=0}^{k-1} u^T A^t \sum_{j \in S_m} r_j e_j e_j^T (A^t)^T u \\ & = \sum_{t=0}^{k-1} \sum_{j \in S_m} |r_j| \|u^T A^t e_j\|^2 > 0 \\ \iff & \exists t \geq 0, j \in S_m : u^T A^t e_j \neq 0 \\ \iff & \exists j \in S_m : u^T [\dots|A^{k-1}e_j|\dots|Ae_j|e_j] \neq 0 \\ \iff & u^T [B_{j_1}|\dots|B_{j_{|S_m|}}] \neq 0 \\ \iff & rank\left([B_{j_1}|\dots|B_{j_{|S_m|}}]\right) = m \end{aligned} \tag{15}$$

□

The theorem III.7 states the condition for having a P.D Σ_X. We now examine whether this condition is likely to be satisfied in a randomized setup. In other words, we are interested in knowing how probable it is to have a dynamical system with this property. For this purpose, we consider graphs whose topologies are fixed, and then the coefficients are randomly generated.

Suppose $G = (V, E)$ is a given directed graph. We generate a *random stable system* with A generated by the following procedure.

$\forall i, j \in E : \hat{A}_{i,j} = Y_{ij}$ in which Y_{ij} is a continuous random variable such that $Y_{ij} \sim \mathcal{N}(0, 1)$ generated i.i.d. for each (i, j). To make sure the system is stable and the covariance matrix of the stationary distribution $\Sigma_x = \sum_{t=0}^{\infty} A^t \Sigma_N (A^t)^T$ for Σ_N exists, we define $A = \frac{1-\epsilon}{\lambda_{max}(\hat{A})}\hat{A}$ and then A will be the transition matrix of the system, i.e. $\forall t > 0 : \underline{X}(t) = A\underline{X}(t-1) + \underline{N}(t)$. The rescaling ensures that the eigenvalues are smaller than $1 - \epsilon$ and hence the system is stable.

Now we show that the *Hamiltonian* property of a graph determines whether Σ_X is non-singular. G is Hamiltonian if it has a Hamiltonian cycle, a cycle with the length of m which meets each vertex exactly once. The Theorem III.10 states that for a randomly-generated linear system, if the graph of the generative model is Hamiltonian, then a single-point noise injection is sufficient for Σ_X to be positive definite with probability 1. Before that, we express two lemmas which is used in the proof of this theorem.

Lemma III.8. *Let $f(X_1, X_2, \dots, X_k)$ be a multivariate polynomial (with finite total-degree). Suppose $\forall i : X_i \sim \mathcal{N}(0, 1)$. Then $Prob\{f(X_1, X_2, \dots, X_k) = 0\} = 0$ if there exists an assignment $X_1 = x_1, X_2 = x_2, \dots, X_k = x_k$ such that $f(x_1, x_2, \dots, x_k) \neq 0$.*

Proof. This follows from standard arguments; see Lemma 2.6 in [14] for example. □

Lemma III.9. *Suppose the hamitonian graph $G = (V, E)$ is given. Then if we assume G to be the causality graph of a linear dynamical system, there is a valid transition matrix A compatible with this graph so that,*

$$det\left([A^{m-1}e_i|\dots|Ae_i|e_i]\right) \neq 0 \; for\ all\ i \in [m]. \tag{16}$$

Proof. Let us denote the Hamiltonian cycle of G with $\mathcal{H}(G) \subset V$. So we define A as the following:

$$A_{j,i} = \begin{cases} 1 & (i, j) \in \mathcal{H}(G) \\ 0 & (i, j) \notin \mathcal{H}(G) \end{cases} \tag{17}$$

So it can be seen that for all $i, j \in [m]$, there is a $k \in [m-1]$ such that $A^k e_i = e_j$. This implies that the matrix $[A^{m-1}e_i|\dots|Ae_i|e_i]$ is a permutation of $I_{m \times m}$ and thus it will be full rank which is equivalent to $det\left([A^{m-1}e_i|\dots|Ae_i|e_i]\right) \neq 0$. □

Theorem III.10. *Let us assume that the causality graph of a linear dynamical system with additive noise is Hamiltonian. Suppose that a single-point noise injection scheme is used, i.e. $\Sigma_N = e_j e_j^T$ for some $j \in [m]$ and the matrix A is generated in the fashion described above so the system is stable. Then for all $j \in [m]$, Σ_x is positive definite with probability 1.*

Proof. Following the same steps as in Theorem III.7, $\Sigma_x > 0$ if and only if the rank of the matrix $B = [\dots|A^2e_j|Ae_j|e_j]$ is m. For the matrix B to have a rank of m it is sufficient that $B_m = [A^{m-1}e_j|\dots|A^2e_j|Ae_j|e_j]$ be full rank, which is equivalent to proving $det(B_m) \neq 0$.

It can be seen that $det(B_m)$ is a polynomial in $A_{i,j}$ with a finite degree. Since the graph is Hamiltonian, from the Lemma III.9 there is an assignment to the matrix A so that $det(B_m) \neq 0$. From Lemma III.8, $Prob\{det(B_m) = 0\} = 0$ which yields the theorem. □

In the next theorem, we prove that if the graph $G = (V, E)$ is not *strongly connected*, i.e. there exists a pair of nodes $i, j \in [m]$ that there is no path from i to j, then single-point noise injection is not sufficient for Σ_X to be P.D.

Theorem III.11. *If the causality graph of a linear dynamical system with additive noise is not strongly connected, then there exists $j \in [m]$ such that for $\Sigma_N = e_je_j^T$, Σ_X is not P.D .*

Proof. We assume that the graph is not strongly connected. So we can conclude that: $\exists i, j \in [m] \forall k \in \mathbb{N} : \left(A^k e_j\right)_i = 0$. In other words, the ith row of $B = [\dots|A^2e_j|Ae_j|e_j]$ is all zeros.

Let us set $\Sigma_N = e_je_j^T$. Similar to the Theorem III.10, showing Σ_X is not P.D is equivalent to showing that $rank\left(B = [\dots|A^2e_j|Ae_j|e_j]\right) < m$. From the graph not being strongly connected we conclude that the ith row of B is all zeros. Hence $rank(B) < m$. □

The Theorem III.10 expresses the sufficiency of the graph being Hamiltonian for Σ_x to be positive definite when the graph is fixed and the coefficients are chosen according to a random distribution. Now, we study a model where both the graph is randomly generated and the coefficients are randomly generated given the graph, as before. The question is: how likely is it for a random graph to be Hamiltonian. This problem is well studied and here we only mention the results. Suppose that G is a random directed graph of n nodes based on the Erdos-Renyi model [15]. The graph is generated simply by considering a complete directed graph with n nodes, and including each edge in G with a probability of p, or removing it with a probability of $1 - p$. We denote such a graph with $D(n, p)$. The theorem below states a condition for which the graph will surely have a Hamiltonian graph.

Theorem III.12. *[16] [17] Let $D(n, p)$ be an Erdos-Renyi graph. If $p \geq (1 + o(1)) \frac{logn}{n}$, then D has a directed Hamiltonian cycle with probability 1.*

Similarly, the Theorem III.11 expresses the necessity of strong connectivity for sufficiency of an arbitrary noise injection to result in a P.D Σ_X. The theorem below states the condition under which the graph will almost always be not strongly connected and hence a single-node noise injection will not be sufficient.

Theorem III.13. *[18] Let $D(n, p)$ be an Erdos-Renyi graph. If $p \leq (1 - o(1)) \frac{logn}{n}$, then D is disconnected with probability 1.*

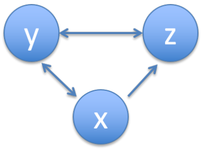

Fig. 1. The causality graph of the Lorenz system

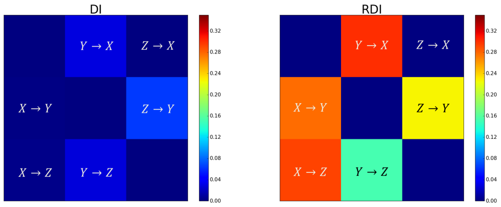

Fig. 2. The heatmap of the directed information (DI) and restricted directed information (RDI) for Lorenz system. Each row represents an inbound node (from up to down: x, y and z respectively) and Each column represents an outbound node (from left to right: x, y and z respectively).

IV. Non-linear Systems

In the section III we discussed about the linear systems. Here we focus on a family of more general functions: non-linear functions. Here, we will concentrate on purely deterministic systems, as in the presence of independent noise at each node, it can be shown that RDI will return the correct graph following steps similar to existing work [4].

In general, a deterministic non-linear dynamical system is described as $\underline{X}(t) = [g_1(\underline{X}(t-1)), \dots, g_n(\underline{X}(t-1))]^T$ in which $\{g_i\}_{i\in[m]}$ are fixed deterministic functions and the alphabet $\mathcal{X}$ is finite. For this system, we would like to study the possibility of inferring the causality graph from RDI method.

The Theorem IV.1 determines the conditions under which the non-linear deterministic system can be inferred by RDI method.

Definition IV.1. Property B*: A causal graph G_c and a probability distribution p_X is said to have property B, if, for all the edges (i, j) in the edge-set of the G_C of a non-linear deterministic system, $H\left(g_j(\underline{X})|\{X_u\}_{u\in[m]-\{i\}}\right) > 0$ under $P_{\underline{X}}$.*

Theorem IV.1. *Consider a deterministic non-linear system in which $\forall t > 0 : \underline{X}(t) = g\left(\underline{X}(t-1)\right)$. Let $P_{\underline{X}}$ be the stationary distribution of the system.*

- *(a) If Property B is true, then $G_{RDI} = G_C$.*
- *(b) If Property B is not true, then no method can return the correct G_C.*

Proof. First we prove (a). For a given (i, j), $RDI\left(X_i \to X_j|\{X_u\}_{u\in[m]-\{i,j\}}\right)$ can be written as:

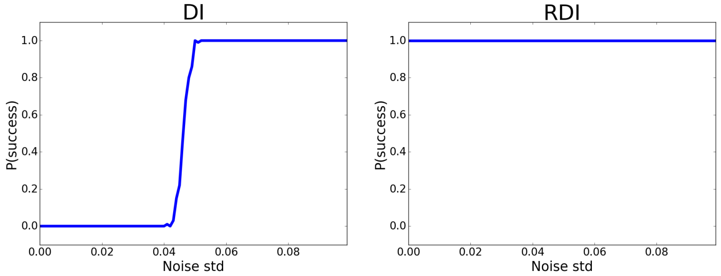

Fig. 3. Probability of successfully extracting the right causality graph for DI and RDI methods.

$$\begin{aligned}
&H\left(X_j(t)|\{X_u(t-1)\}_{u\in[m]-\{i\}}\right)\\
&\quad-\underbrace{H\left(X_j(t)|\{X_u(t-1)\}_{u\in[m]}\right)}_{0}\\
=&H\left(X_j(t)|\{X_u(t-1)\}_{u\in[m]-\{i\}}\right)\\
=&H\left(g_j(\underline{X}(t-1))|\{X_u(t-1)\}_{u\in[m]-\{i\}}\right)\\
=&H\left(g_j(\underline{X})|\{X_u\}_{u\in[m]-\{i\}}\right)
\end{aligned} \tag{18}$$

where the last equation holds under the stationary distribution. If $X_j(t)$ is not a function of $X_i(t-1)$ then $H\left(X_j(t)|\{X_u(t-1)\}_{u\in[m]-\{i\}}\right) = 0$. So $RDI\left(X_i \to X_j|\{X_u\}_{u\in[m]-\{i,j\}}\right) = 0$ and RDI graph correctly does not include the (i,j) edge.

If $X_j(t)$ is a function of $X_i(t-1)$, it implies that (i,j) is in the edge set of the G_C. Given that property B is satisfied, $H\left(g_j(\underline{X})|\{X_u\}_{u\in[m]-\{i\}}\right) > 0$ and (i,j) is included in G_{RDI} hence (a) of the theorem is proved.

Next, we prove (b). Assuming property B is not satisfied means that there exists an edge (i_0,j_0) in G_C for which $H\left(X_j(t)|\{X_u(t-1)\}_{u\in[m]-\{i_0\}}\right) = H\left(g_j(\underline{X})|\{X_u\}_{u\in[m]-\{i\}}\right) = 0$. So $\exists h : X_{j_0}(t) = h\left(\{X_u(t-1)\}_{u\in[m]-\{i_0\}}\right)$.

However, (i_0,j_0) included in G_C means $\exists g_{j_0} : X_{j_0}(t) = g_{j_0}\left(\{X_u(t-1)\}_{u\in[m]-\{i_0\}}, X_{i_0}(t-1)\right)$.

So there is an alternate function $h \neq g_{j_0}$, which yields the same system dynamics. So there is an ambiguity in the system and no algorithm can infer the correct graph. □

While Property B proves the Theorem IV.1, it may not be easy to check. The following lemma will introduce conditions under which property B is satisfied. First we define two concepts which we'll use throughout the proof of the lemma.

Definition IV.2. *A distribution $P(X)$ with $\{X_1, X_2, \ldots, X_n\}$ is said to have a deterministic component X_i if $\exists g : X_i = g(\{X_j\}_{j\in[m]-\{i\}})$, equivalently $H(X_i|\{X_j\}_{j\in[m]-\{i\}}) = 0$.*

Definition IV.3. *Suppose that y is defined as a function of the $m \times 1$ vector $\underline{x}$, i.e. $y = g(\underline{x})$. For a particular x_i, y is called a fully sensitive function of x_i if:*

$$\begin{aligned}
&\forall\{x_j\}_{j\in[m]-\{i\}}\forall x_i, \tilde{x}_i : x_i \neq \tilde{x}_i \iff\\
&g(\{x_j\}_{j\in[m]-\{i\}}, x_i) \neq g(\{x_j\}_{j\in[m]-\{i\}}, \tilde{x}_i)
\end{aligned} \tag{19}$$

In other words, for all fixed $\{x_j\}_{j\in[m]-i}$, there is a bijection between y and x_i.

Lemma IV.2. *For a non-linear deterministic dynamical system, if for all edges (i,j) in the causality graph, X_j is a fully sensitive function of X_i, and the variable X_i is not a deterministic component for the stationary distribution, then property B is satisfied.*

Proof. Consider (i,j) an edge in G_C.

$$\begin{aligned}
&H\left(g_j(\underline{X})|\{X_u\}_{u\in[m]-\{i\}}\right) =\\
&H\left(X_i|\{X_u\}_{u\in[m]-\{i\}}\right) > 0
\end{aligned} \tag{20}$$

The first equality follows from the fact that g_j is fully sensitive and the inequality follows from X_i not being a deterministic component. □

V. Simulations

In the simulation section, we applied RDI method on the *Lorenz System*, as described in Section 1 and compared its performance to DI method. Before describing the simulation setup in greater details, we note that there are many stationary ergodic measures on the Lorenz system; this is because the Lorenz system does have periodic orbits in addition to the strange attractor. Since we are most interested in the strange attractor, there is an invariant measure called the Sinai-Ruelle-Bowen measure [19], [20], which applies to the systems satisfying Axiom A, an example of which is the Lorenz system [21]. Most points near the attractor are attracted towards the strange attractor and result in this particular stationary ergodic measure, which is what we observed in our simulations as well. Empirically, this dataset seems to satisfy the conditions of Lemma IV.2 and therefore satisfies Property B; in particular, the iterations are fully sensitive to the the inputs and the measure seems to have no deterministic component. Proving this formally for the SRB measure is left for future work. By Theorem IV.1, we expect the RDI method to work well in inferring the causal graph of the Lorenz system. We test this hypothesis by a simulation study.

The causality graph of the Lorenz System is shown in Figure 1. As we notice, all nodes influence each other except z to x. The standard values for the parameters of the system are $\sigma = 10$, $\rho = 28$, and $beta = 8/3$ as well as the initial point $\underline{x}(0) = [2,3,4]^T$ which are kept constant throughout the simulations. First, we simulated one run of the system with 1000000 samples and calculated the pairwise conditioned DI's and RDI's. The heatmap of the values is shown in Figure 2. The heatmap qualitatively shows that RDI returns all the corresponding edges in the causality graph of the Lorez system (G_C) correctly, but the DI method does not return the correct answer. Note that the diagonal in both methods is irrelevant (and is set to zero).

In the next simulation, we quantitatively study the performance of network inference in the presence of noise. We assume that i.i.d. Gaussian noise is added to each variable in the system iteration. Recall that both DI and RDI are guaranteed to return the correct solution when noise is present in the system. However, this may require an infinite number of samples. Here, we fix the number of samples to 100000 and compare the reconstruction performance of the two systems as a function of noise variance.

Since both methods return values that maybe non-zero for all the pairs of variables, it becomes necessary to set a threshold in order to declare the causal graph. To calculate these zero thresholds, we calculated the values of conditioned DI and RDI for two independent variables conditioned over a thrid variable, all which were Gaussin zero-mean unit-variance, and from each of which 100000 samples were generated. We repeated the calculations for 100 times, and took the top 5 percentile as the zero threshold. We construct the reconstructed graph to be those edges whose information values are greater than the generated threshold.

Then we defined a *success event* as the event of G_{DI} or G_{RDI} being the same as G_C of the Lorenz system. We changed the standard deviation of the noises from 0.001 to 0.099 with increments 0.001 and calculated the DI and RDI values and observed whether a success happened or not. For each deviation value, we repeated the simulation for 100 times and found the number of total successes as a fraction of total trials (100) and took the resulting value as the *Probability of Success*. The results of simulations is shown in Figure 3.

It can be seen that for small powers of noise, the DI method does not show any success. As the power of the noise increases, around the deviation value of 0.05 we see that DI starts performing well. It shows that the DI method needs a specific amount of noise added at each stage to perform well, and indeed at high SNR the performance of DI method is not satisfactory, while RDI method does not need to rely on the noise to have a good performance and it shows a success probability of 1 for the Lorenz system.

VI. Acknowledgements

The authors would like to thank Xiaojie Qiu and Cole Trapnell for discussions that triggered this work. This work was supported by a startup grant at the University of Washington, Seattle.

References

[1] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[2] C. W. Granger, "Testing for causality: a personal viewpoint," *Journal of Economic Dynamics and control*, vol. 2, pp. 329–352, 1980.

[3] M. Eichler, "Graphical modelling of multivariate time series," *Probability Theory and Related Fields*, vol. 153, no. 1-2, pp. 233–268, 2012.

[4] C. J. Quinn, N. Kiyavash, and T. P. Coleman, "Directed information graphs," *IEEE Transactions on Information Theory*, vol. 61, no. 12, pp. 6887–6909, 2015.

[5] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, pp. 303–305, Citeseer, 1990.

[6] E. N. Lorenz, "Deterministic nonperiodic flow," *Journal of the atmospheric sciences*, vol. 20, no. 2, pp. 130–141, 1963.

[7] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems," *science*, vol. 338, no. 6106, pp. 496–500, 2012.

[8] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical systems and turbulence, Warwick 1980*, pp. 366–381, Springer, 1981.

[9] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[10] W. Gao, S. Kannan, S. Oh, and P. Viswanath, "Conditional dependence via shannon capacity: Axioms, estimators and applications," *arXiv preprint arXiv:1602.03476*, 2016.

[11] W. Gao, S. Oh, and P. Viswanath, "Demystifying fixed k-nearest neighbor information estimators," *arXiv preprint arXiv:1604.03006*, 2016.

[12] A. Chatterjee, A. S. Rawat, S. Vishwanath, and S. Sanghavi, "Learning the causal graph of markov time series," in *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, pp. 107–114, IEEE, 2013.

[13] C. E. Rasmussen, "Gaussian processes for machine learning," 2006.

[14] K. Sreeram, S. Birenjith, and P. V. Kumar, "Dmt of multihop networks: end points and computational tools," *IEEE Transactions on Information Theory*, vol. 58, no. 2, pp. 804–819, 2012.

[15] P. Erdös and A. Rényi, "On random graphs, i," *Publicationes Mathematicae (Debrecen)*, vol. 6, pp. 290–297, 1959.

[16] C. McDiarmid, "Clutter percolation and random graphs," in *Combinatorial Optimization II*, pp. 17–25, Springer, 1980.

[17] A. M. Frieze, "An algorithm for finding hamilton cycles in random directed graphs," *Journal of Algorithms*, vol. 9, no. 2, pp. 181–204, 1988.

[18] I. Palásti, "On the strong connectedness of directed random graphs," *Studia Sci. Math. Hungar*, vol. 1, pp. 205–214, 1966.

[19] Y. G. Sinai, "Gibbs measures in ergodic theory," *Russian Mathematical Surveys*, vol. 27, no. 4, p. 21, 1972.

[20] R. Bowen and D. Ruelle, "The ergodic theory of axiom a flows," in *The Theory of Chaotic Attractors*, pp. 55–76, Springer, 1975.

[21] W. Tucker, "A rigorous ode solver and smales 14th problem," *Foundations of Computational Mathematics*, vol. 2, no. 1, pp. 53–117, 2002.