Learning Minimal Latent Directed Information Trees

Jalal Etesami
Department of Industrial and Enterprising Systems Engineering
University of Illinois
Urbana, Illinois 61801
Email:etesami2@illinois.edu

Negar Kiyavash
Department of Industrial and Enterprising Systems Engineering
University of Illinois
Urbana, Illinois 61801
Email:kiyavash@illinois.edu

Todd P. Coleman
Department of Bioengineering
University of California, San Diego
La Jolla, CA 92093
Email:tpcoleman@ucsd.edu

***Abstract*—THIS PAPER IS ELIGIBLE FOR THE STUDENT PAPER AWARD– We propose a framework for learning the structure of a minimal latent tree with an associated discrepancy measure. Specifically, we apply this algorithm to recover the minimal latent directed information tree on a mixture of set of observed and unobserved random processes. Directed information trees are a new type of probabilistic graphical model based on directed information that represent the casual dynamics among random processes in a stochastic systems. To the best of our knowledge, this is the first approach that recovers these type of latent graphical models where samples are available only from a subset of processes.**

I. Introduction

Latent graphical models refer to a class of probabilistic graphical models that relate a set of observed variables to a set of hidden variables. Introducing latent variables can greatly improve the flexibility of probabilistic modeling, allowing it to address a diverse range of problems with hidden factors, for instance in speech recognition and bio-informatics to name a few.

Graphical models simplify analysis of multivariate probabilistic complex systems. These graphs can be directed, undirected, or a mix. Traditional graphical models explain relationships between groups of random variables without a time axis. Physical systems evolve dynamically with time. To understand causal dynamics, we have recently developed a class of directed graphical models that correspond to random *processes* and reflect its inherent generative model, analogous to such graphs for coupled differential equations [8]. As such, these directed graphical models - termed directed information graphs - have an advantage over undirected graphical models in understanding causation. In this type of graphical models, nodes represent random processes and edges are selected using a criterion that evaluates the directed information [1].

Inference on latent graphical models pertains to when certain random variables are observed, and others are not. In structure learning, it is of interest to discover the topological structure of the graphical model itself (i.e. where edges are present and when there are not), given observations of a subset of the random variables. Some quartet[1]-based distance reconstruction methods have been suggested [2]to discover the structure of latent Markov graphical models [6]. In these type of graphical models, nodes represent random variables and existence or absence of edges represent conditional dependence or independence respectively. Recently, Anandkumar et al. applied this approach to learning linear multivariate tree models when only the leaves are observed [3]. In [3] the nodes are multivariate random vectors and it is assumed that the conditional expected value of each node given its parent is determined by its parent and a deterministic matrix. Recursive grouping (RG) [4] and Chow-Liu recursive grouping (CLRG) [5], [4] are two other distance-based learning algorithms that can recover latent Markov graphical models [6], where some of the observed nodes are internal nodes. Both RG and CLRG can only recover latent models on set of hidden and observed random which are jointly Gaussian or have a symmetric discrete joint distribution [4].

It is the purpose of this paper to develop a framework to perform structure learning on directed information graphs, where we observe subsets of random *processes*. Specifically, we will consider the scenario of *latent directed information trees*, where the directed information graph representing observed and unobserved processes is a tree. In such trees, some of the nodes represent the observed processes while others represent the hidden ones. Therefore learning such trees requires both finding the number of hidden processes as well as recovering the connections among all hidden and observed nodes.

In the aforementioned graphical models literature, nodes are either random variable or multivariate random vectors; hence no notion of time dependency is encoded in this class of graphical models. Consequently, they are unable to capture causal dynamics in stochastic systems. On the contrary, we will recover a fundamentally different graphical model for which the causation among a set of random processes can be encoded. In graphical models of our interest, the nodes encode random processes and the observed nodes can be internal or leaves.

The contributions of this work can be summarized as follows. We propose an algorithm to recover the minimal latent directed information tree on a set of observed and unobserved nodes with an associated discrepancy measure. We

[1]A quartet is an un-rooted binary tree on a set of four observed nodes.

apply our proposed algorithm to finding the latent directed information trees, a new type of probabilistic graphical model based on directed information that may be used to succinctly represent the casual dynamics among random processes in a stochastic systems. In contrast to previous latent tree estimation approaches, (1) our approach is applicable to random processes with temporal dependence structure; (2) we allow the possibility that some observed nodes are internal and (3) we do not require jointly Gaussian or symmetry properties of the joint distribution of nodes.

The remainder of the paper is organized as follows. We formally introduce the minimal latent directed information trees in Section III. In Section IV, we show that given a so-called discrepancy measure on the observed nodes of any tree we can recover the entire latent structure. Specifically in Section V, we show that in case of latent directed information trees, this measure corresponds to the time delays between pairs of observed processes.

II. Notations

Consider a stochastic dynamical system described by m random processes $\underline{\mathbf{X}} = (\mathbf{X}_1, ..., \mathbf{X}_m)$ with joint distribution $P_{\underline{\mathbf{X}}}$ such that each random process contains n random variables. We denote the jth random variable in the ith random process by $X_{i,j}$ and the random process $\mathbf{X}_i$ from time 1 up to time j by $X_{i,1}^j$. Using the chain rule over increasing time indices, the distribution can be factored as

$$P_{\underline{\mathbf{X}}} = \prod_{i=1}^{n} P_{X_{1,i},...,X_{m,i}|X_{1,1}^{i-1},...,X_{m,1}^{i-1}} \ . \tag{1}$$

In many causal systems, given the full past, the future of processes are conditionally independent. In such cases, given the full past of all processes the joint distribution can be simplified to

$$P_{\underline{\mathbf{X}}} = \prod_{i=1}^{n} \prod_{j=1}^{m} P_{X_{j,i}|X_{1,1}^{i-1},...,X_{m,1}^{i-1}} \ . \tag{2}$$

For example, a collection of random processes described by coupled stochastic differential equations satisfies the analogous statement in continuous time.

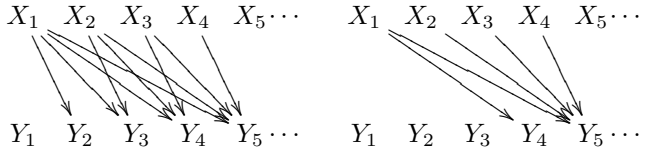
Using casual conditioning notation introduced by Kramer [7],

$$P_{\mathbf{X}_j||\,\underline{\mathbf{X}}_{[j]}}(\mathbf{X}_j||\ \underline{\mathbf{X}}_{[j]}) := \prod_{i=1}^{n} P_{X_{j,i}|X_{1,1}^{i-1},...,X_{m,1}^{i-1}} \ , \tag{3}$$

we can rewrite (2) as

$$P_{\underline{\mathbf{X}}} = \prod_{j=1}^{m} P_{\mathbf{X}_j||\,\underline{\mathbf{X}}_{[j]}}(\mathbf{X}_j||\ \underline{\mathbf{X}}_{[j]}) \ , \tag{4}$$

where $[j] := \{1, ..., m\} \backslash \{j\}$. In this notation the set of random processes $\underline{\mathbf{X}}_{[j]}$ influence the random process $\mathbf{X}_j$ by one time delay. This notation may be generalized to the case for which the dependency has a delay of d time steps. We denote the causal conditioned distribution with a delay of $d \in \mathbb{N}$ time steps as follows



(a) $P_{\mathbf{Y}||\mathbf{X}}$. (b) $P_{\mathbf{Y}||_3\mathbf{X}}$.

Fig. 1: Time dependencies between random processes $\mathbf{X}$ and $\mathbf{Y}$ in two different scenarios. Directed edges are used to show which variables of process $\mathbf{X}$ take part in generating which variable of process $\mathbf{Y}$.

$$P_{\mathbf{X}_1||_d\,\underline{\mathbf{X}}_J} := P_{X_{1,1}^d} P_{X_{1,(d+1)}|X_{1,1}^d,\underline{\mathbf{X}}_{J,1}} \prod_{i=d+2}^{n} P_{X_{1,i}|X_{1,1}^{i-1},\underline{\mathbf{X}}_{J,1}^{i-1}} \ . \tag{5}$$

In equation (5), $\underline{\mathbf{X}}_{J,1}^{i-1}$ stands for $(X_{j_1,1}^{i-1}, ..., X_{j_s,1}^{i-1})$, and $J = \{j_1, ..., j_s\}$ is a set of random processes which influence X_1. Figure 1 illustrates the time dependency in Kramer's notation (a) and equation (5) when $d = 3$ (b). It is easy to check that for $d = 1$, equation (5) becomes Kramer's causal conditioned distribution (3). **actually this is not true. In Kramer, if you are y at time i, you condition on y^{i-1} and x^i. In your case, you condition on x^{i-1}. The punchline is that for continuous-time processes, this is extremely natural - think about coupled ODEs. This is also what Marko implicitly assumed to get his Kirchoff current laws to hold in his 73 paper. Also this leads to a conservation of mutual information $I(X^n; Y^n) = I(X^n \to Y^n) + I(Y^n \to X^n)$ under this assumption. I would change this and use some of the language from the directed information graphs paper. see http://coleman.ucsd.edu/wp-content/uploads/2011/08/QKC_IT_Trans_Apr12.pdf, remark 1.**.

Assumption 1: For the reminder of this paper we only consider a collection of random processes for which (i) there exists a reference measure ϕ such that $\frac{dP_{\underline{\mathbf{X}}}}{d\phi} > 0$ and (ii) the joint distribution is given by (4). For simplicity, we will write $P_{\mathbf{X}||\mathbf{Y}}$ instead of $P_{\mathbf{X}||_1\mathbf{Y}}$.

Denote the Kullback-Leibler divergence for $P_{\underline{\mathbf{X}}}$ and $Q_{\underline{\mathbf{X}}}$ as

$$D(P_{\underline{\mathbf{X}}}||Q_{\underline{\mathbf{X}}}) := \mathbb{E}_{P_{\underline{\mathbf{X}}}}\left[\log \frac{P_{\underline{\mathbf{X}}}}{Q_{\underline{\mathbf{X}}}}\right] \ . \tag{6}$$

Consider two random processes $\mathbf{X}_i$ and $\mathbf{X}_j$ and a set of indices $\mathcal{I}$ such that $\mathcal{I} \subseteq [i, j]$, then the conditional KL divergence, directed information, and conditioned directed information are given, respectively, by

$$D(P_{\mathbf{X}_i||\underline{\mathbf{X}}_{\mathcal{I}}}||\ Q_{\mathbf{X}_i||\underline{\mathbf{X}}_{\mathcal{I}}}|P_{\underline{\mathbf{X}}_{\mathcal{I}}}) := \mathbb{E}_{P_{\underline{\mathbf{X}}_{\mathcal{I}}}}\left[D(P_{\mathbf{X}_i||\underline{\mathbf{X}}_{\mathcal{I}}}||Q_{\mathbf{X}_i||\underline{\mathbf{X}}_{\mathcal{I}}})\right] \ , \tag{7}$$

$$I(\mathbf{X}_j \to \mathbf{X}_i) := D\left(P_{\mathbf{X}_i||\mathbf{X}_j}||\ P_{\mathbf{X}_i}|P_{\mathbf{X}_j}\right) \ , \tag{8}$$

$$I(\mathbf{X}_j \to \mathbf{X}_i||\underline{\mathbf{X}}_{\mathcal{I}}) := D\left(P_{\mathbf{X}_i||\underline{\mathbf{X}}_{\mathcal{I}\cup\{j\}}}||\ P_{\mathbf{X}_i||\underline{\mathbf{X}}_{\mathcal{I}}}|P_{\underline{\mathbf{X}}_{\mathcal{I}\cup\{j\}}}\right) \ . \tag{9}$$

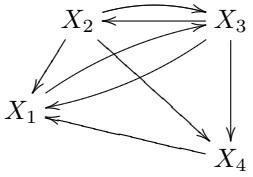


Fig. 2: Generative model graph of Example 1.

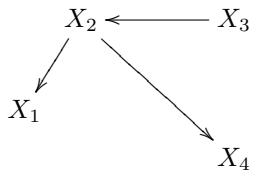


Fig. 3: Directed information tree of Example 2.

III. Generative Models & Directed Information Graphs

Definition 3.1: For a joint distribution a *generative model* is a function $A : \{1,..m\} \to \mathcal{P}(\{1,..m\})$, (power set of $\{1,..m\}$) such that for each process $j \in \{1,..m\}, j \notin A(j)$ and

$$D(P_{\underline{\mathbf{X}}} || \ P_A) = 0 \ ,$$

where $P_A := \prod_{j=1}^{m} P_{\mathbf{X}_j || \ \underline{\mathbf{X}}_{A(j)}}(\mathbf{X}_j || \ \underline{\mathbf{X}}_{A(j)})$.

A directed graph $\overrightarrow{G} = (V, \overrightarrow{E})$ is characterized by a set V of vertices or nodes and a set of ordered pairs of vertices, called arrows or edges $\overrightarrow{E} \subset V \times V$. An undirected graph is called *connected* if there is at least one path between any two nodes and if there is exactly one path between any pair of vertices, then it is called *tree*. A directed tree is a graph such that its undirected underlying graph obtained by replacing all arrows with undirected edges is a tree. In a directed tree $\overrightarrow{T} = (V, \overrightarrow{E})$ a node r without any incoming arrow is called *root* and all of its neighbors are considered to be its children.

Definition 3.2: A *generative model graph* is a directed graph where each node corresponds to a random process, and there is an arrow from i to j, for $i, j \in \{1,..m\}$ if and only if $i \in A(j)$. It is called minimal if for each i, $A(i)$ has minimal cardinality. Under Assumption 1, there is a unique minimal generative model graph [8].

Example 1: Consider the following joint distribution

$$P_{\underline{\mathbf{X}}} = P_{\mathbf{X}_1||\mathbf{X}_4,\mathbf{X}_2,\mathbf{X}_3} P_{\mathbf{X}_2||\mathbf{X}_3} P_{\mathbf{X}_3||\mathbf{X}_1,\mathbf{X}_2} P_{\mathbf{X}_4||\mathbf{X}_2,\mathbf{X}_3} \ .$$

Its generative model graph is depicted in Fig. 2.

Definition 3.3: A *directed information graph* is a directed graph over a set of random processes $\underline{\mathbf{X}}$ where there is an arrow from i to j for $i, j \in \{1, ..., m\}$ if and only if

$$I(\mathbf{X}_i \to \mathbf{X}_j || \ \underline{\mathbf{X}}_{[i,j]}) > 0 \ . \tag{10}$$

Theorem 3.4 ([8]): For any joint distribution $P_{\underline{\mathbf{X}}}$ satisfying Assumption 1, the corresponding minimal generative model graph and directed information graph are equivalent.

In the reminder of this paper we will refer to generative model graphs and directed information graphs interchangeably. Consider a set of random processes $\underline{\mathbf{X}}$ for which the directed information graph is a tree $T = (V, \overrightarrow{E})$, abbreviated as DIT. Denote $\underline{\mathbf{O}} = (\mathbf{X}_1, ..., \mathbf{X}_m)$ as the set of observable processes and their corresponding nodes in the DIT is denoted by O. Likewise, denote $\underline{\mathbf{L}} = (\mathbf{Y}_1, ..., \mathbf{Y}_k)$ as the set of latent processes and their corresponding nodes are denoted by L. Briefly, $\underline{\mathbf{X}} = (\underline{\mathbf{O}}, \underline{\mathbf{L}})$ and $V = O \cup L$.

A probability distribution $P_{\underline{\mathbf{O}}}$ is called *tree-decomposable* if it is a marginal of a tree-structured graphical model $P_{\underline{\mathbf{O}},\underline{\mathbf{L}}}$. In this case, $P_{\underline{\mathbf{O}},\underline{\mathbf{L}}}$ is said to be a tree-extension of $P_{\underline{\mathbf{O}}}$; it is said to have a *redundant latent* node $h \in L$ if we could remove h and the marginal on the set of visible nodes O remains as $P_{\underline{\mathbf{O}}}$. In other words, $h \in L$ is redundant if the directed information graph corresponding to the joint distribution of observed and latent nodes excluding $\mathbf{Y}_h$, $(P_{\underline{\mathbf{O}},\underline{\mathbf{Y}}_{[h]}})$ remains a tree. Finally, a latent directed information tree is called *minimal* if it has no redundant hidden node [6].

Assumption 2: We assume that the joint distribution of the set of observed processes is tree-decomposable and its minimal tree has only one root, i.e., a node without any incoming edge.

Example 2: Fig. 3 demonstrates a directed information tree which satisfies Assumption 2.

$$P_{\underline{\mathbf{X}}} = P_{\mathbf{X}_1||\mathbf{X}_2} P_{\mathbf{X}_2||\mathbf{X}_3} P_{\mathbf{X}_3} P_{\mathbf{X}_4||\mathbf{X}_2} \ .$$

Lemma 3.5: Under Assumption 2, every vertex except the root in the minimal latent directed information tree (LDIT) has exactly one incoming edge. Moreover, all hidden nodes with one incoming and one outgoing edge are redundant hidden nodes. In other words, a minimal LDIT has no internal latent vertex with degree less than three.

Proof: The first statement is immediate from Assumption 2. As for the second claim, we give a proof sketch. Suppose there exists a latent node $(\mathbf{X}_2)$ with one incoming and one outgoing arrow $(\mathbf{X}_1 \to \mathbf{X}_2 \to \mathbf{X}_3)$ in a minimal LDIT. From the definition of a generative model graph, it follows that

$$P_{X_{2,1}^{i-1}|X_{3,1}^{i-1},\mathbf{X}_1} = P_{X_{2,1}^{i-1}|X_{3,1}^{i-1},X_{1,1}^{i-1}} \ . \tag{11}$$

By the chain rule, we have

$$P_{X_{3,i}|X_{3,1}^{i-1},\mathbf{X}_1} = \sum_{X_{2,1}^{i-1}} P_{X_{3,i}|X_{3,1}^{i-1},X_{2,1}^{i-1},\mathbf{X}_1} P_{X_{2,1}^{i-1}|X_{3,1}^{i-1},\mathbf{X}_1} \ . \tag{12}$$

From (11), (12) and the fact that $A(3) = 2$, we obtain $P_{\mathbf{X}_3|\mathbf{X}_1} = P_{\mathbf{X}_3||\mathbf{X}_1}$. ∎

Lemma 3.6: In a system of random processes $\underline{\mathbf{X}}$ with a directed information tree $T = (V, \overrightarrow{E})$. If there is a directed path from $\mathbf{X}_i$ to $\mathbf{X}_j$ and a path from $\mathbf{X}_j$ to $\mathbf{X}_k$, i.e.,

$$\cdots \to \mathbf{X}_i \to \cdots \to \mathbf{X}_j \to \cdots \to \mathbf{X}_k \to \cdots$$

Then the joint distribution of $\mathbf{X}_i, \mathbf{X}_j$ and $\mathbf{X}_k$ factors as follows

$$P_{\mathbf{X}_i,\mathbf{X}_j,\mathbf{X}_k} = P_{\mathbf{X}_i} P_{\mathbf{X}_j||\mathbf{X}_i} P_{\mathbf{X}_k||\mathbf{X}_j} \ , \tag{13}$$

or equivalently,

$$D\left(P_{\mathbf{X}_j|\mathbf{X}_i} || P_{\mathbf{X}_j||\mathbf{X}_i}\right) = D\left(P_{\mathbf{X}_k|\mathbf{X}_j,\mathbf{X}_i} || P_{\mathbf{X}_k||\mathbf{X}_j}\right) = 0 \ . \tag{14}$$

Proof: Without loss of generality, let $\mathbf{X}_1$ to be the root then by the definition of generative model graph and Theorem 3.4, one can write the joint distribution of $\underline{\mathbf{X}}$ as follows

$$P_{\underline{\mathbf{X}}} = P_{\mathbf{X}_1} \cdots P_{\mathbf{X}_i||\mathbf{X}_{A(i)}} \cdots P_{\mathbf{X}_j||\mathbf{X}_{A(j)}} \cdots P_{\mathbf{X}_k||\mathbf{X}_{A(k)}} \cdots \quad (15)$$

By marginalizing over the descendants of $\mathbf{X}_k$ and applying Lemma 3.5 several times for the intermediate nodes between $\mathbf{X}_i$, $\mathbf{X}_j$ and $\mathbf{X}_k$, one can obtain

$$P_{\underline{\mathbf{X}}_{\mathcal{C}_i},\mathbf{X}_i,\mathbf{X}_j,\mathbf{X}_k} = P_{\underline{\mathbf{X}}_{\mathcal{C}_i}} P_{\mathbf{X}_i||\underline{\mathbf{X}}_{\mathcal{C}_i}} P_{\mathbf{X}_j||\mathbf{X}_i} P_{\mathbf{X}_k||\mathbf{X}_j} \ . \quad (16)$$

Where $\underline{\mathbf{X}}_{\mathcal{C}_i}$ is the set of all ancestors and siblings of $\mathbf{X}_i$ in the DIT. The left-hand side of (16) can be expanded by the chain rule as follows

$$P_{\underline{\mathbf{X}}_{\mathcal{C}_i},\mathbf{X}_i,\mathbf{X}_j,\mathbf{X}_k} = P_{\underline{\mathbf{X}}_{\mathcal{C}_i}} P_{\mathbf{X}_i|\underline{\mathbf{X}}_{\mathcal{C}_i}} P_{\mathbf{X}_j|\underline{\mathbf{X}}_{\mathcal{C}_i\cup\{i\}}} P_{\mathbf{X}_k|\underline{\mathbf{X}}_{\mathcal{C}_i\cup\{i,j\}}} \ . \quad (17)$$

Using (16) and (17) and marginalizing over $\underline{\mathbf{X}}_{\mathcal{C}_i}$, (14) is proven. ∎

Lemma 3.7: Consider a system of random processes $\underline{\mathbf{X}}$ with a directed information tree $T = (V, \overrightarrow{E})$. If there is a directed path from j to i of length d, i.e., there is a sequence of nodes $(i_1, ..., i_{d-1})$ where j is the ancestor of i_1, i_k is the ancestor of i_{k+1} for $(1 \leq k \leq d-2)$, and i_{d-1} is the ancestor of i then

$$D(P_{\mathbf{X}_i|\mathbf{X}_j} || \ P_{\mathbf{X}_i||_d\mathbf{X}_j}) = 0 \ . \quad (18)$$

Proof: It is suffices to prove the lemma for $d = 2$, as the case for larger d, may be proven by induction. Consider the case where $d = 2$ ($\mathbf{X} \to \mathbf{Y} \to \mathbf{Z}$). First we prove that

$$P_{Z_m|Z^{m-1},\mathbf{X}} = P_{Z_m|Z^{m-1},X^{m-2}} \ \forall m \leq n \ , \quad (19)$$

Note that if (19) holds, then by multiplying all terms for $m = 1, ..., n$, we obtain

$$P_{\mathbf{Z}|\mathbf{X}} = P_{Z^2} P_{Z_3|Z^2,X_1} \prod_{m=4}^{n} P_{Z_m|Z^{m-1},X^{m-2}} \ ,$$

which proves our claim. Equation (19) can be proved by induction on n. The base case follows from the definition of causal conditioning and Lemma 3.6. Now suppose that (19) holds for m less than k. Note that from the chain rule for any m in particular $m = k$ we have

$$P_{Z_k|Z^{k-1},\mathbf{X}} = \sum_{Y^{k-1}} P_{Z_k|Z^{k-1},Y^{k-1},\mathbf{X}} P_{Z^{k-1}|Y^{k-1},\mathbf{X}} \frac{P_{Y^{k-1}|\mathbf{X}}}{P_{Z^{k-1}|\mathbf{X}}} \quad (20)$$

This equation can be simplified by using induction hypothesis and Lemma 3.6 in particular equations (14) which imply

$$P_{Z_k|Z^{k-1},Y^{k-1},\mathbf{X}} = P_{Z_k|Z^{k-1},Y^{k-1},X^{k-2}} \ , \quad (21)$$

$$P_{Z^{k-1}|Y^{k-1},\mathbf{X}} = P_{Z^{k-1}|Y^{k-1},X^{k-2}} \ , \quad (22)$$

$$P_{Y^{k-1}|\mathbf{X}} = P_{Y^{k-1}|X^{k-2}} \ . \quad (23)$$

Substituting (21), (22), and (23) into the right-hand side of (20) and induction hypothesis prove our claim. ∎

Lemma 3.7 implies that by walking along the path between two random process $\mathbf{X}_i$ and $\mathbf{X}_j$, each time we pass a node, the time dependency between $\mathbf{X}_i$ and $\mathbf{X}_j$ decreases at least by one unit. In the next sections we will see that these time delays will help us to recover the structure of a minimal LDIT.

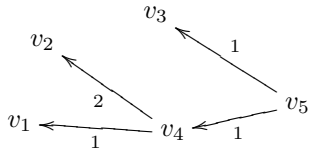


Fig. 4: Directed information tree of Example 3.

IV. Recovery of Latent Trees

A simple observation about the directed tree with one root is that each pair of nodes has a unique common ancestor and also we know by Lemma 3.5 that all leaves in a minimal LDIT are observable. Therefore, if we could find that how far each pair of leaves are from their common ancestor then intuitively, we expect to be able to recover the entire tree. In this section, we define a notion of distance on a tree in order to determine the distance from each pair of nodes to their common ancestor. Moreover, we will show that given these distances for a subset of nodes including leaves, the tree will be recoverable uniquely.

Definition 4.1: Given a tree $T = (V, E)$ with the root r, we define the *discrepancy* between two nodes as a function $\gamma_r(v_1, v_2) : V \times V \to \mathbb{Z}^+$, that assigns a positive integer to the path from v_1 to the common ancestor between v_1 and v_2, such that

1) $\gamma_r(v_1, v_1) = 0$.
2) If v_1 is the ancestor of v_2 then $\gamma_r(v_1, v_2) = 0$.
3) If the path from v_1 to v_3 contains the common ancestor of v_1 and v_2, then

$$\gamma_r(v_1, v_2) < \gamma_r(v_1, v_3).$$

The image of this function can be presented by the discrepancy matrix:

$$\Gamma_V := [\gamma_r(v_i, v_j)] \ , \quad v_i, v_j \in V \ .$$

Note that for a given tree, the discrepancy matrix is not unique. Any function that satisfies Definition 4.1's conditions is a valid discrepancy measure.

Example 3: Consider the tree depicted in Fig. 4 with root v_5 and a discrepancy matrix given by

$$\Gamma_V = \begin{pmatrix} 0 & 1 & 2 & 1 & 2 \\ 2 & 0 & 4 & 2 & 4 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

For instance, looking at the third row, this particular function assigns 1 to the path from v_3 to its common ancestor with v_1 which is v_5. Similarly, the weight of the path from v_3 to its common ancestor with v_2, v_5, is one and so on.

Note that the discrepancy matrix describes the topology of a tree $T = (V, E)$ uniquely. Moreover, under some conditions a tree can be recovered just by knowing the discrepancy matrix of a proper subset of V.

Theorem 4.2: Let $T = (V, E)$ to be a tree with root r and $S \subseteq V$ such that every node $v \in V \setminus (S \cup \{r\})$ has degree at

least 3. Then, the discrepancy matrix of S (Γ_S), will determine V and E uniquely.

Proof: The proof is by induction on $|S|$. Suppose, a tree $T = (V, E)$ can be recovered uniquely, by any $S \subseteq V$ such that every node $v \in V \setminus (S \cup \{r\})$ has degree at least 3 and $|S| \leq k-1$. For the case that $|S| = k$, let $B_v := \arg\min_{u \in S \setminus \{v\}} \gamma_r(v,u)$. If $B_v = S \setminus \{v\}$ for all v then T is a star with one hidden node at the center. Otherwise, fix a vertex $v \in S$ such that $B_v \neq S \setminus \{v\}$. If $\min_{u \in S \setminus \{v\}} \gamma_r(v,u) = 0$, then all the nodes in B_v are the descendants of v. In this case by induction hypothesis, the subtree of T containing v and all its descendant, is recoverable by $B_v \cup \{v\}$ as well as the rest of the tree by $S \setminus B_v$. The case of $\min_{u \in S \setminus \{v\}} \gamma_r(v,u) > 0$ is similar, however, in this case, B_v and $\{v\}$ have a common hidden ancestor which is also connected to v. ∎

V. Discrepancy Measure for Latent Directed Information Trees

In this section, we establish the discrepancy measure for a minimal directed information tree. Lemma 3.7 states that the lag between random processes grow by walking along the directed paths in a minimal DIT. This allows us to have the following definition in a minimal DIT.

Definition 5.1: For any pairs of random processes $(\mathbf{X}_j, \mathbf{X}_k) \in \underline{\mathbf{X}}$, we define the directed measure from $\mathbf{X}_j$ to $\mathbf{X}_k$ denoted by $\gamma(j,k)$ as follows

$$\gamma(j,k) := \max_d \left\{ d : \; D\left(P_{\mathbf{X}_k|\mathbf{X}_j} || \; P_{\mathbf{X}_k||_d \mathbf{X}_j}\right) = 0 \right\} \; , \quad (24)$$

whenever $j \neq k$ and it is zero otherwise.
Clearly, $\gamma(j,k) \neq \gamma(k,j)$. Note, that (24) implies

$$I(\mathbf{X}_j; X_{k,1}^{\gamma(j,k)}) = 0 \;\; \& \;\; I(\mathbf{X}_j; X_{k,1}^{\gamma(j,k)+1}) > 0 \; . \quad (25)$$

Thus if (24) exists, then equivalently

$$\gamma(j,k) = \max_d \left\{ d : \; I(\mathbf{X}_j; X_{k,1}^{d}) = 0 \right\} \; . \quad (26)$$

Theorem 5.2: Let $\underline{\mathbf{X}}$ be a collection of random processes with a graphical model which is a minimal directed information tree $T = (V, \overrightarrow{E})$, then the directed measure defined in (5.1) is a discrepancy on T.

Proof: We show that for a given path $\mathbf{X} \to \mathbf{Y} \to \mathbf{Z}$ in a minimal LDIT, if $\gamma(X,Y) = l$ and $\gamma(Y,Z) = d$ then $\gamma(X,Z) > \max\{l, d\}$.
Let us prove that $\gamma(X,Z) > d$. It suffices to show

$$P_{Z_{d+1}|Z^d,\mathbf{X}} = P_{Z_{d+1}|Z^d} \; . \quad (27)$$

Applying the chain rule to the left-hand side of (27):

$$P_{Z_{d+1}|Z^d,\mathbf{X}} = \sum_{Y_1} P_{Z_{d+1}|Z^d,Y_1,\mathbf{X}} P_{Z^d|Y_1,\mathbf{X}} \frac{P_{Y_1|\mathbf{X}}}{P_{Z^d|\mathbf{X}}} \; . \quad (28)$$

Since $\gamma(Y,Z) = d$, i.e., $D(P_{\mathbf{Z}|\mathbf{Y}} || \; P_{\mathbf{Z}||_d \mathbf{Y}}) = 0$, by using the same argument as in the proof of Lemma 3.7, we obtain

$$P_{Z^d|\mathbf{Y},\mathbf{X}} = P_{Z^d} \; , \; P_{Y_1|\mathbf{X}} = P_{Y_1} \; , \quad (29)$$

$$P_{Z_{d+1}|Z^d,Y_1,\mathbf{X}} = P_{Z_{d+1}|Z^d,Y_1} \; . \quad (30)$$

Finally, the claim follows by substituting (29) and (30) into the right-hand side of (28). The statement $\gamma(X,Z) > l$ may be proven by showing $P_{Z_{l+1}|Z^l,\mathbf{X}} = P_{Z_{l+1}|Z^{l+1}}$ in a similar fashion. ∎

VI. Conclusion and Future Work

This work develops a new approach for learning a latent tree when a certain discrepancy measure is available for the observed nodes. This procedure may be applied to learning latent directed information trees from the samples of observed processes. Our algorithm produces a matrix of integer values based on the samples and uses the elements of the matrix to discover the hidden nodes and the connections between the hidden and observed nodes. Note that Theorem 5.2 showed that a discrepancy measure may be obtained for minimal DITs. This in conjunction to Theorem 4.2 means that we can recover the structure of a minimal LDIT from the samples available at the observed nodes. Equation (25) leads us to extract the directed measures by estimating mutual information between the present of one observed process and past of other observed processes. Future work entails consistently estimate mutual information from data; this is feasible under appropriate statistical assumptions [10], [9], [14].

Acknowledgment

This work was supported in part to N. Kiyavash by AFOSR under grants FA 9550-11-1-0016, FA 9550-10-1-0573, and FA 9550-10-1-0345; and by NSF grant CCF 10-54937 CAR; and to T. Coleman by NSF Science & Technology Center grant CCF-0939370 and NSF grant CCF 10-65352.

References

[1] J. Massey, "Causality, feedback and directed information," in *Proc. Int. Symp. Information Theory Application (ISITA-90)*, 1990, pp. 303-305.

[2] T. Jiang, P. E. Kearney, and M. Li, "A polynomial-time approximation scheme for inferring evolutionary trees from quartet topologies and its application," SIAM J. Comput. 30, pp. 1942-1961, 2001.

[3] A. Anandkumar, K. Chaudhuri, D. Hsu, S. M. Kakade, L. Song, and T. Zhang, "Spectral Methods for Learning Multivariate Latent Tree Structure," 2011, submitted, arXiv:1107.1283.

[4] M.J. Choi, V. Tan, A. Anandkumar, and A. Willsky, "Learning Latent Tree Graphical Models." *Journal of Machine Learning Research,* 2011.

[5] C. Chow, C. Liu, "Approximating discrete probability distributions with dependence trees" *IEEE Transaction on Information Theory,* vol. 14, no. 3, pp. 462-467, 1968.

[6] J. Pearl, "Probabilistic reasoning in intelligent systems: networks of plausible inference." *Morgan Kaufmann*, 1988.

[7] G. Kramer, "Directed information for channels with feedbacks," Ph.D. dissertation, University of Manitoba, Canada, 1998.

[8] C. Quinn, N. Kiyavash, and T. Coleman, "Equivalence Between Minimal Generative Model Graphs and Directed Information Graphs," in *Information Theory Proc. (ISIT), IEEE International Symposium*, 2011, pp. 293-297.

[9] I. Csiszár, Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *Information Theory, IEEE Transactions*, vol. 52, no. 3, pp. 1007-1016, 2006.

[10] H. Cai, S. R. Kulkarni, and S. Verdú, "Universal entropy estimation via block sorting", *Information Theory, IEEE Transactions*, vol.50, no. 7, pp. 1551-1561, 2004.

[11] J. Ziv, A. Lempel, "A Universal Algorithm for Sequential Data Compression'" *IEEE Transactions on Information Theory*, vol. 23, no. 3, pp. 337-343, 1977.

[12] F. Perez-Cruz, "Estimation of information theoretic measures for continuous random variables," NIPS, 2008.

[13] Q. Wang, S.R. Kulkarni, and S. Verdú, "Divergence estimation of continuous distributions based on data-dependent partitions", *Information Theory, IEEE Transactions*, vol. 51, no. 9, pp. 3064-3074, 2005.

[14] X.L. Nguyen, M.J. Wainwright, and M.I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *Information Theory, IEEE Transactions*, vol. 56, no. 11, pp. 5847-5861, 2010.