ISIT'2013 1569720215

Robust Directed Tree Approximations for Networks of Stochastic Processes

Christopher J. Quinn
Dep. of Electrical & Computer Eng.
University of Illinois
Urbana-Champaign, IL
Email: quinn7@illinois.edu

Jalal Etesami and Negar Kiyavash
Dep. of Industrial & Enterprise Systems Eng.
University of Illinois
Urbana-Champaign, IL
Email: {etesami2, kiyavash}@illinois.edu

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

***Abstract*—We develop low-complexity algorithms to robustly identify the best directed tree approximation for a network of stochastic processes in the finite-sample regime. Directed information is used to quantify influence between stochastic processes and identify the best directed tree approximation in terms of Kullback-Leibler (KL) divergence. We provide finite-sample complexity bounds for confidence intervals of directed information estimates. We use these confidence intervals to develop a minimax framework to identify the best directed tree that is robust to point estimation errors. We provide algorithms for this minimax calculation and describe the relationships between exactness and complexity.**

I. Introduction

Identifying the structure of large, directed networks is an important research problem in a variety of disciplines. In the stock market, traders and investors might want to understand how the past stock price fluctuations of one stock might affect those of another. In systems neuroscience, there is increasing interest in understanding the brain as a network of interacting processes. In the above and other fields, scientists investigate the causal dependencies of the nodes in networks, often from noisy data.

When the networks of interest are large, from thousands of stocks to billions of brain cells, identifying the full structure can be prohibitive or unnecessary for the desired understanding. In such cases, working with an approximate network structure, such as a tree, can be beneficial. Trees are easy to visualize and analyze, as there is a single root node with paths from the root to all other nodes. Only pairwise statistics are needed to find the best tree, as compared to the full distribution over all processes in the more general case. Efficient algorithms exist to identify the best trees.

When the data used to infer the network structures is limited/noisy, confidence intervals might be preferred to point estimates for statistical estimates. Small errors in point estimates can lead to large errors in the selected trees. This leads to the important problem of (a) identifying sample complexity limits on estimating pairwise directed information to obtain relevant confidence intervals, and (b) developing a framework to identify optimal tree approximations that operate on intervals, as compared to point estimates, to enable robustness to uncertainty. In this paper, we address both problems. We use directed information to quantify influence between stochastic processes in a network. To address problem (a), we characterize sample complexity of two types of directed information estimators to derive confidence intervals. To address problem (b), we develop a minimax framework to determine robust tree approximations, describe its complexity, and some algorithms for implementation.

Some related works that discuss estimating directed information to identify network structure include [1]–[3]. [4] and [5] study sample complexity for entropy and mutual information estimators respectively. [6] finds the error exponent for learning tree structures for Markov networks.

This paper is organized as follows. Section II describes the problem of finding an optimal directed tree approximation for a network using noisy data. A minimax framework is proposed to ensure the tree approximation is robust to estimation errors. Section III finds sample complexity bounds to derive confidence intervals for directed information estimates, in both parametric and non-parametric settings. Section IV discusses algorithms to identify the optimal robust tree and to efficiently find a near-optimal robust tree. Section V describes simulations of finding robust trees from data.

II. Setup

Consider a network of m random processes $\underline{\mathbf{X}} = \{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ that are jointly stationary, ergodic, and Markov of finite order $l \ll n$. Let $P_{\underline{\mathbf{X}}}$ denote the joint distribution. Denote the time indices as $t \in \{-l+1, \ldots, n\}$.

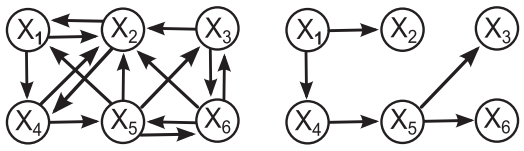
Let $\mathcal{T}$ denote the set of directed spanning trees on m nodes. We will approximate $P_{\underline{\mathbf{X}}}$ with a distribution $P_{\underline{\mathbf{X}};T}$ whose topological structure corresponds to a spanning tree $T \in \mathcal{T}$. Using causal conditioning notation, where for two processes Y^n and Z^n,

$$P_{Y^n \| Z^n}(y^n \| z^n) \triangleq \prod_{t=1}^{n} P_{Y_t | Y^{t-1}, Z^{t-1}}(y_t | y^{t-1}, z^{t-1}), \qquad (1)$$

the approximating tree distribution for network $\underline{\mathbf{X}}$ is

$$P_{\underline{\mathbf{X}};T}(\underline{\mathbf{x}}) \triangleq \prod_{i=1}^{m} P_{\mathbf{X}_{\pi(i)} \| \mathbf{X}_{b(\pi(i))}}(\mathbf{x}_{\pi(i)} \parallel \mathbf{x}_{b(\pi(i))}) \qquad (2)$$

where π is a permutation on $\{1, \ldots, m\}$ and $0 \leq b(i) < i$ with $\mathbf{X}_0$ denoting a deterministic constant (for the root node's dependence). See Fig 1 for an example.



(a) Full topology. (b) Tree approximation.

Fig. 1. Topologies of a full distribution $P_{\underline{\mathbf{X}}}$ and a tree approximation.

It was shown in [7] that the optimal approximate tree distribution in terms of KL divergence is the one that has the maximum sum of directed information values along its edges:

Theorem 1.

$$\arg\min_{T\in\mathcal{T}} D(P_{\underline{\mathbf{X}}}\|P_{\underline{\mathbf{X}};T}) = \arg\max_{T\in\mathcal{T}} \sum_{i=1}^{m} \mathrm{I}(\mathbf{X}_{b(\pi(i))}\to\mathbf{X}_{\pi(i)}). \quad (3)$$

Directed information was defined by Marko [8] as[1]:

$$I(\mathbf{X}\to\mathbf{Y}) \;=\; \int_{\mathbf{x}} D\big(P_{\mathbf{Y}\|\mathbf{X}=\mathbf{x}}\|P_{\mathbf{Y}}\big)\, P_{\mathbf{X}}(d\mathbf{x}). \quad (4)$$

Theorem 1 is analogous to the problem studied by Chow and Liu [9], showing that the best undirected tree approximation for a network of random variables is the one with the maximum sum of mutual informations on its edges. Note that in [9], nodes represent random variables and edges depict conditional dependencies. In this work, nodes are random processes and edges depict statistical causation.

Also similar to [9], given directed information estimates $\widehat{\mathrm{I}}(\mathbf{X}_i\to\mathbf{X}_j)$ for all ordered pairs (i,j), an efficient maximum weight directed spanning tree (MWDST) algorithm can be used to identify the best tree. One such algorithm is Edmonds [10], which runs in $\mathcal{O}(m^2)$ time.

In this paper, we consider the practical setting when the directed information estimates might not be reliable, but confidence intervals are known for the estimates. Suppose that for every ordered pair $(\mathbf{X}_i,\mathbf{X}_j)$ of processes, instead of computing a single value $\widehat{\mathrm{I}}(\mathbf{X}_i\to\mathbf{X}_j)$ for an estimate of $\mathrm{I}(\mathbf{X}_i\to\mathbf{X}_j)$, a confidence interval $\widehat{\mathcal{I}}(\mathbf{X}_i\to\mathbf{X}_j) = [\delta^-_{(i,j)},\delta^+_{(i,j)}]$ is available with an associated joint probability. For example, the intervals $\{\widehat{\mathcal{I}}(\mathbf{X}_i\to\mathbf{X}_j)\}$ used might be 95% confidence intervals, meaning

$$\mathbb{P}\Big(\bigcap_{1\le i\ne j\le m} \{\mathrm{I}(\mathbf{X}_i\to\mathbf{X}_j)\in\widehat{\mathcal{I}}(\mathbf{X}_i\to\mathbf{X}_j)\}\Big)\ge 0.95.$$

Let r index the set $\{(i,j)\}$ of ordered pairs with $i,j\in\{1,\dots,m\}$ and $i\ne j$. Let $R=m(m-1)$ be the total number of such pairs. The rth pair is denoted as (i_r,j_r). Denote the Cartesian product of the intervals as

$$\mathcal{S}\triangleq \bigtimes_{r=1}^{R}\widehat{\mathcal{I}}(\mathbf{X}_{i_r}\to\mathbf{X}_{j_r}). \quad (5)$$

Note that $\mathcal{S}$ is a subset of $\mathbb{R}^R$. Each element $s\in\mathcal{S}$ is a length R vector. The rth coordinate s_r corresponds to a directed information estimate for the rth edge, which we denote as $\widehat{\mathrm{I}}_s(\mathbf{X}_{i_r}\to\mathbf{X}_{j_r})\triangleq s_r$. We refer to each $s\in\mathcal{S}$ as a *scenario*. Given a directed spanning tree $T\in\mathcal{T}$, denote its weight in scenario s as

$$W(T,s)\triangleq \sum_{(i_r,j_r)\in T}\widehat{\mathrm{I}}_s(\mathbf{X}_{i_r}\to\mathbf{X}_{j_r}). \quad (6)$$

In a particular scenario s, denote the MWDST as

$$T^*(s)\triangleq \arg\max_{T\in\mathcal{T}} W(T,s). \quad (7)$$

For a given scenario s, we can efficiently find $T^*(s)$. However, $T^*(s)$ might perform quite poorly for another scenario $s'\in\mathcal{S}$. We want to select a tree T that performs well for all scenarios $s\in\mathcal{S}$. In particular, we want to select the "robust" tree T_{rob} that attains the minimax regret:

$$T_{\mathrm{rob}}\triangleq \arg\min_{T\in\mathcal{T}} \max_{s\in\mathcal{S}} \{W(T^*(s),s)-W(T,s)\}. \quad (8)$$

In Section IV, we discuss the complexity of (8) and algorithms to solve it given $\mathcal{S}$. We next describe how to obtain a set of confidence intervals $\mathcal{S}$ for each of the directed informations. We consider a parametric and a non-parametric estimator. We obtain the confidence intervals by characterizing the sample complexity of both estimators. Sample complexity refers to how the confidence interval width δ decays as the number of samples n grows, for a fixed probability of error and fixed m. We also consider *graph sample complexity,* which refers to how fast n needs to grow as the network size m increases for fixed δ and probability of error.

III. Confidence Intervals

In this section, we consider directed information estimation using a non-parametric empirical estimator for finite-alphabets, as well as a parametric estimator. We compute the confidence intervals and their corresponding probabilities by identifying the sample complexity for the estimators.

Assumption 1. *The network $\underline{\mathbf{X}}$ is jointly stationary, ergodic, and Markov of finite order l. Also, each pair of processes $\{\mathbf{X}_i,\mathbf{X}_j\}$ are jointly stationary and Markov order l.*

Under Assumption 1, the directed information is

$$\begin{aligned}\mathrm{I}(\mathbf{X}_i\to\mathbf{X}_j) &= \frac{1}{n}\sum_{t=1}^{n}\mathrm{I}(X_{j,t};X_{i,t-l}^{t-1}|X_{j,t-l}^{t-1}) &(9)\\ &= \mathrm{I}(X_{j,l+1};X_{i,1}^{l}|X_{j,1}^{l}) &(10)\\ &= \sum_{\{x_i^l,x_j^{l+1}\}\in\mathsf{X}^{(2l+1)}} P_{X_i^l,X_j^{l+1}}(x_i^l,x_j^{l+1})\\ &\qquad\times\log\frac{P_{X_{j,l+1}|X_i^l,X_j^l}(x_{j,l+1}|x_i^l,x_j^l)}{P_{X_{j,l+1}|X_j^l}(x_{j,l+1}|x_j^l)}. &(11)\end{aligned}$$

(9) follows from Markovicity, (10) follows from stationarity, and (11) follows from the definition of mutual information.

[1] Works using directed information in the specific context of communication channels with synchronous input/output condition on Z_t in (1).

We first estimate the pairwise distributions $\widehat{P}_{X_i^l,X_j^{l+1}}$, and then plug those into (11) to obtain a directed information estimate $\widehat{\mathrm{I}}(\mathbf{X}_i \to \mathbf{X}_j)$. The confidence interval is set as

$$\widehat{\mathcal{I}}(\mathbf{X}_i \to \mathbf{X}_j) = \left[\widehat{\mathrm{I}}(\mathbf{X}_i \to \mathbf{X}_j) - \delta, \widehat{\mathrm{I}}(\mathbf{X}_i \to \mathbf{X}_j) + \delta\right]$$

for a given constant $\delta > 0$. Let B_δ denote the event that all pairwise directed informations are within their respective intervals, i.e.,

$$B_\delta \triangleq \{\forall r \in \{1,\dots,R\}, \left|\widehat{\mathrm{I}}(\mathbf{X}_{i_r} \to \mathbf{X}_{j_r}) - \mathrm{I}(\mathbf{X}_{i_r} \to \mathbf{X}_{j_r})\right| < \delta\}. \quad (12)$$

We next examine the sample complexity of these estimators to characterize $\mathbb{P}(B_\delta)$ as a function of n.

A. Non-parametric Empirical estimator for Finite Alphabets

In the setting of finite alphabet X, we will use the "empirical" distribution for a non-parametric estimator. For each ordered pair (i,j) we compute a distribution $\widehat{P}_{X_i^l,X_j^{l+1}}$, where for each possible realization $\{x_i^l, x_j^{l+1}\} \in \mathsf{X}^{2l+1}$ of $\{X_i^l, X_j^{l+1}\}$, the estimates are

$$\widehat{P}_{X_i^l,X_j^{l+1}}(x_i^l, x_j^{l+1}) \triangleq \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}_{\{\{X_{i,t-l}^{t-1}, X_{j,t-l}^{t}\} = \{x_i^l, x_j^{l+1}\}\}}. \quad (13)$$

To ensure convergence of the empirical estimator (13), we will make the following assumption on the mixing time of the process. Denote the state of the network from time $t-l$ to time t by $\underline{\mathrm{V}}_t \triangleq \underline{X}_{t-l}^t$. Then $\{\underline{\mathrm{V}}_t\}_{t=1}^n$ forms a first-order Markov chain. We assume this chain satisfies the following condition which is related to uniform ergodicity [11]:

Assumption 2. *There exists a probability measure $\phi(v)$ on $\mathsf{X}^{m(l+1)}$, a constant $0 < \lambda \leq 1$, and an integer $d \geq 2$ such that for all $v_1 \in \mathsf{X}^{m(l+1)}$, $\mathbb{P}(V_d = v | V_1 = v_1) \geq \lambda\phi(v)$.*

If the Markov chain converges to a stationary distribution $\pi(v)$, then Assumption 2 can be applied with $\phi(v) = \pi(v)$ and $\lambda \approx 1$ for sufficiently large d. There is a trade-off between decreasing d and increasing λ.

Theorem 2. *Under Assumptions 1 and 2, $\mathbb{P}(B_\delta) \geq 1 - \rho$, where*

$$\rho = 8R|\mathsf{X}|^{2l+1}\exp\left(-\frac{(n\epsilon - 2d/\lambda)^2}{2nd^2/\lambda^2}\right). \quad (14)$$

For any $\epsilon' > 0$, the sample complexity of Algorithm 1 is $\delta = \mathcal{O}(n^{-1/2+\epsilon'})$, and the graph sample complexity is $n = \mathcal{O}(\log m)$.

Proof. We first obtain concentrations on the empirical distribution from Hoeffding and union bounds. We then use an L_1 bound on entropy to translate concentrations on entropies to ones on the directed information estimates.

We require the following concentrations on the empirical probability distributions. For every pair (i,j), for every possible realization $\{x_i^l, x_j^{l+1}\} \in \mathsf{X}^{2l+1}$, and for a given $\epsilon > 0$ which we will later fix as a function of δ:

$$|\widehat{P}_{\underline{\mathbf{Z}}}(\underline{\mathbf{z}}) - P_{\underline{\mathbf{Z}}}(\underline{\mathbf{z}})| < \epsilon, \quad (15)$$

for $\underline{\mathbf{Z}} \in \{\{X_i^l, X_j^{l+1}\}, \{X_i^l, X_j^l\}, \{X_j^{l+1}\}, \{X_j^l\}\}$.

From the Hoeffding inequality generalized to uniformly ergodic Markov chains [11], under Assumption 2, for any (i,j) and any realization $\{x_i^{(l+1)}, \underline{x}_{\underline{\mathrm{A}}}^l\} \in \mathsf{X}^{(2l+1)}$,

$$\mathbb{P}\left(\left|\widehat{P}_{X_i^l,X_j^{l+1}}(x_i^l, x_j^{l+1}) - P_{X_i^l,X_j^{l+1}}(x_i^l, x_j^{l+1})\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{(n\epsilon - 2d/\lambda)^2}{2nd^2/\lambda^2}\right). \quad (16)$$

Applying the union bound to (16), the four inequalities in (15) hold for each of the $|\mathsf{X}|^{2l+1}$ realizations for each of the R pairs of processes $\{(i_r, j_r)\}$ with probability ρ, given in (14).

We next find what value of ϵ corresponds to the event B_δ. For simplicity, denote $\{X_i^l, X_j^{l+1}\}$ by $\underline{\mathbf{Z}}$. We want a concentration on $|\widehat{H}(\underline{\mathbf{Z}}) - H(\underline{\mathbf{Z}})|$. First note that

$$\|\widehat{P}_{\underline{\mathbf{Z}}} - P_{\underline{\mathbf{Z}}}\|_1 \triangleq \sum_{\underline{\mathbf{z}} \in \mathsf{X}^{2l+1}} |\widehat{P}_{\underline{\mathbf{Z}}}(\underline{\mathbf{z}}) - P_{\underline{\mathbf{Z}}}(\underline{\mathbf{z}})| \leq |\mathsf{X}|^{2l+1}\epsilon, \quad (17)$$

where (17) follows from (15).

Using an L_1 bound on entropy, if $\|\widehat{P}_{\underline{\mathbf{Z}}} - P_{\underline{\mathbf{Z}}}\|_1 \leq \frac{1}{2}$, then

$$|\widehat{H}(\underline{\mathbf{Z}}) - H(\underline{\mathbf{Z}})| \leq -\|\widehat{P}_{\underline{\mathbf{Z}}} - P_{\underline{\mathbf{Z}}}\|_1 \log \frac{\|\widehat{P}_{\underline{\mathbf{Z}}} - P_{\underline{\mathbf{Z}}}\|_1}{|\mathsf{X}|^{2l+1}}. \quad (18)$$

The bound is of the form $-b\log\frac{b}{c}$, which is concave in b and maximized at $b = \frac{c}{e}$. With $\epsilon \leq \frac{1}{e}$, the upper bound in (17), $|\mathsf{X}|^{2l+1}\epsilon$, is in the interval $(0, |\mathsf{X}|^{2l+1}/e]$ where the bound (18) is increasing. Thus, (18) can be bounded using (17):

$$|\widehat{H}(\underline{\mathbf{Z}}) - H(\underline{\mathbf{Z}})| \leq -|\mathsf{X}|^{2l+1}\epsilon\log\epsilon. \quad (19)$$

Note that directed information (10) decomposes into a linear combination of entropies:

$$\mathrm{I}(X_{j,l+1}; X_i^l | X_j^l) = H(X_j^{l+1}) - H(X_j^l) - H(X_j^{l+1}, X_i^l) + H(X_j^l, X_i^l). \quad (20)$$

Applying the triangle inequality and (19) to (20) gives that for all R pairs (i_r, j_r),

$$|\widehat{\mathrm{I}}(\mathbf{X}_{i_r} \to \mathbf{X}_{j_r}) - \mathrm{I}(\mathbf{X}_{i_r} \to \mathbf{X}_{j_r})| \leq -4|\mathsf{X}|^{2l+1}\epsilon\log\epsilon. \quad (21)$$

Setting $\delta = -4|\mathsf{X}|^{2l+1}\epsilon\log\epsilon$ would conclude the proof. However, to obtain an analytic expression for how ϵ depends on δ, we will bound $\epsilon\log\epsilon$ with a polynomial expression. The function $-\epsilon\log\epsilon$ has a maximum value of $\frac{1}{e}$ on the interval $\epsilon \in (0,1)$. That value is attained at $\epsilon = \frac{1}{e}$. For all $0 < a < 1$,

$$-\epsilon\log\epsilon = \frac{1}{a}\epsilon^{1-a}(-\epsilon^a\log\epsilon^a) \leq \frac{1}{ae}\epsilon^{1-a}. \quad (22)$$

For large ϵ, the bound with larger a is tighter; for small ϵ, the bound with small a is tighter. For all $0 < a < 1$ and all r,

$$|\widehat{\mathrm{I}}(\mathbf{X}_{i_r} \to \mathbf{X}_{j_r}) - \mathrm{I}(\mathbf{X}_{i_r} \to \mathbf{X}_{j_r})| \leq \frac{4|\mathsf{X}|^{2l+1}}{ae}\epsilon^{1-a}. \quad (23)$$

Setting the value of ϵ as $\epsilon = \left(\frac{ae\delta}{4|\mathsf{X}|^{2l+1}}\right)^{\frac{1}{1-a}}$ finishes the proof that $\mathbb{P}(B_\delta) \geq 1-\rho$.

Note that for a fixed probability of error ρ (14), fixed m, and sufficiently large $n\epsilon$, that as n increases, ϵ decays as $n^{-1/2}$ which implies that $\delta = \mathcal{O}(n^{-1/2+\epsilon'})$ for all $\epsilon' > 0$. Alternatively, if m is increasing, to maintain a fixed probability of error ρ with a fixed δ, n needs to increase as $\log m$. □

B. Parametric estimator

Parametric models are widely used for modeling time series in economics, biology, and other fields. In this section, we identify sample complexity results for networks of stochastic processes whose conditional distribution $P_{\underline{\mathbf{X}}_t|\underline{\mathbf{X}}_{t-l}^{t-1};\underline{\theta}^*}$ is characterized by a parameter vector $\underline{\theta}^*$. Since $\underline{\mathbf{X}}$ is stationary by Assumption 1, $P_{\underline{\mathbf{X}}_t|\underline{\mathbf{X}}_{t-l}^{t-1};\underline{\theta}^*}$ determines the stationary distribution and thus the joint distribution. We next discuss conditions for the maximum likelihood estimate (MLE) $\widehat{\underline{\theta}}_n$ to exist.

Let $\mathcal{F}_t$ denote the σ-field generated by $\underline{\mathbf{X}}^t$. Suppose $\underline{\theta}^*$ is an unknown parameter vector in the interior of Θ, a compact subset of $\mathbb{R}^Q$. Let q index the parameter vector $\underline{\theta} = \{\theta_q\}_{q=1}^Q$. Denote the conditional log-likelihood of $\underline{\mathbf{X}}_t$ as

$$L_t(\underline{\theta}) \triangleq \log P_{\underline{\mathbf{X}}_t|\underline{\mathbf{X}}_{t-l}^{t-1};\underline{\theta}}(\underline{\mathbf{X}}_t|\underline{\mathbf{X}}_{t-l}^{t-1}). \tag{24}$$

Define the matrices $A_t(\underline{\theta})$ and $G_t(\underline{\theta})$ evaluated at $\underline{\theta}'$ as

$$A_t(\underline{\theta}') = \left[-\frac{\partial^2 L_t(\underline{\theta})}{\partial\theta_{q_1}\partial\theta_{q_2}}\bigg|_{\underline{\theta}=\underline{\theta}'}\right]_{1\leq q_1,q_2\leq Q} \tag{25}$$

$$G_t(\underline{\theta}') = \left[\frac{\partial L_t(\underline{\theta})}{\partial\theta_{q_1}}\bigg|_{\underline{\theta}=\underline{\theta}'}\frac{\partial L_t(\underline{\theta})}{\partial\theta_{q_2}}\bigg|_{\underline{\theta}=\underline{\theta}'}\right]_{1\leq q_1,q_2\leq Q}. \tag{26}$$

Assumption 3. *$L_n(\underline{\theta})$ is almost surely and continuously twice differentiable in terms of $\underline{\theta}$. $\mathbb{E}[\sup_{\theta\in\Theta}[L_t(\underline{\theta})]] < \infty$ and $\mathbb{E}[L_t(\underline{\theta})]$ has a unique maximizer at $\underline{\theta}^*$. The vector $[\frac{\partial L_t(\underline{\theta})}{\partial\theta_q}|_{\underline{\theta}=\underline{\theta}^*}]_{1\leq q\leq Q}$ is a martingale difference in terms of $\mathcal{F}_t$ with $\mathbb{E}[G_t(\underline{\theta}^*)]$ finite and positive definite. $\mathbb{E}[A_t(\underline{\theta}^*)]$ is positive definite and $\mathbb{E}[\sup_{\underline{\theta}:\|\underline{\theta}-\underline{\theta}^*\|_2<\eta}\|A_t(\underline{\theta})\|_2] < \infty$ for some $\eta > 0$.*

Remark 1. *Several classes of autoregressive (AR) time-series models satisfy Assumption 3 such as threshold AR models, bilinear AR moving averages, and GARCH models [12].*

Define the covariance matrix

$$\Sigma \triangleq [E[A_t(\underline{\theta}^*)]]^{-1} E[G_t(\underline{\theta}^*)] [E[A_t(\underline{\theta}^*)]]^{-1}. \tag{27}$$

Lemma 1. *Under Assumptions 1 and 3,*

$$\sqrt{n}(\widehat{\underline{\theta}}_n - \underline{\theta}^*) \to \mathcal{N}(0,\Sigma) \quad \textit{in distribution}. \tag{28}$$

This follows from [12]. Lemma 1 extends to functions of the parameters. Let $g_r(\underline{\theta})$ denote the directed information (4) of the rth pair (i_r, j_r) computed with $\underline{\theta}$:

$$g_r(\underline{\theta}) \triangleq \mathrm{I}(\mathbf{X}_{i_r} \to \mathbf{X}_{j_r}). \tag{29}$$

Using the $Q\times Q$ parameter covariance matrix $\Sigma = (\sigma_{q,q'})$ (27), define the $R\times R$ covariance matrix $\Sigma' = (\sigma'_{r,r'})$ for the directed information estimates as $\sigma'_{r,r'} = \sum_{q=1}^Q\sum_{q'=1}^Q \sigma_{q,q'}\frac{\partial g_r}{\partial\theta_q}\frac{\partial g_{r'}}{\partial\theta_{q'}}\Big|_{\underline{\theta}=\underline{\theta}^*}$.

Lemma 2. *Under Assumptions 1 and 3,*

$$\sqrt{n}\left[(g_1(\widehat{\underline{\theta}}_n) - g_1(\underline{\theta}^*)),\ldots,(g_R(\widehat{\underline{\theta}}_n) - g_R(\underline{\theta}^*))\right] \to \mathcal{N}(0,\Sigma') \quad \textit{in distribution}. \tag{30}$$

This follows by the multivariate delta method. Note that under Assumptions 1 and 3, the unknown covariance matrices Σ and Σ' in (28) and (30) respectively can be consistently estimated by using $\widehat{\underline{\theta}}_n$ in place of the unknown $\underline{\theta}^*$ [12].

Remark 2. *Analytically calculating Σ' might prove challenging in some cases. We briefly note an alternative, approximate method. The directed information (4) is the difference of two entropy terms $\mathrm{I}(\mathbf{X}\to\mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y}\|\mathbf{X})$. The parameter vectors $\widehat{\underline{\theta}}'_n$ and $\widehat{\underline{\theta}}''_n$ corresponding to the conditional marginals $P_{Y_t|Y_{t-l}^{t-1};\widehat{\underline{\theta}}'_n}$ and $P_{Y_t|Y_{t-l}^{t-1},X_{t-l}^{t-1};\widehat{\underline{\theta}}''_n}$ can be separately estimated. Parameter values from the confidence regions for $\widehat{\underline{\theta}}'_n$ and $\widehat{\underline{\theta}}''_n$ can then be sampled to approximate the corresponding confidence intervals for the entropies and thus for $\widehat{I}(\mathbf{X}\to\mathbf{Y})$.*

We next identify the sample complexity results.

Theorem 3. *Under Assumptions 1 and 3, the sample complexity is $\delta = \mathcal{O}(n^{-1/2})$ and the graph sample complexity is $n = \mathcal{O}(\log m)$.*

Proof. We first find the graph sample complexity. We lower bound $\mathbb{P}(B_\delta)$. Note that the equiprobable contours of $\mathcal{N}(0,\Sigma')$ form ellipsoids with principal axis lengths proportional to the largest eigenvalue of Σ'. Let σ'' denote the largest eigenvalue of Σ'. Define a new diagonal covariance matrix Σ'' whose entries are all σ''. Then the probability of any volume centered at zero under $\mathcal{N}(0,\Sigma')$ will be larger than $\mathcal{N}(0,\Sigma'')$. Also, since Σ'' is diagonal, the corresponding random variables are independent. Thus,

$$\begin{aligned}\mathbb{P}(B_\delta) &= \mathbb{P}\left(\{-\delta \leq g_r(\widehat{\underline{\theta}}_n) - g_r(\underline{\theta}^*) \leq \delta)\}_{r=1}^R\right)\\ &\geq \left[\mathbb{P}\left(-\delta\frac{\sqrt{n}}{\sigma''} \leq \frac{\sqrt{n}}{\sigma''}(g_r(\widehat{\underline{\theta}}_n) - g_r(\underline{\theta}^*)) \leq \delta\frac{\sqrt{n}}{\sigma''}\right)\right]^R \quad (31)\\ &= \left[\mathrm{erf}\left(\delta\frac{\sqrt{n}}{\sqrt{2}\sigma''}\right)\right]^R \quad (32)\end{aligned}$$

where (31) uses the independence of the error estimates under distribution $\mathcal{N}(0,\Sigma'')$ and normalizes them, and (32) uses the "error" function $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,\mathrm{d}t$.

Using the first two terms of the asymptotic expansion of $\mathrm{erf}(x)$ with appropriate constants c_1 and c_2,

$$\begin{aligned}\left[\mathrm{erf}\left(\delta\frac{\sqrt{n}}{\sqrt{2}\sigma''}\right)\right]^R &\approx \left[1 - \frac{c_1}{\sqrt{n}}e^{-c_2 n}\right]^R \quad (33)\\ &\approx 1 - \frac{m^2 c_1}{\sqrt{n}}e^{-c_2 n} \quad (34)\\ &= 1 - c_1 e^{2\log(m) - c_2 n - \frac{1}{2}\log n}. \quad (35)\end{aligned}$$

Equation (34) uses the first two terms in the binomial expansion. Repeat these steps setting σ'' as the minimum eigenvalue of Σ' to get an upper bound in (31). This finishes the proof for the graph sample complexity.

For the sample complexity rate, note the $\delta\sqrt{n}$ terms in the normalized inequalities in (31). Thus $\delta = \mathcal{O}(n^{-1/2})$. □

IV. Algorithms

In this section, we discuss how to determine the robust tree T_{rob} (8) given set of scenarios $\mathcal{S}$ (5). Finding T_{rob} is NP-hard [13]. There is an exact algorithm using branch and bound techniques. There is also an efficient 2-approximation algorithm. Both algorithms use the following property. For a given directed spanning tree $T \in \mathcal{T}$, define $s^T = \{s_r^T\}$ where

$$s_r^T \triangleq \begin{cases} \widehat{\mathrm{I}}(\mathbf{X}_{i_r} \to \mathbf{X}_{j_r}) - \delta, & \text{if } (i_r, j_r) \in T \\ \widehat{\mathrm{I}}(\mathbf{X}_{i_r} \to \mathbf{X}_{j_r}) + \delta, & \text{if } (i_r, j_r) \notin T \end{cases}. \tag{36}$$

Then s^T is the worst-case scenario for T.

Lemma 3. *Define s^T as in* (36)*. Then s^T maximizes the regret, i.e., $s^T = \arg\max_{s \in \mathcal{S}} \{W(T^*(s), s) - W(T, s)\}$.*

This follows from [14]. The branch and bound algorithm was introduced by Conde [13]. The branching process partitions possible sets of directed spanning trees based on whether they include or do not include specific edges. The bounding process finds upper and lower bounds for the maximum regret of the partitioned sets of spanning trees using a particular choice of edge weights and solving the MWDST problem for those edge weights.

There is also a 2-approximation. Let s^{mid} be defined as $s_r^{\text{mid}} \triangleq \widehat{\mathrm{I}}(\mathbf{X}_{i_r} \to \mathbf{X}_{j_r})$, the midpoint of interval $\widehat{\mathcal{I}}(\mathbf{X}_{i_r} \to \mathbf{X}_{j_r})$. $T^*(s^{\text{mid}})$ is at least half as robust as T_{rob}.

Lemma 4. *$T^*(s^{\text{mid}})$ satisfies the following:*

$$\max_{s \in \mathcal{S}} \{W(T^*(s), s) - W(T^*(s^{\text{mid}}), s)\} \leq 2 \max_{s \in \mathcal{S}} \{W(T^*(s), s) - W(T_{\text{rob}}, s)\}.$$

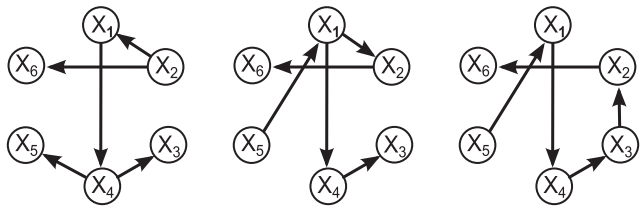
This follows from [14]. Lemma 4 justifies using the estimates $\{\widehat{\mathrm{I}}(\mathbf{X}_{i_r} \to \mathbf{X}_{j_r})\}$ directly to compute a MWDST.

V. Simulations

We simulated a network of $m = 6$ processes with $n = 10^5$. They were modeled as a zero-mean multivariate normal autoregressive time-series such that $\underline{\mathbf{X}}_t = B\underline{\mathbf{X}}_{t-1} + \xi_t$, where ξ_t was i.i.d. Gaussian noise. B was randomly generated.

The MLE $\widehat{B}$ was computed using least squares and the 95% confidence region $\mathcal{S}_B$ for its parameters. We computed $\mathcal{S}$ (5) by sampling $\widehat{B}$ uniformly from $\mathcal{S}_B$ to compute confidence intervals $\widehat{\mathcal{I}}(\mathbf{X}_i \to \mathbf{X}_j)$ for each pair of processes $\{\mathbf{X}_i, \mathbf{X}_j\}$.

Using $\mathcal{S}$, the near-optimal robust MWDST $T^*(s^{\text{mid}})$ was computed (see Lemma 4) and is depicted in Figure 2(a). Additionally, 10^3 scenarios $s \in \mathcal{S}$ were drawn randomly and the corresponding MWDSTs were computed. The most robust MWDSTs found are shown in Figure 2(a-c). The MWDST for the generating distribution, denoted as $T^*(s^*)$ was also calculated. $T^*(s^{\text{mid}})$ is the same tree as $T^*(s^*)$. $T^*(s^{\text{mid}})$ was the most robust. Edges $\mathbf{X}_1 \to \mathbf{X}_4$, $\mathbf{X}_2 \to \mathbf{X}_6$, and $\mathbf{X}_4 \to \mathbf{X}_3$ are common to all trees in Figure 2.



(a) Regret 0.41. $T^*(s^*)$; $T^*(s^{\text{mid}})$; (b) Regret 0.45. (c) Regret 0.48.

Fig. 2. The three most robust trees found by randomly sampling scenarios. The most robust tree (a) was the 2-approximation $T^*(s^{\text{mid}})$ (see Lemma 4) and the MWDST for the true distribution.

Acknowledgment

Christopher Quinn was supported by the DoE Computational Science Graduate Fellowship under grant number DE-FG02-97ER25308. This work was supported in part by AFOSR under grant FA 9550-10-1-0573; and by NSF grants CCF 10-54937 CAR and CNS 08-31488.

References

[1] S. Kim, D. Putrino, S. Ghosh, and E. N. Brown, "A Granger causality measure for point process models of ensemble neural spiking activity," *PLoS Comput Biol*, vol. 7, no. 3, March 2011.

[2] C. Quinn, T. Coleman, N. Kiyavash, and N. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 17–44, 2011.

[3] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, "Universal estimation of directed information," *ArXiv e-prints*, Jan. 2012.

[4] K. Sricharan, R. Raich, and A. O. Hero, III, "Empirical estimation of entropy functionals with confidence," *ArXiv e-prints*, Dec. 2010.

[5] R. Wu, R. Srikant, and J. Ni, "Learning graph structure in discrete Markov random fields," *ArXiv e-prints*, Apr. 2012.

[6] V. Tan, A. Anandkumar, L. Tong, and A. Willsky, "A large-deviation analysis of the maximum-likelihood learning of Markov tree structures," *Information Theory, IEEE Transactions on*, vol. 57, no. 3, pp. 1714–1735, 2011.

[7] C. J. Quinn, T. P. Coleman, and N. Kiyavash, "Efficient Methods to Compute Optimal Tree Approximations of Directed Information Graphs," *IEEE Transactions on Signal Processing*, 2013, forthcoming.

[8] H. Marko, "The bidirectional communication theory–a generalization of information theory," *Communications, IEEE Transactions on*, vol. 21, no. 12, pp. 1345–1351, Dec 1973.

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

[10] J. Edmonds, "Optimum branchings," *J. Res. Natl. Bur. Stand., Sect. B*, vol. 71, pp. 233–240, 1967.

[11] P. Glynn and D. Ormoneit, "Hoeffding's inequality for uniformly ergodic Markov chains," *Statistics & probability letters*, vol. 56, no. 2, pp. 143–146, 2002.

[12] S. Ling and M. McAleer, "A general asymptotic theory for time-series models," *Statistica Neerlandica*, vol. 64, no. 1, pp. 97–111, 2010.

[13] E. Conde, "A branch and bound algorithm for the minimax regret spanning arborescence," *Journal of Global Optimization*, vol. 37, no. 3, pp. 467–480, 2007.

[14] A. Kasperski and P. Zieliński, "An approximation algorithm for interval data minmax regret combinatorial optimization problems," *Information Processing Letters*, vol. 97, no. 5, pp. 177–180, 2006.