Score-and-search methods for the recovery of structure in networks of dynamic systems

Darya Biparva and Donatello Materassi

***Abstract*— A relevant problem in many areas of science is to determine the structure of a network by observing its nodes. A desirable property of any network reconstruction techniques is consistency, namely the convergence of the reconstructed network to the actual structure when the time horizon of the observations goes to infinity. Unfortunately, when feedthrough components are present in the network, multiple structures could give rise to the same observations. Hence, in these situations, the best theoretical result that can be achieved is the determination of all the possible structures compatible with what is being observed. There are some results offering such theoretical guarantees, but these methods rely on a large number of statistical tests making their sample complexity relatively large. This article proposes the adoption of reconstruction techniques where, given the observed data, each structure is evaluated according to a score representing the likelihood that such structure is the actual one. Such a technique is proven to have the same consistency properties of state-of-the-art methods based on statistical tests, while numerical experiments show it to have a lower sample complexity.**

I. Introduction

In many areas of science, it is often of interest to recover a graphical representation that highlights the connections among different components in a distributed system just by using observational data. Indeed, reconstructing the unknown graph underlying the interconnection of a distributed dynamic system with multiple components is an active field of research with applications in medicine [1], economics[2] and climate studies [3]. Given some a priori knowledge about the system, several techniques can provide guarantees of a correct reconstruction of the underlying graph [4], [5], [6], [7], [8], [9]. For example, the work in [5] can reconstruct the underlying graph of a network under the assumption that it has a tree structure with exactly one root. The work in [8] relaxes these assumptions by considering trees with multiple roots while the work in [7] considers directed acyclic graphs. When all the transfer functions in the network are strictly proper there are approaches that allow a reconstruction of the graph even in presence of loops such as the work in [4], [6], [10] or approaches based on variations of Granger causality or directed information [11], [12]. The work in [9] generalizes these techniques since it considers networks with generic topologies (including the presence of feedback loops) making only the mild assumption that every feedback loop includes at least a strictly causal transfer function. A fundamental limitation for all these techniques is given by the fact that, in presence of feedthrough terms, the orientation of all edges might not always be recovered since multiple structures compatible with the data might exist. For example, [5] does not orient the recovered edges for the underlying tree, at all. The results in [9] recover the orientations for the subset of the edges that correspond to strictly causal transfer functions. Even though the orientation of all edges is not always recovered, these works typically guarantee the convergence to the correct undirected topology when the length of the observed time series is sufficiently large. However, these techniques have limited applications in scenarios where the length of the time series tends to be relatively short (e.g. DNA microarray data). Another drawback of these techniques that limits their applicability to real case scenarios is the fact that the detection of each individual edge typically requires to perform a relatively large number of statistical tests resulting in an overall high computational complexity.

Thus, it is desirable to devise better performing techniques requiring at the same time lower sample complexity and lower computational complexity. In the area of graphical models, there is a specific class of methods for selecting a graph that fits i.i.d. observational data with the goal of keeping the sample complexity as low as possible. This class of methods is often referred to as score-and-search methods [13]. For any given graph, score-and-search methods assign a score describing how well the graph fits the observed data and then implement a search to determine the graph with the maximum score. The fundamental contribution of this article is to borrow core ideas behind score-and-search methods and apply them to the reconstruction of networks of dynamic systems. In particular, we show that a specific score function, often referred to as Bayesian Information Criterion (BIC), provides some form of consistency properties in the sense that its maximization, for sufficiently long time series, results in the determination of all possible structures compatible with the data. A related approach for network discovery using the BIC score was recently applied in [10] but only in the context of networks with strictly proper dynamics where the problem of network reconstruction is always well-posed.

The article outline is as follows. In Section II we describe Linear Dynamic influence Models (LDIMs), which are the class of models considered in this article to obtain the main theoretical results. In Section III, we formulate the network reconstruction problem observing that exact recovery of the structure of an LDIM from data is an ill-posed question, in the general case. Section IV provides preliminary notions from the theory graphical models that are then exploited in Section V to obtain consistency guarantees about the topology reconstruction of an LDIM. Section VI illustrates the consistency results via some numerical experiments.

II. Generative class of models

The focus of this article is the reconstruction of the connectivity structure of network models of a specific form which have been extensively studied in the system identification literature [14], [4], [5], [15].

D. Biparva and D. Materassi are with Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55414 bipar001@umn.edu, mater013@umn.edu

Definition 1 (Linear Dynamic Influence Models (LDIMs))**.** *Linear Dynamic Influence Models (LDIMs) are a class of networks representing input/output relations among stochastic processes. An LDIM is defined as a pair $(H(z), e)$ where*

- *$e = (e_1, e_2, ..., e_n)^T$ is a vector of n rationally related random scalar processes that are mutually independent (namely the cross-power spectral density $\Phi_{e_i e_j} = 0$ for $i \neq j$ and $\Phi_{e_i}(z)$ is a rational function for $i = 1, ...n.$)*
- *$H(z)$ is an $n \times n$ transfer matrix with rational entries.*

The output processes $\{y_j\}_{j=1}^n$ of the LDIM are defined as $y_j = e_j + \sum_{i=1}^n H_{ji}(z) Y_i$ or $y = e + H(z)y$ where $y = (y_1, ..., y_n)^T$. Furthermore

- *If the entries of $H(z)$ are all proper, we say that the LDIM is proper.*
- *If the power spectral density matrix of e is a constant, namely $\Phi_e(z) = \Phi_e$, we say that the LDIM has static excitation.*
- *If the system has no algebraic loops, we say that the LDIM is causally well-posed.*

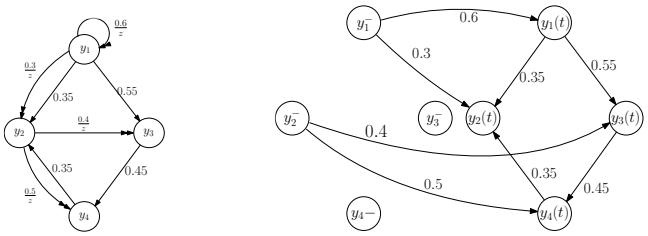
The structure of the connections among the output processes $\{y_j\}_{j=1}^n$ of an LDIM $(H(z), e)$ can be suggestively represented using a graph where each output process is a vertex and the set of edges represents the sparsity structure of the matrix $H(z)$. We assume the reader is already familiar with standard notions of graph theory and we just briefly introduce our notation. We denote a directed graph G as a pair (V, E) where V is the set of vertices and $E \subseteq V \times V$ is the set of directed edges. We represent a directed edge from $y_i \in V$ to $y_j \in V$ as (y_i, y_j), or $y_i \to y_j$, or $y_j \leftarrow y_i$. If $y_i \to y_j \in E$, we say that y_j is a child of y_i or, equivalently, that y_i is a parent of y_j using the notation $pa_{y_j}^G$ to denote the set of all parent of y_j. We say that two edges are adjacent if they share a vertex. If a path goes from one node to itself, the path is a cycle. Furthermore, if all the edges have the same orientation, we say that the path is directed. We note y_j a descendant of y_i if there exists a directed path $y_i \to ... \to y_j$.

The skeleton of a directed graph is the undirected graph obtained by removing all the orientations from the edges in the graph.

In this article, in order to represent the structure of an LDIM conveying information about the presence of potential feedthrough terms, we use a generalized notion of graphs, known as multi-typed graphs [9], [16]. Multi-typed graphs are an extension of standard graphs where there are edges of different types connecting the nodes. Specifically in the representation of an LDIM connectivity, we are going to use two different types of edges to describe the presence of feedthroughs and/or the presence of strictly causal components in each of the entries $H_{ji}(z)$.

Definition 2 (Multi-arrowed graph representation of LDIMs [9])**.** *Let $\mathcal{G} = (H, e)$ be an LDIM with output processes $y_1, ..., y_n$. Consider a multi-typed graph $G = (V, E_1, E_2)$ with the two sets of edges $E_1, E_2 \subseteq V \times V$. We say that G is a graphical representation for LDIM $\mathcal{G}$ if the following two properties hold*

- *$H_{ji}(z)$ has a non-zero direct feedthrough component implies $(y_i, y_j) \in E_1$ and*
- *$H_{ji}(z)$ has a non-zero strictly proper component implies $(y_i, y_j) \in E_2$.*



(a) (b)

Fig. 1: *(a) and (b) respectively depict multi-arrowed graph representation and hybrid representation of an LDIM. The transfer functions associated with each edge are the components of $H(z)$ in the underlying LDIM.*

If the above statements hold in the opposite direction as well, we say that G is the perfect graph representation for $\mathcal{G}$. We refer to E_1 as the direct feedthrough edges and represent its elements with a single-headed edge. We refer to E_2 as the strictly proper component edges and represent its elements with a double-headed edge. If the standard graph (V, E_1) which contains only the single headed edges has no loops, we say that the graphical representation G is recursive.

Since multi-typed graphs are a generalization of standard graphs, typical notions defined for standard graphs, such as parents, children, paths, cycles, extend naturally to multi-typed graphs by considering a standard graph with the same nodes and having as edges the union of all edge sets.

In some situations, it might be more convenient to work with an equally informative representation of an LDIM in terms of a standard graph which we refer to as hybrid graph.

Definition 3 (Hybrid Representation of LDIMs)**.** *Let $G = (V, E_1, E_2)$ be a graphical representation of an LDIM $\mathcal{G}$ with output processes $V := \{y_1, ..., y_n\}$. Let $S^- := \{y_1^-, ..., y_n^-\}$ a set of nodes representing the "past" of each process in V, namely the time series $y_i^- = y_i|_{-\infty}^{t-1}$, $1 \leqslant i \leqslant n$. Let $S^+ := \{y_1(t), ..., y_n(t)\}$ be a set of nodes representing the "present" of each process in V. Let $V^h := S^- \cup S^+$ be the nodes for the hybrid graph G^h. The set of edges E^h is defined as follows:*

- *$(y_i(t), y_j(t)) \in E^h$ if and only if $(y_i, y_j) \in E_1$,*
- *$(y_i^-, y_j(t)) \in E^h$ if and only if $(y_i, y_j) \in E_2$.*

In Figure 1(b) we have the hybrid graph associated to the graphical representation of Figure 1(a). Also, by looking at the transfer functions on the edges of both figures, it is easy to observe that they are both graphical representation of one unique LDIM.

It is straightforward from Definition 3 that given any graphical representation of an LDIM, we can immediately obtain an associated hybrid graph and viceversa.

III. Problem statement

A relevant problem in many areas of science is to determine the structure of a network by observing its nodes. In the case of an LDIM, we could formalize this problem as the reconstruction of its perfect graphical representation from the observation of its output processes. Unfortunately, such a problem is in general ill-posed. Indeed, as Example

1 in Section IV will show, multiple LDIMs with different perfect graphical representations can generate exactly the same output processes. Thus, the best we can achieve is to determine the structures of all LDIMs which are compatible with the observed data. This consideration justifies the following problem statement.

Problem. Determine a method that, from the observation of the output processes generated by an LDIM $\mathcal{G}^*$, consistently recovers the perfect graphical representations of all LDIMs G compatible with the observed data.

IV. Basic notions from the theory of graphical models

In order to tackle the problem formulated in Section III, we need to recall some definitions from the theory of graphical models [13]. In the theory of graphical models, a special role is played by graph substructures often referred to as colliders.

Definition 4 (Colliders, Shielded and unshielded colliders)**.** *Given a graph G, we say that a node y_k is a collider between y_i and y_j if $y_i \to y_k$ and $y_j \to y_k$ are edges in the graph. If either $y_i \to y_j$ or $y_j \to y_i$ are edges in the graph, the collider is shielded. Otherwise we say it is unshielded.*

A set of vertices Z can activate a path according to the following definition.

Definition 5 (Active path)**.** *Let $y_{\pi_0}, ..., y_{\pi_\ell}$ be a path π in graph G. Let Z be a subset of variables in G such that $y_{\pi_0}, y_{\pi_\ell} \notin Z$. The path $y_{\pi_0}, ..., y_{\pi_\ell}$ is active given Z if*

- *In the path π for every collider y_{π_i} of the form $y_{\pi_{i-1}} \to y_{\pi_i} \leftarrow y_{\pi_{i+1}}$, y_{π_i} or one of its descendants are in Z, and*
- *all other nodes in the path are not Z.*

The following notion of d-separation in a graph is central in the theory of graphical models.

Definition 6 (d-separation)**.** *Let Z be a subset of nodes in G. We say that y_i and y_j are d-separated given Z if there's no active path between y_i and y_j given Z. We write this relation as $dsep_G(y_i, y_j|Z)$.*

Observe that the concept of d-separation is purely a "graphical notion" since it can be defined just from the graphical representation of an LDIM. Instead "probabilistic notions" of separation can be defined among stochastic processes in the following way.

Definition 7 (Feedthrough Wiener Separation)**.** *Let $\mathcal{G}$ be an LDIM with output processes $Y = \{y_1, ..., y_n\}$. Consider y_i, y_j and Z such that $y_i, y_j \notin Z$. Define $\bar{Z} := Y \backslash (Z \cup \{i, j\})$. Compute the causal Wiener filter estimating $y_j(t)$ from $y_i|_{-\infty}^{t}$, $y_j|_{-\infty}^{t-1}$, $y_Z|_{-\infty}^{t}$ and $y_{\bar{Z}}|_{-\infty}^{t-1}$. Let $W_{ji}(z)$ be the component associated with y_i. If $W_{ji}(z)$ has no feedthrough component, we say that y_i and y_j are feedthrough-Wiener-separated by Z. We write such a relation as $cwsep^+(y_i, y_j|Z)$*

Definition 8 (Strictly proper Wiener Separation)**.** *Let $\mathcal{G}$ be an LDIM with output processes $Y = \{y_1, ..., y_n\}$. Consider y_i, y_j and Z such that $y_j \notin Z$. Define $\bar{Z} := Y \backslash (Z \cup \{j\}))$. Compute the causal Wiener filter estimating $y_j(t)$ from $y_j|_{-\infty}^{t-1}$, $y_Z|_{-\infty}^{t}$ and $y_{\bar{Z}}|_{-\infty}^{t-1}$. Let $W_{ji}(z)$ be the component associated with y_i. If $W_{ji}(z)$ has no strictly proper component, we say that y_i and y_j are strictly proper-Wiener-separated by Z. We write such a relation as $cwsep^-(y_i, y_j|Z)$*

Observe that Definition 7 and Definition 8 describe relations among stochastic processes in an LDIM that only depend on their statistical/probabilistic properties which can be tested just from data. A fundamental property of LDIMs is that d-separation in the hybrid graph implies causal-Wiener-separation [17].

Theorem 1. *Let $\mathcal{G}$ be an LDIM with static excitations, graphical representation given by G and associated hybrid graph G^h. Then*

$$dsep_{G^h}(y_i(t), y_j(t)|Z) \implies cwsep^+(y_i, y_j|Z)$$
$$dsep_{G^h}(y_i^-, y_j(t)|Z) \implies cwsep^-(y_i, y_j|Z).$$

Proof: The proof is an immediate consequence of the application of Theorem 24 in [17] to the hybrid graph G_h.

If in an LDIM the previous notions of Wiener-separation implies d-separation in its hybrid graph G^h, then we say that the LDIM is faithful to G^h (or equivalently to the graphical representation G associated to G^h).

Definition 9 (Faithfulness)**.** *Let $\mathcal{G}$ be an LDIM with static excitations, graphical representation given by G and associated hybrid graph G^h. We say that $\mathcal{G}$ is faithful to G^h (or equivalently to the graphical representation associated with G) if*

$$cwsep^+(y_i, y_j|Z) \implies dsep_{G^h}(y_i(t), y_j(t)|Z)$$
$$cwsep^-(y_i, y_j|Z) \implies dsep_{G^h}(y_i^-, y_j(t)|Z).$$

As recently shown in [18], for a given structure G, if we consider a parameterization of all LDIMs with perfect graphical representation G, we find that only a zero measure set of the parameter space leads to an unfaithful LDIM. Hence, faithfulness can be considered as a non-restrictive technical assumption that simply prevents the occurrence of pathological cases. In the set of all hybrid graphs, we can introduce a notion of equivalence.

Definition 10 (Class of Equivalent Hybrid Graphs)**.** *We say that hybrid graph G_1^h is equivalent to hybrid graph G_2^h if and only if they have the same skeleton and the same unshielded colliders. All G_i^hs that are equivalent belong to an equivalence class of hybrid graphs.*

Since the hybrid graph and its associate graphical representation can be obtained from each other, we can extend the definition of equivalence to graphical representations. Thus, two multi-arrowed graphical representation are equivalent if and only if the hybrid graphs associated with them belong to a class of equivalent hybrid graphs. Figure 2 illustrates hybrid graphical representations of three different LDIMs $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_3$.

The following is a minimalistic example that shows that the problem of recovering the graphical representation of an LDIM from observational data is an ill-posed problem.

Example 1. *Consider the three LDIMs $\mathcal{G}_a = (H_a(z), e_a)$,*

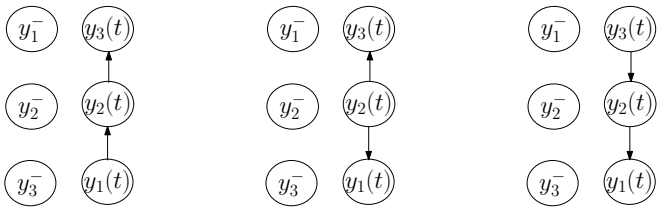


Fig. 2: *Perfect hybrid graphical representations of three LDIMs mentioned in Example 1. The graphs posses the same skeleton and unshielded collider structures. Thus, they belong to the a class of equivalent hybrid graphs.*

$\mathcal{G}_b = (H_b(z), e_b)$, and $\mathcal{G}_c = (H_c(z), e_c)$ defined as follows.

$$H_a(z) = \begin{bmatrix} 0 & 0 & 0 \\ \frac{\sqrt{2}}{2} & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & 0 \end{bmatrix}, \Phi_{e_a} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}$$

$$H_b(z) = \begin{bmatrix} 0 & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & 0 \end{bmatrix}, \Phi_{e_b} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix}$$

$$H_c(z) = \begin{bmatrix} 0 & \frac{\sqrt{2}}{2} & 0 \\ 0 & 0 & \frac{\sqrt{2}}{2} \\ 0 & 0 & 0 \end{bmatrix}, \Phi_{e_c} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Their hybrid representations are illustrated in Figure 2.

Observe that the three output signals y_a, y_b and y_c have all identical statistics. Indeed their power spectral densities are

$$\Phi_{y_a} = \Phi_{y_b} = \Phi_{y_c} = \begin{bmatrix} 1 & \frac{\sqrt{2}}{2} & \frac{1}{2} \\ \frac{\sqrt{2}}{2} & 1 & \frac{\sqrt{2}}{2} \\ \frac{1}{2} & \frac{\sqrt{2}}{2} & 1 \end{bmatrix}$$

Example 1 proves that we can not differentiate between the three structures of Figure 2 based on observational data only since the observed signals could have been generated by an LDIM with any of those three representations.

V. Main results

In the area of graphical models, there are two large classes of methods to infer the structure of a network from data. A first approach is given by constrained based algorithms. Constrained based algorithms use statistical tests to check conditional independence relations between different nodes and on the basis of the results of such tests they determine the presence or absence of individual edges in the graph. The main drawback of such methods is that they rely on a large number of statistical tests and, hence, tend to have a high sample complexity [19]. Another approach is given by score-and-search methods [20]. This second class of methods tries to learn the structure of the underlying graph by maximizing a score function that represents how well the graph fits the observed data. While these methods do not rely on a large number of statistical tests, they need to perform an optimization over the space of all graphs which has a combinatorial complexity in the number of nodes. Thus, these methods offer a lower sample complexity at the price of a higher computational complexity. The scientific literature about structure reconstruction for networks of dynamic systems has mostly focused on constrained based algorithms [9] [12] and there is a smaller number of results employing score-and-search methods [5][10].

In this paper, we provide consistency results that hold in general scenarios which include the presence of direct feedthroughs and feedback loops. Specifically, we are going to prove two results

- Given an LDIM $\mathcal{G}$, faithful to its perfect graphical representation G^*, it is possible to define a likelihood-based score that is maximized by G^* when the time horizon of the observations goes to infinity.
- All graphs maximizing such a likelihood-based score are equivalent to G^* according to Definition 10.

As an immediate consequence of these two results, we obtain that we can consistently recover the equivalence class of the perfect graphical representation of an LDIM.

A. Score definition

The choice of the score is key to achieve a good performance in learning the structure of an LDIM. One intuitive choice is the likelihood function. However, the main problem of the likelihood function is that it tends to overfit the model by selecting denser graphs: indeed, for example, even if we have two independent nodes, the sample correlation between the two associated data streams would never be exactly zero and the likelihood score would be maximized by assuming a connection between the two nodes. To obviate this effect, a common option is to consider the standard likelihood score and modify it by introducing a term penalizing the number of edges in the evaluated graph. In particular, in order to introduce a penalty term we define the dimension of a graphical representation G.

Definition 11 (Dimension of graphical representation of an LDIM)**.** *Given an LDIM with graphical representation $G = (V, E_1, E_2)$, we define the dimension of G as $dim(G) := |E_1| + |E_2|$ where $|E_1|$ and $|E_2|$ are the cardinalities of the edge sets E_1 and E_2.*

A score that is structured as the standard likelihood with a penalty term is the Bayesian Information Criteria (BIC)[21], which we can adapt to our scenario with $dim(G)$ as a penalty term in the following way

$$BIC(\mathcal{D}, G) = \frac{1}{M} 2 \ln(\ell_G(\mathcal{D}; \hat{\theta})) - \frac{\ln(M)}{M} dim(G), \quad (1)$$

where, $\mathcal{D}$ are the observed time series data over a time horizon of length M, G is the representation of an LDIM $\mathcal{G}$ with transfer matrix $H(z)$ parametrized by θ, $\ell_G(\mathcal{D}; \hat{\theta})$ is the maximum likelihood function given by

$$\ell_G(\mathcal{D}; \hat{\theta}) := \max_{\theta} \ell_G(\mathcal{D}; \theta),$$

and $\hat{\theta}$ is the maximum likelihood estimation of θ.

In a way analogous to graphical models, it can be shown that maximum log-likelihood function can be written in the following way for LDIMs [22][13].

$$\ln\left(\ell_G(\mathcal{D}; \hat{\theta})\right) = M \sum_{i=1}^{n} \mathbf{I}(y_i; Pa_G(y_i)) + M \sum_{i=1}^{n} \mathbf{H}(y_i) \quad (2)$$

Where $\mathbf{I}$ is the mutual information rate and $\mathbf{H}$ is the entropy rate of stochastic processes.

The following theorem proves that when the time horizon of the observations M goes to infinity, the perfect graphical representation of the LDIM maximizes the BIC score.

Theorem 2. *Given an LDIM with static excitations and faithful to its perfect graphical representation G^*, the BIC score is consistently maximized by G^* (that is, when the time horizon of the observed time series goes to infinity, G^* maximizes the BIC score).*

Proof: The proof follows identical steps as in [13] (Chapter 18, Theorem 18.2), but it is applied to LDIMs which have stochastic processes as nodes instead of graphical models where nodes represent simple random variables. Thus, instead of using mutual information, our proof uses mutual information rate. Let G^* be the perfect graphical representation of the LDIM $\mathcal{G}^* = (H^*(z), e^*)$ with output y^*. Let G be the graph of the LDIM $\mathcal{G}$ with output y for which we are computing the BIC score. Let also θ be the set of parameters parametrizing the transfer function $H(z)$ and the noise variance matrix of e. In the expression of likelihood of Equation 2 the term $M\sum_{i=1}^{n}\mathbf{H}(y_i^*)$ is the same for any choice of the graph G and hence it can be neglected in the following arguments. If no choice of θ is such that $\Phi_y(z)$ matches $\Phi_{y^*}(z)$, then we have

$$\sum_{i=1}^{n}\mathbf{I}(y_i^*; Pa_{G^*}(y_i^*)) > \sum_{i=1}^{n}\mathbf{I}(y_i^*; Pa_G(y_i^*)).$$

Let us define

$$\Delta := \sum_{i=1}^{n}\mathbf{I}(y_i^*; Pa_{G^*}(y_i^*)) - \sum_{i=1}^{n}\mathbf{I}(y_i^*; Pa_G(y_i^*)) > 0.$$

Then, for $M \to \infty$, we have

$$\begin{aligned} &BIC(G^*) - BIC(G) \simeq \\ &\quad = 2\Delta - \frac{\ln(M)}{M}\left(Dim(G^*) - Dim(G)\right). \end{aligned}$$

Hence, for $M \to \infty$, the score associated with G^* is larger than the score associated with G. If instead, the parameterization of the graph G can reproduce the PSD matrix $\Phi_y(z)$, then we have

$$\sum_{i=1}^{n}\mathbf{I}(y_i^*; Pa_{G^*}(y_i^*)) = \sum_{i=1}^{n}\mathbf{I}(y_i^*; Pa_G(y_i^*)).$$

In this case, for $M \to \infty$, we get

$$BIC(G^*) - BIC(G) \simeq -\frac{\ln(M)}{M}\left(Dim(G^*) - Dim(G)\right).$$

Thus, the higher score is associated with the graph with smaller dimension. Since G^* is the perfect graphical representation and $\mathcal{G}^*$ is faithful to G^*, necessarily $dim(G^*)$ is smaller or equal to $dim(G)$.

Instead, the following theorem complements Theorem 2 by showing that, when the time horizon of the observations M goes to infinity, the only graphs maximizing the BIC score are the ones equivalent to G^*.

Theorem 3. *Given the LDIM $\mathcal{G}$ which is faithful to its perfect graphical representation G^*, for $M \to \infty$, all graphs that maximize the BIC score are equivalent to G^*.*

Proof: First we want to show that if G is equivalent to G^*, then the BIC scores are the same for $M \to \infty$. Let G be equivalent to G^* and let G^h and G^{*h} be their associated hybrid graphs. Consider all the conditional independence relations in G^{h*} and let us denote them as I^*. Consider the ordering σ given by G^h and the set of all relations I^*. By applying Pearl-Verma theorem [23] to σ and I^*, we get the graph $G^{h'}$. Now, consider all the d-separation relations in G^h and let us denote them as I. Trivially, by applying Pearl-Verma theorem to σ and I, we get G^h. Since G^h and G^{*h} are equivalent and $\mathcal{G}$ is faithful to G^*, then I^* and I are the same. Hence G^h and $G^{h'}$ are the same, as well. Furthermore, there is a set of parameters for an LDIM $\mathcal{G}$ with graphical representation G that has the same outputs of $\mathcal{G}^*$. Thus, the graph G has the same score as G^* for $M \to \infty$.

Now, we want to show that if G is not equivalent to G^*, then its BIC score is strictly less than the score of G^*. As shown in the proof of Theorem 2, if there no set of transfer functions such that an LDIM $\mathcal{G}$ with graphical representation G has the same power spectral densities as the outputs of G^*, then the BIC score of G is strictly less than the score of G^*, for $M \to \infty$. If G is not equivalent to G^*, but there are transfer functions such that the outputs of $\mathcal{G}$ have the same power spectral densities as in $\mathcal{G}^*$, then all the d-separation relations in G are also d-separation relations in G^* because of faithfulness. Hence all edges in the skeleton of G^* are in G (see Theorem 1 in [24]). However, since G is not equivalent to G^*, there is at least a d-separation relation in G^* that is not in G, implying that there is an edge in G that is not in G^*. Hence $dim(G) > dim(G^*)$ given a lower score for G than for G^*.

VI. Numerical Experiments

We used the benchmark model described in [9] to test the consistency properties of the score-and-search method proposed in this article. The simulated LDIM was defined by

$$H(z) = \begin{bmatrix} 0.4\frac{1}{z} & 0 & 0 & a_{14}\frac{1}{z} \\ 0 & 0.2\frac{1}{z} & 0 & a_{24} \\ a_{31}(1+\frac{1}{z}) & a_{32}\frac{1}{z} & 0 & 0 \\ a_{41} & 0 & a_{43} & 0 \end{bmatrix}$$

$$\Phi_e = \begin{bmatrix} 0.4 & 0 & 0 & 0 \\ 0 & 0.3 & 0 & 0 \\ 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0.3 \end{bmatrix}$$

with graphical representation depicted in Figure 3. It can be verified that such a graphical representation is the only element in its equivalence class. Thus, in this specific case, the problem of topology reconstruction is well-posed. We have simulated the network for different time horizons (M from 50 to $10,000$) and for each time horizon we have run $2,500$ simulations. Then we have used the data and applied a maximum likelihood approach to estimate the network parameters. We defined graph error rate as the estimated probability that a detected edge is actually in the graphical representation plus the estimated probability that an undetected edge is not in the graphical representation and divided the result by 2. Thus, the graph error rate is a score in the interval $[0, 1]$ where 0 represents a flawless reconstruction. The

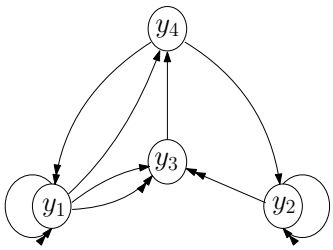
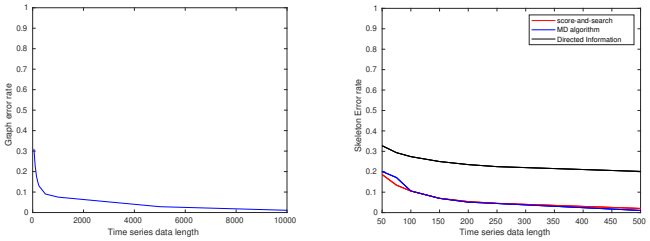


Fig. 3: *Graphical representation of the benchmark model of Section VI.*



(a) (b)

Fig. 4: *(a) visualizes the graph error rate as a function of the data quantity M. The error rate converges to 0 for $M \to \infty$. (b) compares the error rate of determining the skeleton for the score-and-search method of this article and the techniques in [9] and [12].*

results are reported on Figure 4(a) validating the consistency results of Theorem 2. Using a similar simulation scheme, we wanted to compare the performance of our method with the techniques in [9] and [12]. Those techniques do not reconstruct the multi-typed graphical representation that we introduced in this article. In order to have a meaningful comparison, we decided to only evaluate the performance of all the methods considering only the skeleton of the reconstructed topology, since this is information common to all the techniques. As alluded in the introduction, the main reason to adopt a score-and-search method is to improve the performance when the observation horizon is short. Thus, we only considered time series of length M in the range $[50, 500]$. We defined the skeleton error rate as the estimated probability that a detected edge is actually in the skeleton of the graphical representation plus the estimated probability that an undetected edge is not in skeleton of the graphical representation and divided the result by 2. The results are reported on Figure 4(b). We notice that for shorter time series, the score-and-search method is the best performing, however once the time series become long enough, the constrained based method of [9] performs better.

VII. Conclusion

The main contribution of this paper is extending score-and-search methods from the area of graphical models to the reconstruction of the underlying graph of dynamic networks. We adapted the Bayesian Information Criterion (BIC) in order to consistently reconstruct the topology of a class of linear networks. In the general case, it is shown that multiple topologies are compatible with the data. In such a case all such topologies are proven to maximize the BIC score recovering a well-defined class of equivalence of graphical representations. Numerical experiments validate the theoretical results and indicate that this method is more accurate than other methodologies when the data is scarce.

References

[1] Scutari, M. and Nagarajan, R., "Identifying significant edges in graphical models of molecular networks," *Artificial intelligence in medicine*, vol. 57, no. 3, pp. 207–217, 2013.

[2] Giudici, P. and Spelta, A., "Graphical network models for international financial flows," *Journal of Business & Economic Statistics*, vol. 34, no. 1, pp. 128–138, 2016.

[3] Ebert-Uphoff, I. and Deng, Y., "Causal discovery for climate research using graphical models," *Journal of Climate*, vol. 25, no. 17, pp. 5648–5665, 2012.

[4] Gonçalves, J. and Warnick, S., "Necessary and sufficient conditions for dynamical structure reconstruction of lti networks," *IEEE Transactions on Automatic Control*, vol. 53, no. 7, pp. 1670–1674, 2008.

[5] Materassi, D. and Innocenti, G., "Topological identification in networks of dynamical systems," *IEEE Transactions on Automatic Control*, vol. 55, no. 8, pp. 1860–1871, 2010.

[6] Yuan, Y., Stan, G.-B., Warnick, S., and Goncalves, J., "Robust dynamical network structure reconstruction," *Automatica*, vol. 47, no. 6, pp. 1230–1235, 2011.

[7] Materassi, D. and Salapaka, M. V., "Reconstruction of directed acyclic networks of dynamical systems," in *2013 American Control Conference*. IEEE, 2013, pp. 4687–4692.

[8] Sepehr, F. and Materassi, D., "Blind learning of tree network topologies in the presence of hidden nodes," *IEEE Transactions on Automatic Control*, vol. 65, no. 3, pp. 1014–1028, 2019.

[9] Dimovska, M. and Materassi, D., "A control theoretic look at granger causality: extending topology reconstruction to networks with direct feedthroughs," *IEEE Transactions on Automatic Control*, vol. 66, no. 2, pp. 699–713, 2020.

[10] Shi, S., Bottegal, G., and Van den Hof, P. M., "Bayesian topology identification of linear dynamic networks," in *2019 18th European Control Conference (ECC)*. IEEE, 2019, pp. 2814–2819.

[11] Materassi, D. and Salapaka, M. V., "On the problem of reconstructing an unknown topology via locality properties of the wiener filter," *IEEE transactions on automatic control*, vol. 57, no. 7, pp. 1765–1777, 2012.

[12] Quinn, C. J., Coleman, T. P., Kiyavash, N., and Hatsopoulos, N. G., "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.

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

[14] Ghahramani, Z., "Learning dynamic bayesian networks," *International School on Neural Networks, Initiated by IIASS and EMFCSC*, pp. 168–197, 1997.

[15] Van den Hof, P. M., Dankers, A., Heuberger, P. S., and Bombois, X., "Identification of dynamic models in complex networks with prediction error methods—basic methods for consistent module estimates," *Automatica*, vol. 49, no. 10, pp. 2994–3006, 2013.

[16] Shafie, T., "A multigraph approach to social network analysis." *Journal of Social Structure*, vol. 16, 2015.

[17] Materassi, D. and Salapaka, M. V., "Signal selection for estimation and identification in networks of dynamic systems: a graphical model approach," *IEEE Transactions on Automatic Control*, vol. 65, no. 10, pp. 4138–4153, 2019.

[18] Dimovska, M. and Materassi, D., "Granger-faithfulness and link orientation in network reconstruction," *IEEE Transactions on Control of Network Systems*, pp. 1–1, 2022.

[19] Abbeel, P., Koller, D., and Ng, A. Y., "Learning factor graphs in polynomial time and sample complexity," *The Journal of Machine Learning Research*, vol. 7, pp. 1743–1788, 2006.

[20] Glymour, C., Zhang, K., and Spirtes, P., "Review of causal discovery methods based on graphical models," *Frontiers in genetics*, vol. 10, p. 524, 2019.

[21] Schwarz, G., "Estimating the dimension of a model," *The annals of statistics*, pp. 461–464, 1978.

[22] Lasserre, M., Lebrun, R., and Wuillemin, P.-H., "Learning continuous high-dimensional models using mutual information and copula bayesian networks," in *The Thirty-Fifth AAAI Conference on Artificial Intelligence*, vol. 35, no. 13, 2021, pp. 12 139–12 146.

[23] Verma, T. and Pearl, J., "Causal networks: Semantics and expressiveness," in *Machine intelligence and pattern recognition*. Elsevier, 1990, vol. 9, pp. 69–76.

[24] ——, "Equivalence and synthesis of causal models," in *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, 1990, pp. 255–270.