2017 IEEE 56th Annual Conference on Decision and Control (CDC)
December 12-15, 2017, Melbourne, Australia

Granger-Causality Meets Causal Inference in Graphical Models: Learning Networks via Non-Invasive Observations

Mihaela Dimovska & Donatello Materassi (mdimovsk@vols.utk.edu & dmateras@utk.edu)
Department of Electrical Engineering and Computer Science,
University of Tennessee, Knoxville * (1520 Middle Dr, 37996, Knoxville TN)

Abstract— **Algorithms developed in the area of graphical models have proven to be useful tools for the analysis and reconstruction of networks of dynamic systems with directed acyclic structure. However, such techniques cannot typically provide a consistent reconstruction for networks of dynamic systems in presence of feedback mechanisms. On the other hand, reconstruction techniques based on causal estimators and one-step-ahead predictors, such as Granger causality, are capable of reconstructing loopy networks, but only if all the operators defining the input/output structure of the network are strictly causal. In this work, we develop a novel reconstruction method for network topologies that combines the advantages of graphical models and causal estimator techniques unifying both approaches under a single framework. The fundamental result is a generalization of Granger causality which provides a consistent reconstruction of the topology of a network of linear dynamic systems under the milder assumption that every loop contains at least one strictly causal transfer function.**

I. Introduction

Networked and modular systems have become an irreplaceable modeling tool in areas as diverse as biology [1], [2], neuroscience [3], economics [4], political science [5], [6], and social networks [7]. In many application scenarios, direct measurements of the dynamics of the individual components of the network are available, and the objective of the investigation is to infer how these components affect and influence each other, often representing these relations with a graph. There are a large number of methodologies that obtain a graph structure from experimental data. Some of these are based on heuristics, thus they do not typically provide any form of guarantees about the correctness of the reconstruction [8], [9], [10], [11], [12]. Other methods guarantee the reconstruction of relatively large classes of networks, but rely on the possibility of having access to, or directly manipulating, the inputs of the system under investigation [13], [14], [15], [16], [17], [18]. A fundamentally more challenging situation, that is also the main focus of this work, arises when data are only collected via non-invasive observations, namely when the system is operating under standard conditions, input signals are unknown, and it is not possible or just too impractical to manipulate them. Several algorithms have been developed to learn the structure of a network from non-invasive observations. These algorithms originated mostly from the area of graphical models [19], [20], [21]. Only more recently many of these techniques have been ported and extended to the domain of stochastic processes to detect input/output relations among dynamic systems [22], [23]. Indeed, even though graphical models of random variables and networks of dynamic systems have significantly different underlying semantics (i.e. graphical models typically consider static operators), it has been shown that many learning techniques can be applied to both [24].

One of the most versatile algorithms to learn a network is Peter-Clark (PC) algorithm which provides a consistent reconstruction of a graphical model of random variables with directed acyclic structure [20]. It has been shown that a variation of PC algorithm can be implemented to reconstruct networks of dynamic systems, but this variation mantains the limitation of not being able to cope with structures with directed loops [25]. In the context of networks of dynamic systems this can be a severe limitation since network structures with feedback loops are inherent in many large scale systems. However, networks of dynamic systems, as opposed to graphical models of random variables, tend to have a richer structure, allowing for different approaches. Indeed, there are several methodologies for network reconstruction that are based on causal estimators and one-step-ahead predictors. Among these methodologies, techniques making use of Granger causality [26] or directed information [27] are definitely prevalent. These techniques tend to rely on the basic fact that causes precede their effects and assume that data are sampled at a frequency sufficient to capture such delays [28], [29]. For example, [24] shows that if delays can be detected on each link of a network with linear dynamics, Granger causality consistently reconstructs the network topology. Analogous results have been proven in the nonlinear case [30]. Compared to graphical model methods, estimator methods can deal with directed loops, but cannot cope with data sampled at a frequency where propagation delays between connected nodes become unobservable.

In this article we derive a methodology for the reconstruction of networks of dynamic systems that can be considered a successor of both PC algorithm and Granger causality, since it subsumes the two methods into one which, overall, requires weaker assumptions to guarantee a consistent reconstruction. Indeed, the method is shown to consistently reconstruct a network of linear dynamic systems under the much milder assumption that every directed loop in the network contains at least one strictly causal transfer function. To our best knowledge, there is no other methodology that can guarantee a consistent reconstruction of a comparably large class of networks.

The article is organized as follows: in Section II we provide preliminary notions and introduce Linear Dynamic

*This work is partially supported by NSF (CISE/CAREER:1553504).

Influence Models (LDIMs), the class of networks which is the focus of our work; in Section III we present the problem of reconstructing the topology of a network, describing the limitations of current methods; in Section IV we provide the main result that consists of a method allowing the exact reconstruction of the topology of a LDIM where each loop contains at least one strictly causal transfer function.

II. Background notions

In this section we recall some fundamental notions of graph theory which are functional to the subsequent developments and then introduce the class of Linear Dynamic Influence Models (LDIMs) that will be the focus of the network reconstruction methodology of the main section.

A. Basic notions of graph theory

We recall the definition of undirected and directed graphs.

Definition 1 (Directed and Undirected Graphs): A directed (undirected) graph G is a pair (V, E) where V is a set of vertices or nodes and E is a set of edges or arcs, which are ordered (unordered) subsets of two distinct elements of V.

It is possible to associate an undirected graph to any directed graph by removing the orientation of its links.

Definition 2 (Skeleton): Given a directed graph $G = (V, E)$, we define its skeleton as the undirected graph $(V, \overline{E})$ obtained by removing the orientation of its edges.

In a directed graph we define paths and chains.

Definition 3 (Paths and chains): In a directed (undirected) graph G a path from node j to i is an ordered sequence of distinct contiguous edges (edges with one node in common) connecting j and i. If the edges have all the same orientation the path is called a chain.

On a directed graph we specify the notions of ancestry and descendance.

Definition 4 (Parents, children, ancestors, descendants): Consider a graph $G = (V, E)$. A vertex j is a *parent* of a vertex i if there is a directed edge from j to i. In such a case i is a *child* of j. Also j is an *ancestor* of i if there is a chain from i to j. In such a case j is a *descendant* of i.

B. Linear Dynamic Influence Models

We now define the class of stochastic processes used in the development of our theoretical framework.

Definition 5 (Rationally Related Random Processes): Let $\mathcal{E}$ be a set containing discrete-time scalar, zero-mean, jointly wide-sense stationary random vector processes such that, for any $e_i, e_j \in \mathcal{E}$, the power spectral density $\Phi_{e_ie_j}(z)$ exists, is real-rational with no poles on the unit circle and given by $\Phi_{e_ie_j}(z) = A(z)^{-1}B(z)$, where $A(z)$ and $B(z)$ are matrix polynomials with real coefficients such that $det(A(z)) \neq 0$ for any $z \in \mathbf{C}$, with $|z| = 1$. Then, we say that $\mathcal{E}$ is a set of rationally related random processes.

Definition 6: The set $\mathcal{F}$ is defined as the set of real-rational transfer matrices that are analytic and invertible on the unit circle $\{z \in \mathbf{C}| \quad |z| = 1\}$. Given a transfer matrix $H(z) \in \mathcal{F}$, it can be represented in the frequency domain as $H(z) = \sum_{k=-\infty}^{\infty} h_k z^{-k}$. If$h_k = 0$ for every $k < 0$, then we say that the transfer matrix is causal, and in addition if $h_k = 0$ for $k = 0$ we say the transfer matrix is strictly causa. We define the space of causal transfer functions as $\mathcal{F}^+$.

Definition 7: As a notation we define $y = H(z)e$ as the process obtained by filtering e with $H(z)$, namely $y(t) = \sum_{-\infty}^{\infty} h_{t-k} e(k)$ for all $t \in \mathbb{Z}$. For example, $\frac{1}{z}e$ denotes the process e delayed by one time step.

It is possible to filter rationally related processes with transfer functions in $\mathcal{F}^+$ to obtain a space of rationally related processes.

Definition 8 (Causal transfer function spans): Given a set $\mathcal{E}$ of rationally related random processes, we define the causal transfer function span as

$$\text{c-tf-span}\{\mathcal{E}\} := \mathcal{F}^+\mathcal{E} = \left\{ y = \sum_k H_k(z) e_k | e_k \in \mathcal{E}, H_k \in \mathcal{F}^+ \right\}.$$

The fact that $\mathcal{F}^+\mathcal{E}$ is a space of rationally related processes is an immediate consequence of the Wiener-Khinchin Theorem (see [31]). Also, $\mathcal{F}^+\mathcal{E}$ equipped with the inner product

$$\left\langle y_i, y_j \right\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \Phi_{y_iy_j}(e^{i\omega}) d\omega$$

is a pre-Hilbert space [32]. Such inner product induces the norm $||y||^2 = \langle y_i, y_i \rangle$ in the usual way.

The following definition provides a class of models for a network of dynamical systems. It is assumed that the dynamics of each agent (node) in the network is represented by a vector random process $\{y_i\}_{i=1}^n$ that is given by the superposition of an autonomous behavior e_j and the "influences" of some other "parent nodes" through dynamic links. The autonomous behaviour on each node is assumed independent of the others. If a certain agent "influences" another one a directed edge can be drawn and a directed graph can be obtained.

Definition 9 (Linear Dynamic Influence Models): A Linear Dynamic Influence Models $\mathcal{G}$ is defined as a pair $(H(z), e)$ where

- $e = (e_1|...|e_n)^T$ is a vector of N rationally related random vector processes e_1, ... e_n of dimensions $n_1, ..., n_n$ respectively with $N = n_1 + ... + n_n$, such that $\Phi_e(z)$ is block diagonal, namely $\Phi_{e_ie_j} = 0$ for $i \neq j$. The vector of positive integers $(n_1, ..., n_n)$ is the "partition" of the LDIM.
- $H(z)$ is a $N \times N$ transfer matrix partitioned in $n \times n$ blocks such that H_{ji} is of dimension $n_j \times n_i$ for $i, j \in \{1, ..., n\}$. $H(z)$ is termed as the "dynamics" of the LDIM.

The output vector 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 in a more compact way

$$y(t) = e(t) + H(z)y(t) \tag{1}$$

where $y = (y_1|...|y_n)^T$.

Definition 10: (Graph associated to a LDIM) Let $\mathcal{G} = (H, e)$ be a LDIM with output processes $y_1, ..., y_n$ Let $V := \{1, ..., y_n\}$ and let E be a subset of $V \times V$ such that $(i, j) \notin E$ implies $H_{ji} = 0$. We say that the graph $G = (V, E)$ is a graphical representation of the LDIM. Furthermore, if $G = (V, E)$ is a graphical representation of the LDIM and $(i, j) \in E$ implies $H_{ji} \neq 0$, we say that $G = (V, E)$ is an exact graphical representation of the LDIM. By extension, when considering a specific graphical representation of a LDIM, nodes and edges of a LDIM will mean nodes and edges of the specific graphical representation.

Observe that the graphical representation of a LDIM provides information about blocks of the matrix $H_{ji}(z)$ that are identically zero. Indeed, if $(i, j) \notin E$ then $H_{ji}(z)$ is definitely zero. However, unless it is known that the graphical representation is exact, the presence of (i, j) in E has to be interpreted as $H_{ji}(z)$ is "potentially different" from zero.

Definition 11: A LDIM $\mathcal{G}$ is well-posed if $\mathcal{I} - H_{II}$ is causally invertible for every set of indeces $I \subseteq \{1, ..., n\}$, where H_{II} is the submatrix of H obtained by selecting the rows and columns with indeces in I.

Definition 12: A LDIM $\mathcal{G}$ is topologically detectable if $\Phi_{e_i e_i}(e^{i\omega}) > 0$ for any $\omega \in [-\pi, \pi]$ and for any $i = 1, ..., n$.

III. The Problem of Reconstructing a Network

In this article we focus on the following problem.

Skeleton Reconstruction Problem: From the observations of the nodes $y_1, ..., y_n$ of a LDIM G, determine the skeleton of the graph associated with G.

We review two main approaches towards this problem through the lens of Wiener filtering. One approach is a variation of the PC algorithm for dynamic systems [25], and the other is Granger causality [32]. We first define the notion of conditional uncorrelatedness for stochastic processes.

Definition 13 (Wiener Uncorrelated Processes): Let $y_1, ...y_n$ be processes in the space $\mathcal{F}^+\mathcal{E}$. Let W_{ji} be entry of the Wiener filter associated with y_i when estimating y_j from $\{y_i\} \cup S$, where $S \subseteq \{y_1, ..., y_n\} \setminus \{y_i, y_j\}$. If W_{ji} is zero, we say that y_j is Wiener uncorrelated with y_i given S. We also say that then S Wiener-separates y_i and y_j.

A variation of the PC algorithm that reconstructs the exact topology when the underlying network is acyclic is formulated in [32]. The enabling result for the algorithm in [32] is the follwing proposition.

Proposition 3.1: (*Consistent reconstruction of Acyclic Networks*): Let G be a well-posed and topologically detectable acyclic LDIM with output processes $Y = \{y_1, ...y_n\}$. If y_i and y_j are not directly connected in the LDIM, then there exists a set of nodes $S \subseteq Y$ such that given S, y_i and y_j are Wiener-uncorrelated.

As observed in [33], Proposition 3.1 is not strictly a necessary and sufficient condition, but only in pathological situations. When these pathological situations do not occur, the LDIM is said to be faithful. Thus, if the LDIM is acyclic, under the mild assumption of faithfulness, Proposition 3.1 enables to infer the LDIM skeleton by removing edges between pairs of nodes that are Wiener separated by an appropriate S. This method is analogous to the standard PC-algorithm for random variables that uses the notion of independence [34] instead of Wiener-uncorrelatedness.

A different approach makes use of Granger causality. The intuition behind Granger causality is that a cause and its effect can be distinguished, by knowing that a cause precedes its effect [26]. This intuition is translated in the following result, derived in [32].

Proposition 3.2: (*Consistent Reconstruction of Strictly Causal Networks*): Consider a well-posed and topologically detectable LDIM $(H(z), e)$ where all entries of $H(z)$ are strictly causal. Let the output processes of the LDIM be $\{y_1, ...y_n\}$ and define the space Y = c-tf-span$(\frac{1}{z}y_1, ..., \frac{1}{z}y_n)$. Consider the problem of estimating y_j using an element $\hat{y}_j \in Y$:

$$\min_{\hat{y}_j \in Y} \left\| y_j - \hat{y}_j \right\|^2 .$$

Then the optimal solution $\hat{y}_j$ exists, is unique and is given by $\hat{y}_j = \sum_{i=1}^n W^g_{ji}(z)\frac{1}{z}y_i$, where W^g_{ji} denotes the causal Wiener filter that estimates y_j from $\frac{1}{z}y_i$. Further, for $i \neq j$, $W^g_{ji} \neq 0$ is necessary and sufficient condition for y_i being a parent of y_j. If $W^g_{ji} = 0$ and $W^g_{ij} = 0$, we say that y_i and y_j are Granger uncorrelated.

Since W^g_{ji} can be computed from the power spectral densities of $y_{k_{k=1}^{n}}$, Proposition 3.2 provides an effective way to learn the topology of a LDIM from non-invasive data. The two approaches above manage to consistently reconstruct very specific classes of networks: directed acyclic graphs and strictly causal dynamic networks respectively. We now demonstrate an example where both of these approaches fail to reconstruct the true skeleton of a network that belongs in the class of networks that we are considering, in which every cycle has at least one delay.

Example 3.3: Let G be a LDIM with exact graphical representation as in Figure 1, where every transfer function is different than zero. Let H_{11}, H_{14} and H_{32} be strictly causal, and let the other transfer functions be causal but not strictly causal. Observe that the nodes y_1 and y_2 are not directly connected, thus a method for the consistent reconstruction of the network topology should be able to identify the absence of a link between y_1 and y_2. The variation of PC algorithm for dynamic systems fails in this respect, since it can be shown that no set S makes the processes y_1 and y_2 Wiener uncorrelated as the underlying network is cyclic. Also, y_1 and y_2 are not Granger uncorrelated: $y_1(t-\tau)$ influences $y_2(t)$ for some $\tau > 0$ through the path $y_1 \to y_3 \to y_4 \to y_2$ because the transfer function H_{11} is strictly causal, while H_{31}, H_{43}, H_{24} are not strictly causal. As a result, neither the method in [25] nor Granger causality are able to correctly determine the skeleton of this network.

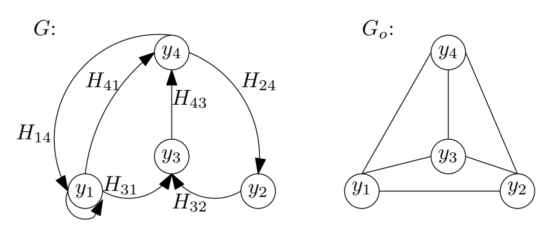

Fig. 1. A LDIM G with one delay in each cycle, and the incorrect skeleton G_o that both PC and Granger method infer.

This example motivates our investigation: can we combine the notion of causality that is at the base of the PC algorithm with the notion of Granger causality under a single framework and obtain a method that is capable of reconstructing a larger class of networks? In the next section we answer this question in the positive.

IV. Main Results

We now present two tests that are instrumental for the definition of out main result: a methodology capable of subsuming and extending PC algorithm and Granger causality.

The first result shows that if two time series variables are independent at time t, then there is a set that makes them Wiener-uncorrelated. The second result allows to consistently determine if a time-series "causes" another without requiring all operators to be strictly causal, as in the case of Granger causality.

Theorem 4.1 (Separation in the present): Consider a well-posed, causal and topologically detectable LDIM with output signals $Y = \{y_1, ..., y_n\}$, such that in every directed loop there is at least one strictly causal transfer function. Let y_i and y_j be such that the transfer functions H_{ji} and H_{ij} are strictly causal. There exists a set of processes $S = S^+ \cup S^-$, with $S \subseteq Y \setminus \{y_i\}$ and $S^- \subseteq \frac{1}{z}Y$, such that when y_j is estimated from $S \cup \{y_i\}$ the component W_{ji} of the Wiener filter is strictly causal.

Proof: The proof is reported in the Appendix. ■

Intuitively, the above statement says that if the random variables $y_i(t)$ do not directly influence $y_j(t)$ each other, then the component W_{ji} of the Wiener filter, given S, is strictly causal. This condition is not a sufficient and a necessary condition only in pathological cases (when the LDIM is not "faithful"). Thus Theorem 4.1 can be considered a practical tool to detect if the transfer functions H_{ji} and H_{ij} are both strictly causal. However, this is not enough to show that y_i and y_j are not connected, namely that H_{ji} and H_{ij} are both zero. We present a theorem that would complement the detection of edge absence in the graph of a LDIM.

Theorem 4.2 (Separation in the past): Consider a well-posed, causal and topologically detectable LDIM with output signals $Y = \{y_1, ..., y_n\}$, such that in every directed loop there is at least one strictly causal transfer function. Let y_i and y_j be such that the transfer functions H_{ji} and H_{ij} are known to be strictly causal. Then if $H_{ji} = 0$, there exist two disjoint sets $S_c \subseteq Y \setminus \{i, j\}$ and $S_s \subseteq (Y \setminus \{i, j\})$ such that y_j and $\frac{1}{z}y_i$ are Wiener uncorrelated given $S = S_c \cup \frac{1}{z}S_s$.

Proof: The proof is reported in the Appendix. ■

Again, Theorem 4.2 is not a necessary and sufficient condition only in pathological cases (when the LDIM is not "faithful"). Thus, given that H_{ji} is a strictly causal transfer function, we can use Theorem 4.2 as a practical tool to determine is $H_{ji} = 0$ by checking if the component W_{ji} of the Wiener filter that estimates y_j from y_i given the signals in $S = S_c \cup \frac{1}{z}S_s$ These two theorems provide a way to reconstruct the skeleton of a LDIM.

Mixed Delay Algorithm (MD Algorithm)

Test Link Presence(y_i , y_j):
Test 1: test if there exists S^+ and S^- as in Theorem 4.1 that lead to W_{ji} strictly causal
If output of Test 1 is negative:
The link between y_i and y_j is present in the skeleton
Else:
Test 2.1: test if there exist S_c and S_s as in Theorem 4.2 such that $y_j(t)$ is independent of $\frac{1}{z}y_i$ (Extended Granger causality)
If output of Test 2.1 is negative:
The link between y_i and y_j is present in the skeleton
Else:
Test 2.2: test if there exist S_c and S_s as in Theorem 4.2 such that $y_i(t)$ is independent of $\frac{1}{z}y_j$ (Extended Granger causality)
If output of Test 2.1 is negative:
The link between y_i and y_j is present in the skeleton
Else:
The link between y_i and y_j is not present in the skeleton

Theorem 4.3 (Extended Granger Causality): If a LDIM has at least one strictly causal transfer function in each of its loops and is faithful, MD Algorithm consistently reconstructs its skeleton.

Proof: The statement follows as a consequence of Theorem 4.1 and Theorem 4.2. ■

Note that Theorem 4.3 places itself as an extension of Granger causality as it only requires the presence of a delay in each loop instead of a delay in each link in order to guarantee a consisten reconstruction of the skeleton.

V. A numerical Example

We give a numerical example with LDIM as in Fig. (1).

$$
\begin{aligned}
y_1(t) &= 0.4y_1(t-1) + 0.4y_4(t-1) + 0.4e_1(t)\\
y_2(t) &= 0.5y_2(t-1) + 0.4y_4(t) + 0.3e_2(t)\\
y_3(t) &= 0.45y_2(t-1) + 0.6y_1(t) + 0.4e_3(t)\\
y_4(t) &= 0.4y_3(t) + 0.4y_1(t) + 0.3e_4(t)\,.
\end{aligned}
$$

We generated a data sample consisting of 5000 observations. The first step of our algorithm finds $S^c_{12} = \{x_2(t-1), x_4(t)\}$ such that $y_1(t)$ and $y_2(t)$ are separated given S^c_{12}. With the second step we found that $y_2(t)$ is separated from $y_1(t-1)$ given $S^g_{12} = \{y_2(t-1), y_1(t)\}$. Also, as expected, our second step finds that $y_1(t)$ and $y_2(t-1)$ are separated by $S^g_{21} = \{y_1(t-1), y_4(t-1)\}$. Thus we achieve what the state of the art methods of the PC-algorithm and Granger causality could not: we determined the absence of a link between y_1 and y_2 in the network of Figure (1). The full code of the numerical example that shows the entire correct reconstruction of the skeleton of the network can be found at osf.io/gz9dn. .

VI. Conclusions

The article provides an algorithm for the consistent reconstruction of the topology of a network of linear dynamic systems, under the mild assumption that there is at least one strictly causal transfer function in every directed cycle. To the best knowledge of the authors, no other methodology is capable of learning an equally broad class of networks using exclusively non-invasive observations. This novel methodology unifies under a single framework other major reconstruction techniques, such as Granger causality and techniques imported from the area of graphical models, such as PC algorithm. The method consists of two tests that rely on specific variations of Wiener filtering. The first test checks if there is a relationship with direct throughput between two processes. If there is no such relationship, the second test checks for the presence of links with strictly causal transfer functions.

References

[1] D. Del Vecchio and R. M. Murray, *Biomolecular feedback systems*. Princeton University Press, 2015.

[2] J. R. Pomerening, "Positive-feedback loops in cell cycle progression," *FEBS letters*, vol. 583, no. 21, pp. 3388–3396, 2009.

[3] M. Kamiński, M. Ding, W. A. Truccolo, and S. L. Bressler, "Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance," *Biological cybernetics*, vol. 85, no. 2, pp. 145–157, 2001.

[4] M. J. Naylor, L. C. Rose, and B. J. Moyle, "Topology of foreign exchange markets using hierarchical structure methods," *Physica A: Statistical Mechanics and its Applications*, vol. 382, no. 1, pp. 199–208, 2007.

[5] J. R. Freeman, "Granger causality and the times series analysis of political relationships," *American Journal of Political Science*, pp. 327–358, 1983.

[6] M. D. Ward, K. Stovel, and A. Sacks, "Network analysis and political science," *Annual Review of Political Science*, vol. 14, pp. 245–264, 2011.

[7] J. Scott, *Social network analysis*. Sage, 2012.

[8] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[9] V. M. Carvalho, "From micro to macro via production networks," *The Journal of Economic Perspectives*, vol. 28, no. 4, pp. 23–47, 2014.

[10] D. Napoletani and T. Sauer, "Reconstructing the topology of sparsely connected dynamical networks," *Phys. Rev. E*, vol. 77, p. 026103, 2008.

[11] F. Lillo, R. N. Mantegna, and S. Miccichè, "Complex networks in air transport," *Complexity Science in Air Traffic Management*, p. 23, 2016.

[12] R. Simon, M. D. Radmacher, K. Dobbin, and L. M. McShane, "Pitfalls in the use of dna microarray data for diagnostic and prognostic classification," *Journal of the National Cancer Institute*, vol. 95, no. 1, pp. 14–18, 2003.

[13] D. Hayden, Y. H. Chang, J. Goncalves, and C. J. Tomlin, "Sparse network identifiability via compressed sensing," *Automatica*, vol. 68, pp. 9–17, 2016.

[14] V. Chetty, D. Hayden, J. Goncalves, and S. Warnick, "Robust signal-structure reconstruction," in *Decision and Control (CDC), 2013 IEEE 52nd Annual Conference on*. IEEE, 2013, pp. 3184–3189.

[15] S. Prabakaran, J. Gunawardena, and E. Sontag, "Paradoxical results in perturbation-based signaling network reconstruction," *Biophysical journal*, vol. 106, no. 12, pp. 2720–2728, 2014.

[16] M. Nabi-Abdolyousefi and M. Mesbahi, "Network identification via node knockout," *Automatic Control, IEEE Transactions on*, vol. 57, no. 12, pp. 3214–3219, 2012.

[17] S. Shahrampour and V. M. Preciado, "Topology identification of directed dynamical networks via power spectral analysis," *IEEE Transactions on Automatic Control*, vol. 60, no. 8, pp. 2260–2265, 2015.

[18] F. Morbidi and A. Y. Kibangou, "A distributed solution to the network reconstruction problem," *Systems & Control Letters*, vol. 70, pp. 85–91, 2014.

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

[20] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.

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

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

[23] A. Dankers, P. M. Van den Hof, X. Bombois, and P. S. Heuberger, "Identification of dynamic models in complex networks with prediction error methods: predictor input selection," *IEEE Transactions on Automatic Control*, vol. 61, no. 4, pp. 937–952, 2016.

[24] D. Materassi and M. Salapaka, "On the problem of reconstructing an unknown topology via locality properties of the wiener filter," *Automatic Control, IEEE Transactions on*, vol. 57, no. 7, pp. 1765–1777, 2012.

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

[26] C. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, pp. 424–438, 1969.

[27] C. D. Charalambous, P. A. Stavrou, and C. K. Kourtellaris, "Directed information on abstract spaces: properties and extremum problems," in *Coordination Control of Distributed Systems*. Springer, 2015, pp. 307–315.

[28] L. Barnett and A. K. Seth, "Granger causality for state-space models," *Physical Review E*, vol. 91, no. 4, p. 040101, 2015.

[29] P.-O. Amblard and O. J. Michel, "On directed information theory and granger causality graphs," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 7–16, 2011.

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

[31] A. V. Oppenheim, A. S. Willsky, and S. Nawab, *Signals and systems 2nd ed.* New Jersey: Prentice Hall, 1997.

[32] D. Materassi and M. V. Salapaka, "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.

[33] J. Pearl, *Causality*. Cambridge university press, 2009.

[34] J. Pearl and T. S. Verma, "A theory of inferred causation," *Studies in Logic and the Foundations of Mathematics*, vol. 134, pp. 789–811, 1995.

[35] T. Kailath, A. H. Sayed, and B. Hassibi, *Linear estimation*. Prentice Hall Upper Saddle River, NJ, 2000, vol. 1.

Appendix

A. Proof of Theorem 4.1

We first introduce a Lemma.

Lemma 6.1: Consider a well-posed, causal and topologically detectable LDG with output signal $Y = \{y_1, ..., y_n\}$. Let y_i and y_k be such that in every chain from y_i to y_k there is at least one strictly causal transfer function. Let $\hat{y}_k$ be the one-step ahead estimate of y_k using processes in $\frac{1}{z}S$, $S \subseteq Y$. Let $y_k^{\perp} = y_k - \hat{y_k}$ and let $\epsilon_i = e_i - \hat{e}_i$, where $\hat{e}_i$ is the one-step ahead estimate of e_i from $\frac{1}{z}e_i$. Then $y_k^{\perp} \perp \epsilon_i$.

Proof: Denote P_k^-: the strictly causal parents of y_k, except y_i, and P_k^+: the causal but not strictly causal parents of y_k. Note $y_i \notin P_k^+$. As there is no delay from any signal in P_k^+ to y_k we have that there must be a delay on every chain from y_i to any of the signals in P_k^+ and their ancestors. As we can write every $y_{p_k} \in P_k^+$ as $y_{p_k} = e_{p_k} + \sum_{y_a \in \{Pa(y_{p_k})\} \setminus y_i} H_{p_k a} y_a + H_{p_k i} y_i$, where $H_{p_k i}$ is strictly causal, we have $P_k^+ \in$ c-tf-span$\left(e_k, \{e_m\}_{m \neq i}, \frac{1}{z}e_i\right)$. As $y_k^{\perp} \in$ c-tf-span$\left(e_k, y_k, \frac{1}{z}S\right) \subseteq$ c-tf-span$\left(e_k, P_k^+, \frac{1}{z}P_k^-, \frac{1}{z}y_i, \frac{1}{z}S\right)$, we have that $y_k^{\perp} \in$ c-tf-span$\left(e_k, \{e_m\}_{m \neq i}, \frac{1}{z}e_i\right)$. On the other hand, $\epsilon_i \in$ c-tf-span$(e_i(t))$ and $\epsilon_i \perp \frac{1}{z}e_i$. Because Φ_{ee} is diagonal we have that $e_i \perp$ c-tf-span$(e_k, \{e_m\}_{m \neq i})$. This implies that $\epsilon_i \perp$ c-tf-span$(e_k, \{e_m\}_{m \neq i})$. Thus, $y_k^{\perp} \perp \epsilon_i$. ■

Proof: Partition the set of nodes $\{y_1, y_2, ...y_n\} \setminus \left\{y_i, y_j\right\}$ in the following subsets:

- P_{ij} : the set of parents of both y_i and y_j.
- $P_{j\bar{i}}^+$: the set of causal but not strictly causal parents of y_j, which are not parents of y_i.
- $P_{j\bar{i}}^-$: the set of strictly causal parents of y_j, which are not parents of y_i.
- $P_{i\bar{j}}^+$: the set of causal but not strictly causal parents of y_i, which are not parents of y_j.
- $P_{i\bar{j}}^-$: the set strictly causal parents of y_i, which are not parents of y_j; $P_{\bar{i}\bar{j}}$: all other nodes.

Let us compute the MLSE for $y_j(t)$ considering:

- $S^{+}_{j\bar{i}} := \left\{ y_k | \, y_k \in P^{+}_{j\bar{i}} \right\}$
- $S^{-}_{j\bar{i}} := \left\{ \frac{1}{z} y_k \,|\, y_k \in P^{+}_{j\bar{i}} \right\} \cup \left\{ y_k \,|\, y_k \in P^{-}_{j\bar{i}} \right\}$
- $S_{j\bar{i}} := S^{+}_{j\bar{i}} \cup S^{-}_{j\bar{i}}$; $S^{+}_{i\bar{j}} := \left\{ y_k \,|\, y_k \in P^{+}_{i\bar{j}} \right\}$
- $S^{-}_{i\bar{j}} := \left\{ \frac{1}{z} y_k \,|\, y_k \in P^{+}_{i\bar{j}} \right\} \cup \left\{ y_k \,|\, y_k \in P^{-}_{i\bar{j}} \right\}$; $S_{i\bar{j}} := S^{+}_{i\bar{j}} \cup S^{-}_{i\bar{j}}$.
- $S^{+}_{ji} := \left\{ y_k \,|\, y_k \in P_{ji} \text{ and } H_{ik} \text{ or } H_{jk} \text{ not strictly causal} \right\}$
- $S^{-}_{ji} := \left\{ \frac{1}{z} y_k \,|\, y_k \in P_{ji} \text{ and } H_{ki} \text{ or } H_{kj} \text{ not strictly causal} \right\}$ $\cup \left\{ y_k \,|\, y_k \in P_{ji} \text{ and } H_{ki} \text{ or } H_{kj} \text{ strictly causal } \right\}$
- $S_{ji} := S^{+}_{ji} \cup S^{-}_{ji}$, $S^{-}_{j} := \frac{1}{z} y_j$, $S^{-}_{i} := \frac{1}{z} y_i$
- $S = S_{j\bar{i}} \cup S_{i\bar{j}} \cup S_{ji} \cup S^{-}_{j} \cup S^{-}_{i}$

Let $\epsilon_j = e_j - \hat{e}_j$, e_j is the independent component of y_j and $\hat{e}_j = F_j^{-1}(z)\frac{1}{z} e_j$ is the one step ahead predictor for e_j from the past of e_j [35]. Let $\epsilon_i = e_i - \hat{e}_i$, e_i is the independent component of y_i and $\hat{e}_i = F_i^{-1}(z)\frac{1}{z} e_i$ is the one step ahead predictor for e_i from the past of e_i. Let $y_k^{\perp} = y_k - \hat{y}_k$, $\hat{y}_k$ is the one-step ahead predictor of $y_k(t)$, $k \neq \{i, j\}$, using the processes in $S^{-}_{j\bar{i}}, S^{-}_{i\bar{j}}, S^{-}_{ji}, S^{-}_{i}, S^{-}_{j}$. We group the $y_k^{\perp}$'s:

- $S^{\perp}_{j\bar{i}} := \left\{ y_k^{\perp} \,|\, y_k \in S^{+}_{j\bar{i}} \right\}$; $S^{\perp}_{i\bar{j}} := \left\{ y_k^{\perp} \,|\, y_k \in S^{+}_{i\bar{j}} \right\}$
- $S^{\perp}_{ji} := \left\{ y_k^{\perp} \,|\, y_k \in S^{+}_{ji} \right\}$

Note: $\text{c-tf-span}\left(S_{j\bar{i}}, S_{i\bar{j}}, S_{ij}, S^{-}_{i}, S^{-}_{j}\right) = \text{c-tf-span}\left(S^{-}_{j\bar{i}}, S^{-}_{i\bar{j}}, S^{-}_{ji}, S^{-}_{i}, S^{-}_{j}, S^{\perp}_{j\bar{i}}, S^{\perp}_{i\bar{j}}, S^{\perp}_{ji}\right)$.

Also, $\text{c-tf-span}\left(S^{\perp}_{j\bar{i}}, S^{\perp}_{i\bar{j}}, S^{\perp}_{ji}\right)$ is orthogonal to $\text{c-tf-span}\left(S^{-}_{j\bar{i}}, S^{-}_{i\bar{j}}, S^{-}_{ji}, S^{-}_{i}, S^{-}_{j}\right)$. We now show a claim:

Claim 6.2: The set $\text{c-tf-span}\left(S_{j\bar{i}}, S_{i\bar{j}}, S_{ji}, S^{-}_{i}, S^{-}_{j}, y_i\right)$ is equal to $\text{c-tf-span}\left(S_{j\bar{i}}, S_{i\bar{j}}, S_{ji}, S^{-}_{i}, S^{-}_{j}, \epsilon_i\right)$.

Proof:

We can write y_i as: $y_i = e_i + \sum_{y_k \in parents(y_i)} H_{ik} y_k = e_i + \sum_{y_k \in (S_{ij} \cup S_{i\bar{j}})} H_{ik} y_k$. Thus, $y_i \in \text{c-tf-span}(e_i, S_{ij} \cup S_{i\bar{j}})$. We can further write e_i as $e_i = \epsilon_i + F_i^{-1}(z)\frac{1}{z} e_i$, which implies that $e_i \in \text{c-tf-span}(\epsilon_i, \frac{1}{z} e_i)$. On the other hand, we can write $\frac{1}{z} e_i$ as the following: $\frac{1}{z} e_i = \frac{1}{z} y_i - \frac{1}{z} \sum_{y_k \in parents(y_i)} H_{ik} y_k \subseteq \text{c-tf-span}\left(S^{-}_{i} \cup S_{ij} \cup S_{i\bar{j}}\right)$ which implies $e_i \in \text{c-tf-span}\left(S_{j\bar{i}}, S_{i\bar{j}}, S_{ji}, S^{-}_{i}, S^{-}_{j}, \epsilon_i\right)$. ■

Now we obtain the optimal estimate for $y_j(t)$ using the processes in S and $y_i(t)$:

$$
\begin{aligned}
\hat{y}_j(t) &= \underset{q \in \text{c-tf-span}(S, y_i(t))}{\arg\min} \left\| y_j(t) - q \right\|^2 = \sum_{k \in S_{ij}} H_{jk} y_k + \sum_{k \in S_{j\bar{i}}} H_{jk} y_k \\
&+ \underset{q \in \text{c-tf-span}\left(S_{j\bar{i}}, S_{i\bar{j}}, S_{ji}, S^{-}_{i}, S^{-}_{j}, y_i(t)\right)}{\arg\min} \left\| e_j(t) - q \right\|^2 \\
&= \sum_{k \in S_{ij}} H_{jk} y_k + \sum_{k \in S_{j\bar{i}}} H_{jk} y_k + \hat{e}_j \\
&+ \underset{q \in \text{c-tf-span}\left(S^{-}_{j\bar{i}}, S^{-}_{i\bar{j}}, S^{-}_{ji}, S^{-}_{i}, S^{-}_{j}, S^{\perp}_{j\bar{i}}, S^{\perp}_{i\bar{j}}, S^{\perp}_{ji}, \epsilon_i(t)\right)}{\arg\min} \left\| \epsilon_j(t) - q \right\|^2 .
\end{aligned}
$$

Now we have the following observations.

- The innovation ϵ_j is orthogonal to $\text{c-tf-span}\left(S^{-}_{j\bar{i}}, S^{-}_{i\bar{j}}, S^{-}_{ji}, S^{-}_{i}, S^{-}_{j}\right)$ as $\epsilon_j \perp \text{c-tf-span}(\frac{1}{z} e_j)$, and as Φ_{ee} is diagonal, we have $\frac{1}{z} e_j \perp \frac{1}{z} e_k$, for all $k \neq j$.
- $\text{c-tf-span}(S^{\perp}_{j\bar{i}}, S^{\perp}_{i\bar{j}}, S^{\perp}_{ji}, \epsilon_i(t)) \perp (S^{-}_{j\bar{i}}, S^{-}_{i\bar{j}}, S^{-}_{ji}, S^{-}_{i}, S^{-}_{j})$.

Thus, we have $q \in \text{c-tf-span}\left(S^{\perp}_{j\bar{i}}, S^{\perp}_{i\bar{j}}, S^{\perp}_{ji}, \epsilon_i(t)\right)$. Hence:

$$
\begin{aligned}
\hat{y}_j(t) &= \underset{q \in \text{c-tf-span}(S, y_i(t))}{\arg\min} \left\| y_j(t) - q \right\|^2 = \sum_{k \in S_{ij}} H_{jk} y_k + \sum_{k \in S_{j\bar{i}}} H_{jk} y_k \\
&+ \hat{e}_j + \underset{q \in \text{c-tf-span}\left(S^{\perp}_{j\bar{i}}, S^{\perp}_{i\bar{j}}, S^{\perp}_{ji}, \epsilon_i(t)\right)}{\arg\min} \left\| \epsilon_j(t) - q \right\|^2 .
\end{aligned}
\tag{2}
$$

Note that if we were estimating $y_i(t)$ from $y_j(t)$, following analogous steps we would get $\epsilon_i(t)$ to be estimated by q belonging in the analogous $\text{c-tf-span}\left(S^{\perp}_{j\bar{i}}, S^{\perp}_{i\bar{j}}, S^{\perp}_{ji}, \epsilon_j(t)\right)$.

Now we observe that $\epsilon_i \perp \text{c-tf-span}(S^{\perp}_{j\bar{i}}, S^{\perp}_{i\bar{j}}, S^{\perp}_{ji})$. For that we use the hypothesis that there is at least one delay in each cycle. This implies that there is at least one delay on every directed path from y_i to y_j, or there is at least one delay on every path from y_j to y_i. Without loss of generality, assume that there is a delay in every path from y_i to y_j; otherwise we switch y_i and y_j in the whole argument. By considering the possible paths between y_i and y_j and by applying Lemma (6.1) we have that $\epsilon_i \perp \text{c-tf-span}(S^{\perp}_{j\bar{i}}, S^{\perp}_{i\bar{j}}, S^{\perp}_{ji})$. We also have that $\epsilon_i \perp \epsilon_j$. Going back to (2), the estimator of ϵ_i in the considered c-tf-span is zero, which shows that if H_{ij} and H_{ji} are strictly causal, $y_j(t)$ and $y_i(t)$ are uncorrelated given S. ■

B. *Proof of Theorem 4.2*

Proof: Define

$$
\begin{aligned}
S^{+} &= \left\{ y_k | \; y_k \in \text{not strictly causal parents of } y_j, k \neq i \right\} \\
S^{-} &= \left\{ y_k | \; y_k \in \text{strictly causal parents of } \; y_j \right\} \\
&\cup \left\{ \frac{1}{z} y_k | \; y_k \in \text{causal but not strictly causal parents of } y_j \right\} \\
S^{-}_{i} &= \frac{1}{z} y_i , \; S^{-}_{j} = \frac{1}{z} y_j , \; S = S^{+} \cup S^{-} \cup S^{-}_{j} \cup S^{-}_{i} .
\end{aligned}
$$

Let $\epsilon_j = e_j - \hat{e}_j$, where $\hat{e}_j = F_j^{-1}(z)\frac{1}{z} e_j$ is the one step ahead predictor for e_j, where e_j is the independent component of y_j. Let $y_k^{\perp} = y_k - \hat{y}_k$, where $\hat{y}_k$ is the one-step ahead predictor of $y_k \in S^{+}$ using the processes in $S^{-}, S^{-}_{j}, S^{-}_{i}$. The set of $y_k^{\perp}$ is denoted by $S^{\perp}$. We make the following observations:

- For $y_k \in S^{+}$, $y_k^{\perp} \perp \text{c-tf-span}\left(S^{-}, S^{-}_{j}, S^{-}_{i}\right)$.
- $\text{c-tf-span}(S) = \text{c-tf-span}(S^{\perp} \cup S^{-} \cup S^{-}_{j} \cup S^{-}_{i})$.
- $\text{c-tf-span}(S^{\perp})$ and $\epsilon_j \perp \text{c-tf-span}(\left\{ S^{-} \cup S^{-}_{j} \cup S^{-}_{i} \right\})$.

If y_k is a strictly causal parent of y_j, we have $y_k^{\perp} \notin S^{\perp}$. If y_k is a causal, but not strictly causal, parent of y_j, then $y_k^{\perp} = y_k - \hat{y}_k \in S^{\perp}$. As there is no delay from y_k to y_j, there must be at least one delay on every path from y_j to y_k. By Lemma 6.1, we have $\epsilon_j \perp S^{\perp}$. Thus we have that $\hat{y}_j \in S = \text{ c-tf-span } \left(S^{+}, S^{-}, \frac{1}{z} y_j\right)$, where $y_i, e_i \notin S$. As in the proof of Theorem 4.1 this means that the component of ϵ_i in the considered span is zero, implying that y_j is Wiener-uncorrelated with $\frac{1}{z} y_i$ given S.

■