Graph Topological Aspects of Granger Causal Network Learning

R. J. Kinnear
ryan@kinnear.ca
`https://github.com/RJTK`

R. R. Mazumdar
mazum@uwaterloo.ca

July 16, 2019

Abstract

We study Granger-causality in the context of wide-sense stationary time series, where our focus is on the topological aspects of the underlying causality graph. We establish sufficient conditions (in particular, we develop the notion of a "strongly causal" graph topology) under which the true causality graph can be recovered via pairwise causality testing alone, and provide examples from the gene regulatory network literature suggesting that our concept of a strongly causal graph may be applicable to this field. We implement and detail finite-sample heuristics derived from our theory, and establish through simulation the efficiency gains (both statistical and computational) which can be obtained (in comparison to LASSO algorithms) when structural assumptions are met. Finally, we provide an example application where we accurately discriminate between subjects in an EEG study based purely on the Granger-causality graphs inferred from data by our algorithms, demonstrating that meaningful features are captured by the Granger-causal graph topology.

Keywords— causality graph, EEG, gene regulatory networks, Granger-causality, LASSO, network learning, time series, vector autoregression

Acknowledgement We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number 518418-2018]. Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), [numéro de référence 518418-2018].

1 Introduction and Review

In this paper we study the notion of Granger-causality [1] [2] as a means of uncovering an underlying causal structure in multivariate time series. Though the underlying causality graph cannot be observed directly, we will infer it's presence as a latent structure among our observed time series data. This notion is leveraged in a variety of applications e.g. in Neuroscience as a means of recovering interactions amongst brain regions [3], [4], [5]; in the study of the dependence and connectedness of financial institutions [6]; gene expression networks [7], [8], [9], [10]; and power system design [11], [12].

Granger-causality can generally be formulated by searching for the "best" graph structure consistent with observed data, which is in general an extremely challenging problem (i.e. it may be framed as a best subset selection problem, see [13], [14]), moreover, the comparison of quality between different structures, and hence the notion of "best" needs qualification. In applications where we are interested merely in minimizing the mean squared error of a linear one-step-ahead predictor, then we will be satisfied with an entirely dense graph of connections, since each edge can only serve to reduce estimation error. However, since the number of edges scales quadratically in n (the number of nodes) it becomes imperative to infer a sparse causality graph for large systems, both to avoid overfitting observed data, as well as to aid the interpretability of the results.

A fairly early approach to the problem in the context of large systems is provided by [15], where the authors apply a local search heuristic to the Whittle likelihood with an AIC penalization. The local search heuristic where at each iteration an edge is either added, removed, or reversed is a common approach to combinatorial optimization due to it's simplicity, but is liable to get stuck in shallow local minima.

A second and wildly successful heuristic is the LASSO regularizer [16], which can be understood as a natural convex relaxation to penalizing the count of the non-zero edges. The LASSO enjoys fairly strong theoretical guarantees [17], extending largely to the case of stationary time series data with a sufficiently fast rate of dependence decay [18] [19] [20], and variations on the LASSO have been applied in a number of different time series contexts as well as Granger-causality [21] [22] [23] [24] [9]. One of the key improvements to the original LASSO algorithm is the adaptive (i.e. weighted) "adaLASSO" [25], for which oracle results (i.e. asymptotic support recovery) are established under less restrictive conditions than for the vanilla LASSO. Our experimental comparisons in Section 4 are against the adaLASSO.

1.1 Contributions

In the context of time series data, sparsity assumptions remain important, but there is significant additional structure that may arise as a result of considering the topology of the underlying Granger-causality graph, which to our knowledge remains largely unexplored. The focus of this paper is to shed light on some of these topological questions, in particular, we study a particularly simple notion of causality graph topology which we term "strongly causal" and show that stationary times series whose underlying causality graph has this structure satisfy natural intuitive notions of "information flow" through the graph. Moreover, we show that such graphs are perfectly recoverable with only *pairwise* Granger-causality tests, which would otherwise suffer from serious confounding problems. Our finite sample results are based on simulations, where we show promising results for these graph topologies where our algorithm performs substantially better in our simulation setup than do competing LASSO algorithms, even for graphs that do not exactly satisfy our strongly-causal topology assumptions.

In the case of gene expression networks, we show examples from the literature which suggest our concept of a "strongly causal graph" topology may have application in this field (see Section 2.5).

The principle contributions of this paper are as follows: firstly, in section 2 we study *pairwise* Granger-causality relations, providing novel theorems connecting the structure of the causality graph to the pairwise "causality flow" in the system, as well as an interpretation in terms of the graph topology of the sparsity pattern of matrices arising in the Wold decomposition, generalizing in some sense the notion of "feedback-free" processes studied by [26] in close connection with Granger-causality. We establish sufficient conditions (sections 2.5, 2.6) under which a fully conditional Granger-causality graph can be recovered from pairwise tests alone (sec 2.7). Secondly, we propose in section 3 a graph search heuristic which implements our theoretical results to finite data samples, specifying and summarizing appropriate methods for hypothesis testing (section 3.1), model order selection (section 3.2), computationally efficient estimation (section 3.3), and error rate controls (section 3.4). Our heuristics are compared against the adaLASSO algorithm in Section 4. We stress the scalability of our algorithm which is capable of comfortably handling hundreds or thousands of nodes on a single machine, as opposed to standard LASSO algorithms which do not take advantage of the special structure associated with stationary time series data. In section 5 we develop an example application where a classifier is constructed to accurately discriminate between subjects in an EEG study based only on the Granger-causality graph topology inferred by our algorithms. Concluding remarks on further open problems and extentions are provided in Section 6.

2 Theory

2.1 Formal Setting

Consider the space $L_2(\Omega)$, the usual Hilbert space of finite variance random variables over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ having inner product $\langle x, y \rangle = \mathbb{E}[xy]$. We will work with a discrete time and wide-sense stationary (WSS) n-dimensional vector valued process $x(t)$ (with $t \in \mathbb{Z}$) where the n elements take values in L_2. We suppose that $x(t)$ has zero mean, $\mathbb{E}x(t) = 0$, and has absolutely summable matrix valued covariance sequence $R(\tau) \overset{\Delta}{=} \mathbb{E}x(t)x(t-\tau)^{\mathsf{T}}$ with an absolutely continuous spectral density $S(\omega)$:

$$\begin{aligned} R(\tau) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} S(\omega)e^{j\tau\omega}\mathsf{d}\omega, \\ S(\omega) &= \sum_{\tau=-\infty}^{\infty} R(\tau)e^{-j\tau\omega}. \end{aligned} \tag{1}$$

We will also work frequently with the spaces spanned by the values of such a process

$$\begin{aligned} \mathcal{H}_t^x &= \mathsf{cl}\ \{\sum_{\tau=0}^{p} a_\tau^{\mathsf{T}} x(t-\tau) \mid a_\tau \in \mathbb{R}^n, p \in \mathbb{N}\} \subseteq L_2(\Omega) \\ H_t^x &= \{ax(t) \mid a \in \mathbb{R}\} \subseteq L_2(\Omega), \end{aligned} \tag{2}$$

where the closure is naturally in mean-square. We will often omit the superscript x which should be clear from context. Evidently these spaces are separable, and as closed subspaces of a Hilbert space they are themselves Hilbert. We will denote the spaces generated in analogous ways by particular components of x as e.g. $\mathcal{H}_t^{(i,j)}$, $\mathcal{H}_t^i$ or by all but particular components as $\mathcal{H}_t^{-j}$.

As a consequence of the Wold decomposition theorem [27], every WSS sequence has the moving average $MA(\infty)$ representation

$$x(t) = c(t) + \sum_{\tau=0}^{\infty} A(\tau)v(t-\tau), \tag{3}$$

where $c(t)$ is a purely deterministic sequence[1], $v(t)$ is an uncorrelated sequence and $A(0) = I$. We will assume that $c(t) = 0$, which in practice is to say that the process has been detrended. We additionally require that this representation can be inverted to yield the $\mathsf{VAR}(\infty)$ form

$$x(t) = \sum_{\tau=1}^{\infty} B(\tau)x(t-\tau) + v(t). \tag{4}$$

The equations (3), (4) can be represented as $x(t) = \mathsf{A}(z)v(t) = \mathsf{B}(z)x(t) + v(t)$ via the action (convolution) of the operators (LTI filters) $\mathsf{A}(z) \overset{\Delta}{=} \sum_{\tau=0}^{\infty} A(\tau)z^{-\tau}$ and $\mathsf{B}(z) \overset{\Delta}{=} \sum_{\tau=1}^{\infty} B(\tau)z^{-\tau}$ where the operator z^{-1} is the back shift operator acting on $\ell_2^n(\Omega, \mathcal{F}, \mathbb{P})$, that is:

$$\mathsf{B}_{ij}(z)x_j(t) \overset{\Delta}{=} \sum_{\tau=1}^{\infty} B_{ij}(\tau)x_j(t-\tau). \tag{5}$$

Finally, since $||z^{-1}|| = 1$ we have the inversion formula

$$\mathsf{A}(z) = (I - \mathsf{B}(z))^{-1} = \sum_{k=0}^{\infty} \mathsf{B}(z)^k. \tag{6}$$

[1] the purely deterministic sequence $c(t)$ is one which lies in the remote past $\bigcap_{\tau=1}^{\infty} \mathcal{H}_{t-\tau}^x$ of the process. For such processes a single sample $c(t_0)$ is enough to determine $c(t)$ for every t. For example, $c(t) = \sin(2\pi t + \Theta);\ \Theta \sim \mathcal{U}[-\pi, \pi]$

The aforementioned assumptions are quite weak. The strongest assumption we require is finally that Σ_v is a diagonal matrix, which is referred to as a lack of instantaneous feedback in $x(t)$. We formally state our setup to conclude this section:

Definition 1 (Basic Setup)**.** The process $x(t)$ is an n dimensional wide sense stationary process having invertible $\mathsf{VAR}(\infty)$ representation (4) where $v(t)$ is sequentially uncorrelated and has a diagonal covariance matrix. The $MA(\infty)$ representation of equation (3) has $c(t) = 0$ and $A(0) = I$.

2.2 Granger Causality

Definition 2 (Granger Causality)**.** For the WSS series $x(t)$ satisfying the assumptions of Definition 1 we will say that component x_j *Granger-Causes* (GC) component x_i (with respect to x) and write $x_j \overset{\text{GC}}{\rightarrow} x_i$ if given Hilbert spaces $\mathcal{H}_{t-1}$, $\mathcal{H}_{t-1}^{-j}$

$$\xi[x_i(t) \mid \mathcal{H}_{t-1}] < \xi[x_i(t) \mid \mathcal{H}_{t-1}^{-j}], \tag{7}$$

where $\xi[x \mid \mathcal{H}] = \mathbb{E}(x - \hat{\mathbb{E}}[x \mid \mathcal{H}])^2$ is the mean squared estimation error and $\hat{\mathbb{E}}[x \mid \mathcal{H}] = \text{proj}_{\mathcal{H}}(x)$ denotes the (unique) projection onto the Hilbert space $\mathcal{H}$.

This notion captures the idea that the process x_j provides information about x_i that is not available from elsewhere. The caveat "with respect to x" is important in that GC relations can change when components are added to or removed from our collection x of observations, e.g. new GC relations can arise if we remove the observations of a common cause, and existing GC relations can disappear if we observe a new mediating series.

The notion is closely related to the information theoretic measure of transfer entropy, indeed, if the distribution of $v(t)$ is known to be Gaussian then they are equivalent [28].

We require two technical facts we refer to in the sequel, the notion of conditional orthogonality is used throughout.

Lemma 1 ([27] Proposition 2.4.2)**.** *Consider three closed subspaces of a Hilbert space $\mathcal{A}$, $\mathcal{B}$, $\mathcal{X}$. The following statements are equivalent*

1. $\mathcal{A} \perp \mathcal{B} \mid \mathcal{X}$
2. $\hat{\mathbb{E}}[\beta \mid \mathcal{A} \vee \mathcal{X}] = \hat{\mathbb{E}}[\beta \mid \mathcal{X}]\ \forall \beta \in \mathcal{B}$.

Where $\mathcal{A} \perp \mathcal{B} \mid \mathcal{X}$ denotes conditional orthogonality:

$$\langle a - \hat{\mathbb{E}}[a \mid \mathcal{X}], b - \hat{\mathbb{E}}[b \mid \mathcal{X}]\rangle = 0\ \forall a \in \mathcal{A}, b \in \mathcal{B}.$$

Lemma 2. *Let $x \in \mathcal{H}$ where $\mathcal{H}$ is a separable Hilbert space having inner product $\langle \cdot, \cdot \rangle$ and let $\{\mathcal{H}_n\}_{n=0}^N$ be a collection of subspaces of $\mathcal{H}$ such that $x \perp \mathcal{H}_0$. Then*

$$\hat{\mathbb{E}}[x \mid \bigvee_{n=0}^{N} \mathcal{H}_n] = \hat{\mathbb{E}}[x \mid \bigvee_{n=1}^{N} \mathcal{H}_n],$$

where $\mathcal{H}_1 \vee \mathcal{H}_2 = \mathsf{cl}\ \{\alpha + \beta \mid \alpha \in \mathcal{H}_1, \beta \in \mathcal{H}_2\}$ is the closed sum of subspaces.

Proof. Apply the Gram-Schmidt process and directly calculate the projections. □

Theorem 1 (Granger Causality Equivalences)**.** *Let $x(t)$ be a WSS process with absolutely summable covariance sequence, and spectral density uniformly bounded above and below. Denote by $\xi_{ij} \overset{\Delta}{=} \xi[x_i(t) \mid \mathcal{H}^{-j}]$ and $\xi_i \overset{\Delta}{=} \xi[x_i(t) \mid \mathcal{H}_t]$ for distinct i, j. Then, the following are equivalent:*

1. $x_j \overset{GC}{\nrightarrow} x_i$
2. $\forall \tau \in \mathbb{N}_+ \; B_{ij}(\tau) = 0$ *i.e.* $\mathsf{B}_{ij}(z) = 0$
3. $F_{ij} \overset{\Delta}{=} \big(\frac{\xi_i}{\xi_{ij}} - 1\big) = 0$
4. $\mathcal{H}_t^i \perp \mathcal{H}_{t-1}^j \mid \mathcal{H}_{t-1}^{-j} \iff H_t^i \perp \mathcal{H}_{t-1}^j \mid \mathcal{H}_{t-1}^{-j}$
5. $\hat{\mathbb{E}}[x_i(t) \mid \mathcal{H}_{t-1}^{-j}] = \hat{\mathbb{E}}[x_i(t) \mid \mathcal{H}_{t-1}]$

Proof. (1) $\iff$ (3) is a tautology, and (3) $\iff$ (5) is immediate from the definitions i.e. the projections are equal if and only if the errors are equal. We have (4) $\iff$ (5) as a result of Lemma 1

(1) ⇒ (2): The projection for $x_i(t)$ onto $\mathcal{H}_{t-1}$ is (by definition) given by a form similar to equation 4:

$$\mathbb{E}|x_i(t) - \hat{\mathbb{E}}[x_i(t) \mid \mathcal{H}_{t-1}]|^2 = \mathbb{E}|v_i(t) + \sum_{\tau=1}^{\infty}\sum_{k=1}^{n}(B_{i,k}(\tau) - \hat{B}_{i,k}(\tau))x_k(t-\tau)|^2.$$

Since $v(t)$ in equation 4 is temporally uncorrelated, it follows that the optimal projection is given by the model coefficients themselves. This holds similarly for the projection onto $\mathcal{H}_{t-1}^{-j}$. Then by (1) we have

$$\mathbb{E}|x_i(t) - \sum_{\tau=1}^{\infty}\sum_{k=1}^{n} B_{i,k}(\tau)x_k(t-\tau)|^2 = \mathbb{E}|x_i(t) - \sum_{\tau=1}^{\infty}\sum_{k \neq j} B_{i,k}(\tau)x_k(t-\tau)|^2$$

By the uniqueness of the projection we must have $\forall \tau \; B_{i,j}(\tau) = 0$.

(2) ⇒ (4): In computing $(y - \hat{\mathbb{E}}[y \mid \mathcal{H}_{t-1}^{-j}])$ for $y \in H_t^i$ it is sufficient to consider $y = x_i(t)$ by linearity, then since $H_{t-1}^i \subseteq \mathcal{H}_{t-1}^{-j}$ we have $(x_i(t) - \hat{\mathbb{E}}[x_i(t) \mid \mathcal{H}_{t=1}^{-j}]) = v_i(t)$ since $\mathsf{B}_{ij}(z) = 0$. (4) then follows since $v_i(t) \perp \mathcal{H}_{t-1}$ and $\forall z \in \mathcal{H}_{t-1}^j \; (z - \hat{\mathbb{E}}[z \mid \mathcal{H}_{t-1}^{-j}]) \in \mathcal{H}_{t-1}$ □

2.3 Granger Causality Graphs

We first need to establish some graph theoretic notation and terminology, collected formally in definitions for the reader's convenient reference.

Definition 3 (Graph Theory Review)**.** A *graph* $\mathcal{G} = (V, \mathcal{E})$ is simply a tuple of sets respectively called *nodes* and *edges*. Throughout this paper, we have in all cases $V = [n] \overset{\Delta}{=} \{1, 2, \ldots, n\}$. We will also focus solely on *directed* graphs, where the edges $\mathcal{E} \subseteq V \times V$ are *ordered* pairs.

A (directed) *path* (of length r) from node i to node j, denoted $i \rightarrow \cdots \rightarrow j$, is a sequence $a_0, a_1, \ldots, a_{r-1}, a_r$ with $a_0 = i$ and $a_r = j$ such that $\forall\ 0 \leq k \leq r\ (a_k, a_{k+1}) \in \mathcal{E}$, and where (a_k, a_{k-1}) are *distinct* for $0 \leq k < r$.

A *cycle* is a path of length 2 or more between a node and itself. An edge between a node and itself $(i, i) \in \mathcal{E}$ (which is not a cycle) is referred to as a *loop*.

A graph $\mathcal{G}$ is a *directed acyclic graph* (DAG) if it is a directed graph and does not contain any cycles.

Definition 4 (Parents, Grandparents, Ancestors)**.** A node j is a *parent* of node i if $(j, i) \in \mathcal{E}$. The set of all i's parents will be denoted $pa(i)$, and we explicitly exclude loops as a special case, that is, $i \notin pa(i)$ even if $(i, i) \in \mathcal{E}$.

The set of level ℓ *grandparents* of node i, denoted $gp_\ell(i)$, is the set such that $j \in gp_\ell(i)$ if and only if there is a *directed path* of length ℓ in $\mathcal{G}$ from j to i. Clearly, $pa(i) = gp_1(i)$.

Finally, the set of *level ℓ ancestors* of i: $\mathcal{A}_\ell(i) = \bigcup_{\lambda \leq \ell} gp_\lambda(i)$ is the set such that $j \in \mathcal{A}_\ell(i)$ if and only if there is a directed path of length ℓ *or less* in $\mathcal{G}$ from j to i. The set of *all ancestors* of i (i.e. $\mathcal{A}_n(i)$) is denoted simply $\mathcal{A}(i)$.

Recall that we do not allow a node to be it's own parent, however unless $\mathcal{G}$ is a DAG, a node can be it's own ancestor. We will ocassionally need to explicitly exclude i from $\mathcal{A}(i)$, in which case we will write $\mathcal{A}(i) \setminus \{i\}$.

Our principle object of study will be a graph determined by Granger-causality relations as follows.

Definition 5 (Causality graph)**.** We define the Granger-causality graph $\mathcal{G} = ([n], \mathcal{E})$ to be the directed graph formed on n vertices where an edge $(j, i) \in \mathcal{E}$ if and only if x_j Granger-causes x_i (with respect to x). That is,

$$(j, i) \in \mathcal{E} \iff j \in pa(i) \iff x_j \overset{\text{GC}}{\rightarrow} x_i.$$

The edges of the Granger-causality graph $\mathcal{G}$ can be given a general notion of "weight" by associating an edge (j, i) with the *strictly causal* LTI filter $\mathsf{B}_{ij}(z)$ (see eqn (5)). Thence, the matrix $\mathsf{B}(z)$ is analogous to a *weighted adjacency matrix*[2] for the graph $\mathcal{G}$. And, in the same way that the k^{th} power of an adjacency matrix counts the number of paths of length k between nodes, $(\mathsf{B}(z)^k)_{ij}$ is a filter isolating the "action" of j on i at a time lag of k steps, this is exemplified in the inversion formula 6.

An elementary theorem connecting the adjacency matrix with paths will allow us to deduce the sparsity pattern of $\mathsf{A}(z)$. Proof follows easily by induction:

[2] We are using the convention that $\mathsf{B}_{ij}(z)$ is a filter with input x_j and output x_i so as to write the action of the system as $\mathsf{B}(z)x(t)$ with $x(t)$ as a column vector. This competes with the usual convention for adjacency matrices where $A_{ij} = 1$ if there is an edge (i, j). In our case, the sparsity pattern of B_{ij} is the *transposed* conventional adjacency matrix.

Lemma 3. *Let S be the transposed adjacency matrix of the Granger-causality graph $\mathcal{G}$. Then, $(S^k)_{ij}$ is the number of paths of length k from node j to node i. Evidently, if $\forall k \in \mathbb{N},\ (S^k)_{ij} = 0$ then $j \notin \mathcal{A}(i)$.*

From the VAR representation of $x(t)$ there is clearly a tight relationship between each node and it's parent nodes, the relationship is quantified through the sparsity pattern of $B(z)$. Similarly, the following proposition is analogous to the definition of feedback free processes of [26] and provides an interpretation of the sparsity pattern of $A(z)$ (from the MA representation of $x(t)$) in terms of the causality graph $\mathcal{G}$.

Proposition 1 (Ancestor Expansion)**.** *The component $x_i(t)$ of $x(t)$ can be represented in terms of it's parents in $\mathcal{G}$:*

$$x_i(t) = v_i(t) + \mathsf{B}_{ii}(z)x_i(t) + \sum_{k \in pa(i)} \mathsf{B}_{ik}(z)x_k(t). \tag{8}$$

Moreover, x_i can be expanded in terms of it's ancestor's $v(t)$ components only:

$$x_i(t) = \mathsf{A}_{ii}(z)v_i(t) + \sum_{\substack{k \in \mathcal{A}(i) \\ k \neq i}} \mathsf{A}_{ik}(z)v_k(t), \tag{9}$$

where $\mathsf{A}(z) = \sum_{\tau=0}^{\infty} A(\tau)z^{-\tau}$ is the filter from the Wold decomposition representation of $x(t)$, equation (3).

This statement is ultimately about the sparsity pattern in the Wold decomposition matrices $A(\tau)$ since $x_i(t) = \sum_{\tau=0}^{\infty}\sum_{j=1}^{n} A_{ij}(\tau)v_j(t-\tau)$. The proposition states that if $j \notin \mathcal{A}(i)$ then $\mathsf{A}_{ij}(z) = 0$.

Proof. Equation (8) is immediate from the $\mathsf{VAR}(\infty)$ representation of (4) and Theorem 1, we are left to demonstrate (9).

From equation (4), which we are assuming throughout the paper to be invertible, we can write

$$x(t) = (I - \mathsf{B}(z))^{-1}v(t),$$

where $(I - \mathsf{B}(z))^{-1} = \mathsf{A}(z)$ due to the uniqueness of (3). Since $\mathsf{B}(z)$ is stable we have

$$(I - \mathsf{B}(z))^{-1} = \sum_{k=0}^{\infty} \mathsf{B}(z)^k. \tag{10}$$

Invoking the Cayley-Hamilton theorem allows writing the infinite sum of (10) in terms of *finite* powers of B.

Let S be a matrix with elements in $\{0, 1\}$ which represents the sparsity pattern of $\mathsf{B}(z)$, from lemma 3 S is the transpose of the adjacency matrix for $\mathcal{G}$ and hence $(S^k)_{ij}$ is non-zero if and only if $j \in gp_k(i)$, and therefore $\mathsf{B}(z)^k_{ij} = 0$ if $j \notin gp_k(i)$. Finally, since $\mathcal{A}(i) = \bigcup_{k=1}^{n} gp_k(i)$ we see that $\mathsf{A}_{ij}(z)$ is zero if $j \notin \mathcal{A}(i)$.

Thence

$$\begin{aligned}
x_i(t) &= [(I - \mathsf{B}(z))^{-1}v(t)]_i \\
&= \sum_{j=1}^{n} \mathsf{A}_{ij}(z)v_j(t) \\
&= \mathsf{A}_{ii}(z)v_i(t) + \sum_{\substack{j\in\mathcal{A}(i)\\ j\neq i}} \mathsf{A}_{ij}(z)v_j(t)
\end{aligned}$$

□

2.4 Pairwise Granger Causality

Recall that Granger-causality in general must be understood with respect to a particular universe of observations. If $x_j \overset{\text{GC}}{\to} x_i$ with respect to x_{-k}, it may not hold with respect to x. For example, x_k may be a common ancestor which when observed, completely explains the connection from x_j to x_i. In this section we study *pairwise* Granger-causality, and seek to understand when knowledge of pairwise relations is sufficient to deduce the true fully conditional relations of $\mathcal{G}$.

Definition 6 (Pairwise Granger-causality)**.** We will say that x_j pairwise Granger-causes x_i and write $x_j \overset{\text{PW}}{\to} x_i$ if x_j Granger-causes x_i with respect only to (x_i, x_j).

This notion is of interest for a variety of reasons. From a purely conceptual standpoint, we will see how the notion can in some sense capture the idea of "flow of information" in the underlying graph, in the sense that if $j \in \mathcal{A}(i)$ we expect that $j \overset{\text{PW}}{\to} i$. It may also be useful for reasoning about the conditions under which *unobserved* components of $x(t)$ may or may not interfere with inference in the actually observed components. Finally, motivated from a practical standpoint to analyze causation in large systems, we will seek to construct practical estimation procedures based purely on pairwise causality tests since the computation of such pairwise relations is somewhat easier.

Lemma 4. *Consider distinct nodes i, j in a Granger-causality graph $\mathcal{G}$. If*

(a) $j \not\in \mathcal{A}(i)$ and $i \not\in \mathcal{A}(j)$

(b) $\mathcal{A}(i) \cap \mathcal{A}(j) = \emptyset$

then $\mathcal{H}_t^{(i)} \perp \mathcal{H}_t^{(j)}$, that is, $\forall s, \tau \in \mathbb{Z}_+ \; \mathbb{E}[x_i(t-s)x_j(t-\tau)] = 0$. Moreover, this means that $j \overset{PW}{\nrightarrow} i$ and $\hat{\mathbb{E}}[x_j(t) \mid \mathcal{H}_{t-1}^i] = 0$.

Remark 1. The possibility that there exist nodes sharing both i and j as ancestors is not excluded, the point being that the temporal nature of Granger-causality eliminates the problems caused by "colliders" (in the language of Pearl's causal calculus [29]) in static causal inference.

Proof. We show directly that $\forall s, \tau \in \mathbb{Z}_+$ $\mathbb{E}[x_i(t-s)x_j(t-\tau)] = 0$. To this end, fix $s, \tau \geq 0$, then by expanding with equation (9) we have

$$\begin{aligned}
\mathbb{E}x_i(t-s)x_j(t-\tau) &= \mathbb{E}\big(\mathsf{A}_{ii}(z)v_i(t-s)\big)\big(\mathsf{A}_{jj}(z)v_j(t-\tau)\big) \\
&+ \sum_{\substack{k \in \mathcal{A}(i) \\ k \neq i}} \mathbb{E}[\big(\mathsf{A}_{ik}(z)v_k(t-s)\big)\big(\mathsf{A}_{jj}(z)v_j(t-\tau)\big)] \\
&+ \sum_{\substack{k \in \mathcal{A}(j) \\ k \neq j}} \mathbb{E}[\big(\mathsf{A}_{ii}(z)v_i(t-s)\big)\big(\mathsf{A}_{jk}(z)v_k(t-\tau)\big)] \\
&+ \sum_{\substack{k \in \mathcal{A}(i) \\ k \neq i}} \sum_{\substack{\ell \in \mathcal{A}(j) \\ \ell \neq j}} \mathbb{E}[\big(\mathsf{A}_{ik}(z)v_k(t-s)\big)\big(\mathsf{A}_{j\ell}(z)v_\ell(t-\tau)\big)].
\end{aligned}$$

Keeping in mind that $v(t)$ is an isotropic and uncorrelated sequence we see that each of these above four terms are 0: the first term since $i \neq j$, the second and third since $j \notin \mathcal{A}(i)$ and $i \notin \mathcal{A}(j)$ and finally the fourth since $\mathcal{A}(i) \cap \mathcal{A}(j) = \emptyset$. □

It is possible for components of $x(t)$ to be correlated at some time lags without resulting in pairwise causality:

Lemma 5. *Consider distinct nodes i, j in a Granger-causality graph $\mathcal{G}$. If*

(a) $j \notin \mathcal{A}(i)$

(b) $\mathcal{A}(i) \cap \mathcal{A}(j) = \emptyset$

then $j \overset{PW}{\nrightarrow} i$.

Proof. By Theorem 1 it suffices to show that

$$\langle x_i(t) - \hat{\mathbb{E}}[x_i(t) \mid \mathcal{H}_{t-1}^i], x_j(t) - \hat{\mathbb{E}}[x_j(t) \mid \mathcal{H}_{t-1}^i]\rangle = 0,$$

which by the orthogonality principle is equivalent to

$$\langle x_i(t), x_j(t) - \hat{\mathbb{E}}[x_j(t) \mid \mathcal{H}_{t-1}^i]\rangle = 0. \tag{11}$$

Define the disjoint sets

$$\begin{aligned}
C_0(u) &= \{k \in pa(u) \mid i \notin \mathcal{A}(k), k \neq i\} \\
C_1(u) &= \{k \in pa(u) \mid i \in \mathcal{A}(k) \text{ or } k = i\}.
\end{aligned}$$

We can then expand the parents of j using Equation 8 as

$$x_j(t) = v_j(t) + \sum_{k \in C_0(j)} B_{jk}(z)x_k(t) + \sum_{k \in C_1(j)} B_{jk}(z)x_k(t),$$

which when substituted into the left hand side of Equation 12 results (by Lemma 4) in

$$\langle x_i(t), \sum_{k \in C_1(j)} B_{jk}(z)x_k(t) - \sum_{k \in C_1(j)} \hat{\mathbb{E}}[B_{jk}(z)x_k(t) \mid \mathcal{H}_{t-1}^i]\rangle.$$

We can continue this process recursively (i.e. split each $k \in C_1(j)$ into $C_0(k)$ and $C_1(k)$) which must eventually terminate with $C_1(u) = \{i\}$. Therefore, there exists some causal filter $\Phi(z)$ such that Equation 12 is equivalent to

$$\langle x_i(t), \Phi(z)x_i(t) - \hat{\mathbb{E}}[\Phi(z)x_i(t) \mid \mathcal{H}_{t-1}^i]\rangle, \tag{12}$$

which is 0 since $\hat{\mathbb{E}}[\Phi(z)x_i(t) \mid \mathcal{H}_{t-1}^i] = \Phi(z)x_i(t)$. □

The previous proposition can be strengthened significantly; notice that it is possible to have some $k \in \mathcal{A}(i) \cap \mathcal{A}(j)$ where still $j \overset{\text{PW}}{\nrightarrow} i$, an example is furnished by the three node graph $k \to i \to j$ where clearly $k \in \mathcal{A}(i) \cap \mathcal{A}(j)$ but $j \overset{\text{PW}}{\nrightarrow} i$. We must introduce the concept of a *confounding* variable, which effectively eliminates the possibility presented in this example.

Definition 7 (Confounder)**.** A node k will be referred to as a *confounder* of nodes i, j (neither of which are equal to k) if $k \in \mathcal{A}(i) \cap \mathcal{A}(j)$ and there exists a path $k \to \cdots \to i$ not containing j, and a path $k \to \cdots \to j$ not containing i.

A simple example is furnished by the "fork" graph $i \leftarrow k \to j$.

Proposition 2. *If in a Granger-causality graph $\mathcal{G}$ where $j \overset{PW}{\to} i$ then $j \in \mathcal{A}(i)$ or $\exists k \in \mathcal{A}(i) \cap \mathcal{A}(j)$ which is a confounder of (i, j).*

Remark 2. The interpretation of this proposition is that for $j \overset{\text{PW}}{\to} i$ then there must either be "causal flow" from j to i ($j \in \mathcal{A}(i)$) or there must be a confounder k through which common information is received.

Proof. We will prove by way of contradiction. To this end, suppose that j is a node such that: (a) $j \notin \mathcal{A}(i)$ and (b) for every $k \in \mathcal{A}(i) \cap \mathcal{A}(j)$ every $k \to \cdots \to j$ path contains i.

Firstly, notice that every $u \in \big(pa(j) \setminus \{i\}\big)$ necessarily inherits these same two properties. This follows since if we also had $u \in \mathcal{A}(i)$ then $u \in \mathcal{A}(i) \cap \mathcal{A}(j)$ so that every $u \to \cdots \to j$ path must contain i, but $u \in pa(j)$, so this is not the case since $u \to j$ is a path that doesn't contain i; moreover, if we consider $w \in \mathcal{A}(i) \cap \mathcal{A}(u)$ then we also have $w \in \mathcal{A}(i) \cap \mathcal{A}(j)$ so every $w \to \cdots \to j$ path must contain i. These properties therefore extend inductively to every $u \in \big(\mathcal{A}(j) \setminus \{i\}\big)$.

In order to deploy a recursive argument, define the following partition of $pa(u)$, for some node u:

$$\begin{aligned}
C_0(u) &= \{k \in pa(u) \mid i \notin \mathcal{A}(k), \mathcal{A}(i) \cap \mathcal{A}(k) = \emptyset, k \neq i\}\\
C_1(u) &= \{k \in pa(u) \mid i \in \mathcal{A}(k) \text{ or } k = i\}\\
C_2(u) &= \{k \in pa(u) \mid i \notin \mathcal{A}(k), \mathcal{A}(i) \cap \mathcal{A}(k) \neq \emptyset, k \neq i\}.
\end{aligned}$$

We notice that for any u having the properties $(a), (b)$ above, we must have $C_2(u) = \emptyset$ since if $k \in C_2(u)$ then $\exists w \in \mathcal{A}(i) \cap \mathcal{A}(k)$ s.t. $i \notin \mathcal{A}(k)$ and therefore there must be a path $w \to \cdots \to k \to u$ which does not contain i.

Using this partition, we will expand $x_j(t)$ in terms of it's parents, and recursively expand nodes in C_1 until we reach a case where $C_1 = \emptyset$. For the first step equation (8) gives us:

$$x_j(t) = \mathsf{A}_{jj}(z)\Big(v_j(t) + \sum_{k \in C_0(j)} \mathsf{B}_{ik}(z)x_k(t) + \sum_{k \in C_1(j)} \mathsf{B}_{ik}(z)x_k(t)\Big). \tag{13}$$

Using this representation we choose an arbitrary $\Phi(z)x_i(t-1) \in \mathcal{H}_{t-1}^{(j)}$ and show that

$$\langle x_i(t) - \hat{\mathbb{E}}[x_i(t) \mid \mathcal{H}_{t-1}^{(i)}], \Phi(z)x_j(t-1) - \hat{\mathbb{E}}[\Phi(z)x_j(t-1) \mid \mathcal{H}_{t-1}^{(i)}]\rangle = 0, \tag{14}$$

which will imply (by Theorem 1) that $j \overset{\text{PW}}{\nrightarrow} i$ and for which it is equivalent to show that

$$\langle x_i(t), \Phi(z)x_j(t-1) - \hat{\mathbb{E}}[\Phi(z)x_j(t-1) \mid \mathcal{H}_{t-1}^{(i)}]\rangle = 0, \tag{15}$$

by the orthogonality principle since $\hat{\mathbb{E}}[x_i(t) \mid \mathcal{H}_{t-1}^{(i)}] \in \mathcal{H}_{t-1}^{(i)}$. Substituting (13) into (15) and starting with the first term we have

$$\begin{aligned}
&\langle x_i(t), \Phi(z)\mathsf{A}_{jj}(z)v_j(t-1) - \hat{\mathbb{E}}[\Phi(z)\mathsf{A}_{jj}(z)v_j(t-1) \mid \mathcal{H}_{t-1}^{(i)}]\rangle \\
&\overset{(\alpha)}{=} \langle \mathsf{A}_{ii}(z)v_i(t) + \sum_{k \in \mathcal{A}(i)} \mathsf{A}_{ik}(z)v_k(t), \Phi(z)\mathsf{A}_{jj}(z)v_j(t-1)\rangle \\
&\overset{(\beta)}{=} 0,
\end{aligned}$$

where (α) follows by expanding $x_i(t)$ with (9) and $\hat{\mathbb{E}}[\Phi(z)\mathsf{A}_{jj}(z)v_j(t-1) \mid \mathcal{H}_{t-1}^{(i)}] = 0$ because $\forall \tau, s$

$$\mathbb{E}v_j(t-\tau)x_i(t-s) = \mathbb{E}v_j(t-\tau)\sum_{k \in \mathcal{A}(i) \cup \{i\}} \mathsf{A}_{ik}(z)v_k(t-s) = 0,$$

since $j \notin \mathcal{A}(i)$; (β) follows similarly, that is, $j \notin \mathcal{A}(i)$. Secondly we see that $\forall k \in C_0(j)$

$$\langle x_i(t), \Phi(z)\mathsf{A}_{jj}(z)\mathsf{B}_{ik}(z)x_k(t-1) - \hat{\mathbb{E}}[\Phi(z)\mathsf{A}_{jj}(z)\mathsf{B}_{ik}(z)x_k(t-1) \mid \mathcal{H}_{t-1}^{(i)}]\rangle = 0,$$

which follows from Lemma 5. Finally, for $k \in C_1(j)$ the case $k = i$ is immediate (since the error in estimating $\Phi(z)\mathsf{B}_{ii}(z)x_i(t)$ given $\mathcal{H}_{t-1}^{(i)}$ is 0), so suppose $k \neq i$. We know from above that k inherits the key properties referred to as (a) and (b) above and therefore we can recursively expand k in the same way as in equation (13). Continuing this recursion for each $k \in C_1(j)$ (where $k \neq i$) must eventually terminate since $i \in \mathcal{A}(k)$. □

An interesting corollary is the following:

Corollary 1. *If the graph $\mathcal{G}$ is a DAG then* $j \overset{PW}{\to} i, i \overset{PW}{\to} j \implies \exists k \in \mathcal{A}(i) \cap \mathcal{A}(j)$ *confounding* (i, j).

It seems reasonable to expect a converse of proposition 2 to hold, i.e. $j \in \mathcal{A}(i) \Rightarrow j \overset{\text{PW}}{\to} i$. Unfortunately, this is not the case in general, as different paths through $\mathcal{G}$ can lead to cancellation (see example 1). In fact, we do not even have $j \in pa(i) \Rightarrow j \overset{\text{PW}}{\to} i$ (see example 2).

Example 1. Firstly, on $n = 4$ nodes, "diamond" shapes can lead to cancellation on paths of length 2:

$$x(t) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ a & 0 & 0 & 0 \\ -a & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{bmatrix} x(t-1) + v(t),$$

with $\mathbb{E}v(t) = 0,\ \mathbb{E}v(t)v(t-\tau)^{\mathsf{T}} = \delta_\tau I.$
By directly calculating

$$\begin{aligned} x_4(t) &= x_2(t-1) + x_3(t-1) + v_4(t) \\ &= ax_1(t-2) + av_2(t-1) - ax_1(t-2) - av_3(t-1) + v_4(t) \\ &= a(v_2(t-1) - v_3(t-1)) + v_4(t), \end{aligned}$$

we see that, since $v(t)$ is isotropic white noise, $1 \overset{\text{PW}}{\nrightarrow} 4$. The problem here is that there are multiple paths from x_1 to x_4.

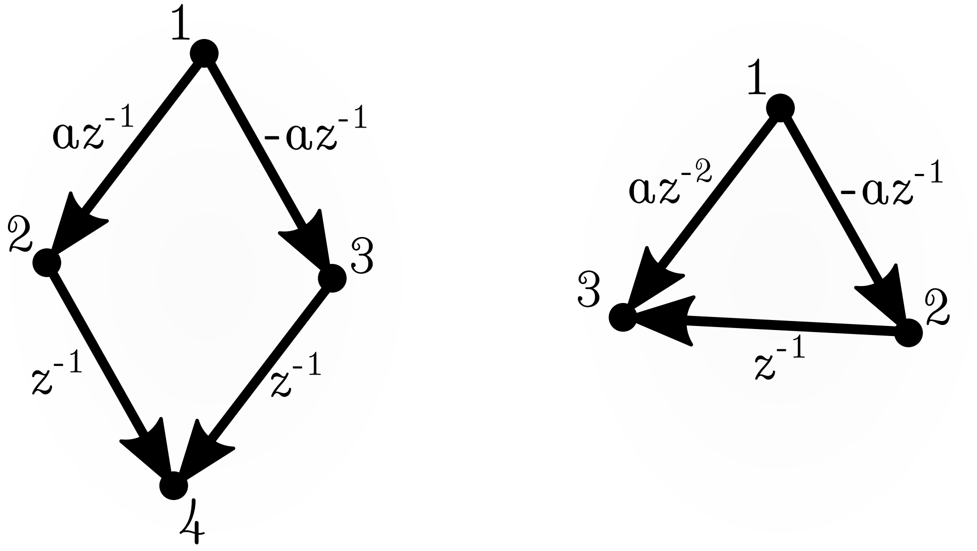
Example 2. A second example on $n = 3$ nodes is also worth examining, in this case cancellation is a result of differing time lags.

$$x(t) = \begin{bmatrix} 0 & 0 & 0 \\ -a & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} x(t-1) + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ a & 0 & 0 \end{bmatrix} x(t-2) + v(t)$$

Then

$$\begin{aligned} x_2(t) &= v_2(t) - ax_1(t-1) \\ x_3(t) &= v_3(t) + x_2(t-1) + ax_1(t-2) \\ \Rightarrow x_3(t) &= v_2(t-1) + v_3(t), \end{aligned}$$

and again $1 \overset{\text{PW}}{\nrightarrow} 3$.



(a) Graph Corresponding to Example 1
$j \in \mathcal{A}(i) \not\Rightarrow j \overset{\text{PW}}{\to} i$

(b) Graph Corresponding to Example 2
$j \in pa(i) \not\Rightarrow j \overset{\text{PW}}{\to} i$

2.5 Strongly Causal Graphs

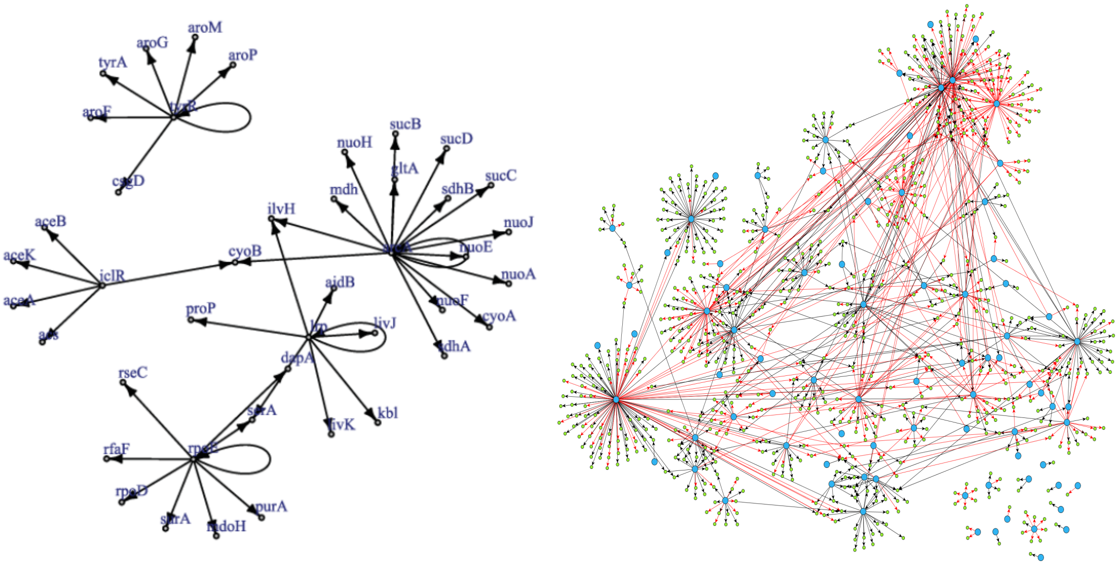
In this section and the next we will seek to understand when converse statements of Proposition 2 *do* hold. One possibility is to restrict the coefficients of the system matrix, e.g. by requiring that $B_{ij}(\tau) \geq 0$. Instead, we think it more meaningful to focus on the defining feature of time series networks, that is, the topology of $\mathcal{G}$.

Definition 8 (Strongly Causal)**.** We will say that a Granger-causality graph $\mathcal{G}$ is *strongly causal* if there is at most 1 directed path between any two nodes. Strongly Causal Graphs will be referred to as SCGs.

Examples of strongly causal graphs include directed trees (or forests), DAGs where each node has at most one parent, and figure 3 of this paper. A complete bipartite graph with $2n$ nodes is also strongly causal, demonstrating that the number of edges of such a graph can still scale quadratically with the number of nodes. It is evident that the strong causal property is inherited by subgraphs.

Example 3. Though examples of SCGs are easy to construct in theory, should practitioners expect SCGs to arise in application? While a positive answer to this question is not *necessary* for the concept to be useful, it is certainly sufficient. Though the answer is likely to depend upon the particular application area, examples appear to be available in biology, in particular, the authors of [10] cite an example of the so called "transcription regulatory network of *E.coli*", and [30] study a much larger regulatory network of *Saccharomyces cerevisiae*. These networks, which we reproduce[3] in figure 2 appear to have at most a small number of edges which violate the strong-causality condition.

Figure 2: Transcription Regulatory Networks



(a) *E.Coli* Network of [10]

(b) *Saccharomyces cerevisiae* Network of [30]

For later use, and to get a feel for the topological implications of strong causality, we explore a number of properties of such graphs before moving into the main result of this section. The following important property essentially strengthens proposition 2 for the case of strongly causal graphs.

Proposition 3. *In a strongly causal graph if $j \in \mathcal{A}(i)$ then any $k \in \mathcal{A}(i) \cap \mathcal{A}(j)$ is not a confounder, that is, the unique path from k to i contains j.*

Proof. Suppose that there is a path from k to i which does not contain j. In this case, there are multiple paths from k to i (one of which *does* go through j since $j \in \mathcal{A}(i)$) which contradicts the assumption of strong causality. $\square$

Corollary 2. *If $\mathcal{G}$ is a strongly causal DAG then $i \overset{PW}{\to} j$ and $j \in \mathcal{A}(i)$ are alternatives, that is $i \overset{PW}{\to} j \Rightarrow j \notin \mathcal{A}(i)$.*

Proof. Suppose that $i \overset{\text{PW}}{\to} j$ and $j \in \mathcal{A}(i)$. Then since $\mathcal{G}$ is acyclic $i \notin \mathcal{A}(j)$, and by proposition 2 there is some $k \in \mathcal{A}(i) \cap \mathcal{A}(j)$ which is a confounder. However, by proposition 3 k cannot be a confounder, a contradiction. $\square$

[3] Figure 2a is reproduced under the Creative Commons Attribution Non-Commercial License (`http://creativecommons.org/licenses/by-nc/2.5`) and Figure 2b under the Creative Commons Attribution License (`https://creativecommons.org/licenses/by/4.0/`)

Corollary 3. *If $\mathcal{G}$ is a strongly causal DAG such that $i \overset{PW}{\to} j$ and $j \overset{PW}{\to} i$, then $i \notin \mathcal{A}(j)$ and $j \notin \mathcal{A}(i)$. In particular, a pairwise bidirectional edge indicates the absence of any edge in $\mathcal{G}$.*

Proof. This follows directly from applying proposition 2 to $i \overset{\text{PW}}{\to} j$ and $j \overset{\text{PW}}{\to} i$. □

In light of proposition 3, the following provides a partial converse to proposition 2, and supports the intuition of "causal flow" through paths in $\mathcal{G}$.

Proposition 4. *If $\mathcal{G}$ is a strongly causal DAG then $j \in \mathcal{A}(i) \Rightarrow j \overset{PW}{\to} i$.*

Proof. We will show that for some $\psi \in \mathcal{H}_{t-1}^{(j)}$ we have

$$\langle \psi - \hat{\mathbb{E}}[\psi \mid \mathcal{H}_{t-1}^{(i)}], x_i(t) - \hat{\mathbb{E}}[x_i(t) \mid \mathcal{H}_{t-1}^{(i)}] \rangle \neq 0 \tag{16}$$

and therefore that $H_t^{(i)} \not\perp \mathcal{H}_{t-1}^{(j)} \mid \mathcal{H}_{t-1}^{(i)}$, which by theorem (1) is enough to establish that $j \overset{\text{PW}}{\to} i$.

Firstly, we will establish a representation of $x_i(t)$ that involves $x_j(t)$. Denote by $a_{r+1} \to a_r \to \cdots \to a_1 \to a_0$ with $a_{r+1} \overset{\Delta}{=} j$ and $a_0 \overset{\Delta}{=} i$ the *unique* $j \to \cdots \to i$ path in $\mathcal{G}$, we will expand the representation of equation (8) backwards along this path:

$$\begin{aligned}
x_i(t) &= v_i(t) + \mathsf{B}_{ii}(z)x_i(t) + \sum_{k \in pa(i)} \mathsf{B}_{ik}(z)x_k(t) \\
&= \underbrace{v_{a_0}(t) + \mathsf{B}_{a_0a_0}(z)x_i(t) + \sum_{\substack{k \in pa(a_0) \\ k \neq a_1}} \mathsf{B}_{a_0k}(z)x_k(t)}_{\overset{\Delta}{=}\widetilde{\alpha}(a_0,a_1)} + \mathsf{B}_{a_0a_1}(z)x_{a_1}(t) \\
&= \widetilde{\alpha}(a_0, a_1) + \mathsf{B}_{a_0a_1}(z)\big[\widetilde{\alpha}(a_1, a_2) + \mathsf{B}_{a_1a_2}(z)x_{a_2}(t)\big] \\
&\overset{(a)}{=} \sum_{\ell=0}^{r} \underbrace{\Big(\prod_{m=0}^{\ell-1} \mathsf{B}_{a_ma_{m+1}}(z)\Big)}_{\overset{\Delta}{=}F_\ell(z)} \widetilde{\alpha}(a_\ell, a_{\ell+1}) + \Big(\prod_{m=0}^{r} \mathsf{B}_{a_ma_{m+1}}(z)\Big)x_{a_{r+1}}(t) \\
&= \sum_{\ell=0}^{r} F_\ell(z)\widetilde{\alpha}(a_\ell, a_{\ell+1}) + F_{r+1}(z)x_j(t)
\end{aligned}$$

where (a) follows by a routine induction argument and where we define $\prod_{m=0}^{-1} \bullet \overset{\Delta}{=} 1$ for notational convenience.

Using this representation to expand equation (16), we obtain the following cumbersome expression:

$$\begin{aligned}&\langle \psi - \hat{\mathbb{E}}[\psi \mid \mathcal{H}_{t-1}^{(i)}], F_{r+1}(z)x_j(t) - \hat{\mathbb{E}}[F_{r+1}(z)x_j(t) \mid \mathcal{H}_{t-1}^{(i)}]\rangle\\ &- \langle \psi - \hat{\mathbb{E}}[\psi \mid \mathcal{H}_{t-1}^{(i)}], \hat{\mathbb{E}}[\sum_{\ell=0}^{r} F_\ell(z)\widetilde{\alpha}(a_\ell, a_{\ell+1}) \mid \mathcal{H}_{t-1}^{(i)}]\rangle\\ &+ \langle \psi - \hat{\mathbb{E}}[\psi \mid \mathcal{H}_{t-1}^{(i)}], \sum_{\ell=0}^{r} F_\ell(z)\widetilde{\alpha}(a_\ell, a_{\ell+1})\rangle.\end{aligned}$$

Note that by the orthogonality principle, $\psi - \hat{\mathbb{E}}[\psi \mid \mathcal{H}_{t-1}^{(i)}] \perp \mathcal{H}_{t-1}^{(i)}$, the middle term above is 0. Choosing now the particular value $\psi = F_{r+1}(z)x_j(t) \in \mathcal{H}_{t-1}^{(j)}$ we arrive at

$$\begin{aligned}&\langle \psi - \hat{\mathbb{E}}[\psi \mid \mathcal{H}_{t-1}^{(i)}], x_i(t) - \hat{\mathbb{E}}[x_i(t) \mid \mathcal{H}_{t-1}^{(i)}]\rangle\\ &= \mathbb{E}|F_{r+1}(z)x_j(t) - \hat{\mathbb{E}}[F_{r+1}(z)x_j(t) \mid \mathcal{H}_{t-1}^{(i)}]|^2\\ &+ \langle F_{r+1}(z)x_j(t) - \hat{\mathbb{E}}[F_{r+1}(z)x_j(t) \mid \mathcal{H}_{t-1}^{(i)}], \sum_{\ell=0}^{r} F_\ell(z)\widetilde{\alpha}(a_\ell, a_{\ell+1})\rangle,\end{aligned}$$

which by the Cauchy-Schwarz inequality is 0 if and only if

$$\sum_{\ell=0}^{r} F_\ell(z)\widetilde{\alpha}(a_\ell, a_{\ell+1}) \overset{\text{a.s.}}{=} \hat{\mathbb{E}}[F_{r+1}(z)x_j(t) \mid \mathcal{H}_{t-1}^{(i)}] - F_{r+1}(z)x_j(t),$$

or by rearranging and applying the representation obtained earlier, if and only if

$$x_i(t) \overset{\text{a.s.}}{=} \hat{\mathbb{E}}[F_{r+1}(z)x_j(t) \mid \mathcal{H}_{t-1}^{(i)}],$$

but this is impossible since $x_i(t) \notin \mathcal{H}_{t-1}^{(i)}$. □

We immediately obtain the corollary, which we remind the reader is, surprisingly, not true in a general graph.

Corollary 4. *If $\mathcal{G}$ is a strongly causal DAG then $j \overset{GC}{\to} i \Rightarrow j \overset{PW}{\to} i$.*

Remark 3. As we have seen and as is true in much of statistics, confounding nodes pose challenges for Granger-causality. However, as opposed to Pearl's causal calculus [31], pairwise Granger-causality does not suffer any difficulty with so-called "colliders", that is, the topology $i \to k \leftarrow j$ will never result in $i \overset{\text{PW}}{\to} j$ or $j \overset{\text{PW}}{\to} i$. This is evidently an advantage of the *temporal* nature of Granger-causality – there is no backwards causal flow along the edges of $\mathcal{G}$.

Example 4. As a final remark of this subsection we note that a complete converse to proposition 2 is not possible without additional conditions. Consider the "fork" system on 3 nodes (i.e. $2 \leftarrow 1 \to 3$) defined by

$$x(t) = \begin{bmatrix} 0 & 0 & 0 \\ a & 0 & 0 \\ a & 0 & 0 \end{bmatrix} x(t-1) + v(t).$$

In this case, node 1 is a confounder for nodes 2 and 3, but $x_3(t) = v_3(t) - v_2(t) + x_2(t)$ and $2 \overset{\text{PW}}{\nrightarrow} 3$ (even though $x_2(t)$ and $x_3(t)$ are contemporaneously correlated)

If we were to augment this system by simply adding an autoregressive component (i.e. some "memory") to $x_1(t)$ e.g. $x_1(t) = v_1(t) + bx_1(t-1)$ then we *would* have $2 \overset{\text{PW}}{\rightarrow} 3$ since then $x_3(t) = v_3(t) + av_1(t-1) - bv_2(t-1) + bx_2(t-1)$. We develop this idea further in the next section.

2.6 Persistent Systems

In section 2.5 we obtained a converse to part (a) of proposition 2 via the notion of a strongly causal graph topology. In this section, we complete a converse by adding the additional requirement we refer to as "persistence".

Definition 9 (Lag Function)**.** Given a causal filter $\mathsf{B}(z) = \sum_{\tau=0}^{\infty} b(\tau) z^{-\tau}$ define

$$\tau_0(\mathsf{B}) = \min\{\tau \in \mathbb{Z}_+ \mid b(\tau) \neq 0\}, \tag{17}$$

$$\tau_\infty(\mathsf{B}) = \sup\{\tau \in \mathbb{Z}_+ \mid b(\tau) \neq 0\}. \tag{18}$$

$$\tag{19}$$

i.e. the "first" and "last" coefficients of the filter $\mathsf{B}(z)$, where $\tau_\infty(\mathsf{B}) \overset{\Delta}{=} \infty$ if the filter has an infinite length, and $\tau_0(\mathsf{B}) \overset{\Delta}{=} \infty$ if $\mathsf{B}(z) = 0$.

This interpretation of the following persistence condition is that each node stores some "memory" of the past.

Definition 10 (Persistent)**.** We will say that the process $x(t)$ with Granger-causality graph $\mathcal{G}$ is *persistent* if for every $i \in [n]$ and every $k \in \mathcal{A}(i)$ we have $\tau_0(\mathsf{A}_{ik}) < \infty$ and $\tau_\infty(\mathsf{A}_{ik}) = \infty$.

Remark 4. In the context of Granger-causality, "most" systems should be persistent. In particular, $\mathsf{VAR}(p)$ models are likely to be persistent since these naturally result in an equivalent $\mathsf{MA}(\infty)$ representation, except for the pathological case where $\mathsf{B}(z)$ is Nilpotent.

Moreover, persistence is not the weakest condition necessary for the results of this section, the condition $\tau_0(\mathsf{A}_{jk}) < \tau_\infty(\mathsf{A}_{ik})$ for each i, j, k such that $k \in \mathcal{A}(i) \cap \mathcal{A}(j)$ is enough. The intuition being that nodes i and j are not receiving temporally disjoint information from k.

Example 5. process $x(t)$ generated by the $\mathsf{VAR}(1)$ model[4] having $\mathsf{B}(z) = Bz^{-1}$, we will examine conditions that guarantee persistence. Pick any $i \in [n], j \in \mathcal{A}(i) \setminus \{i\}$, then the stability of B allows us to write

$$\mathsf{A}(z) = \sum_{k=0}^{\infty} B^k z^{-k},$$

[4] Recall that any $\mathsf{VAR}(p)$ model with $p < \infty$ can be written as a $\mathsf{VAR}(1)$ model, so we lose little generality in considering this case.

whereby we see that $\exists k > 0$ such that $[B^k]_{ij} \neq 0$ (since $j \in \mathcal{A}(i)$). Then consider

$$\begin{aligned} e_i^{\mathsf{T}} B^{rk} e_j &\overset{(a)}{=} \big((P^{\mathsf{T}} e_i)^{\mathsf{T}} J^{rk} P^{-1} e_j\big) \\ &= \mathsf{tr}[(P^{\mathsf{T}} e_i)^{\mathsf{T}} J^{rk} P^{-1} e_j] \\ &\overset{(b)}{=} \mathsf{tr}[(J^{rk})(vu^{\mathsf{T}})], \end{aligned}$$

where (a) utilizes the Jordan Normal Form of B, and (b) denotes $u = P^{\mathsf{T}} e_i$ and $v = P^{-1} e_j$. In order for $\tau_\infty(\mathsf{A}_{ij}) < \infty$, there must be some $N > 1$ such that $\forall r \geq N$, the above term is 0. This may be the case for instance if B is a nilpotent matrix.

Let us suppose now that B is diagonalizable (i.e. J is a diagonal matrix) with at least 2 distinct eigenvalues; in this case B is also *not* nilpotent. We can then rewrite the above as

$$f(r) \overset{\Delta}{=} \mathsf{tr}[(J^{rk})(vu^{\mathsf{T}})] = \sum_{\nu=1}^{n} \lambda_\nu^{rk} v_\nu u_\nu \overset{\Delta}{=} \sum_{\nu=1}^{n} \lambda_\nu^{rk} \beta_\nu$$

where λ_ν denotes the eigenvalues of B and $\beta_\nu = u_\nu v_\nu$. Note that $f(0) = 0$ since $i \neq j$, u is a row of P and v is a column of P^{-1}. Moreover, $f(1) \neq 0$ by hypothesis. But, in order for $f(r) = 0 \; \forall r \geq N$, it would need to be the case that

$$\mathsf{Dg}(\boldsymbol{\lambda})^r \boldsymbol{\lambda} = Vz$$

had a solution in z for every $r \geq N$, where V is an $n \times n - 1$ full-rank matrix whose columns span the nullspace of β, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$. That is, iterates of $\mathsf{Dg}(\boldsymbol{\lambda})$ applied to $\boldsymbol{\lambda}$ would need to remain inside β's nullspace. This would imply that

$$VV^{\dagger} \boldsymbol{\lambda}^{r+1} = \boldsymbol{\lambda}^{r+1},$$

i.e. that $\boldsymbol{\lambda}^{r+1}$ is an eigenvector of $VV^{\dagger}$ for an infinite number of integers r (the exponentiation is to be understood as a pointwise operation). However, since there can only be a finite number of (unit length) eigenvectors, this cannot be the case unless every eigenvalue $(\lambda_1, \ldots, \lambda_n)$ were equal.

We see from this example that the collection of $\mathsf{VAR}(1)$ systems which are not persistent are pathological, in the sense that their system matrices have zero measure when viewed as a subset of $\mathbb{R}^{n^2}$.

Lemma 6. *Suppose $v(t)$ is a scalar sequence with unit variance and zero autocorrelation and let $\mathsf{A}(z), \mathsf{B}(z)$ be nonzero and strictly causal (i.e. $\tau_0(\mathsf{A}) \geq 1$) linear filters. Then,*

$$\langle F(z)\mathsf{A}(z)v(t), \mathsf{B}(z)v(t)\rangle = 0 \; \forall \textit{ strictly causal filters } F(z) \tag{20}$$

if and only if $\tau_0(\mathsf{A}) \geq \tau_\infty(\mathsf{B})$.

Proof. We have

$$\begin{align}
\langle \mathsf{A}(z)v(t), \mathsf{B}(z)v(t)\rangle &= \sum_{\tau=1}^{\infty}\sum_{s=1}^{\infty} a(\tau)b(s)\mathbb{E}[v(t-s)v(t-\tau)] \tag{21}\\
&= \sum_{\tau=\max(\tau_0(\mathsf{A}),\tau_0(\mathsf{B}))}^{\min(\tau_\infty(\mathsf{A}),\tau_\infty(\mathsf{B}))} a(\tau)b(\tau) \tag{22}\\
& \tag{23}
\end{align}$$

due to the uncorrelatedness assumptions on $v(t)$. This expression is 0 if and only if $\tau_0(\mathsf{A}) \geq 1 + \tau_\infty(\mathsf{B})$ or if $\tau_0(\mathsf{B}) \geq 1 + \tau_\infty(\mathsf{A})$ or if the coefficients are orthogonal along the common support.

Specializing this fact to $\langle F(z)\mathsf{A}(z)v(t), \mathsf{B}(z)v(t)\rangle = 0$ we see that the coefficients cannot be orthogonal for every choice of F, and that $\sup_F \tau_\infty(F\mathsf{A}) = \infty$, leaving only the possibility that

$$\begin{aligned}
\tau_0(F\mathsf{A}) \geq 1 + \tau_\infty(\mathsf{B}) \forall F \quad &\overset{(a)}{\iff} \tau_0(\mathsf{A}) \geq 1 + \tau_\infty(\mathsf{B}) - \min_F \tau_0(F)\\
&\overset{(b)}{\iff} \tau_0(\mathsf{A}) \geq \tau_\infty(\mathsf{B}),
\end{aligned}$$

where (a) follows since $\tau_0(F\mathsf{A}) = \tau_0(F) + \tau_0(\mathsf{A})$, and (b) since $\min_F \tau_0(F) = 1$. □

Corollary 5. *For $k \in \mathcal{A}(i) \cap \mathcal{A}(j)$ we have*

$$\begin{aligned}
&\hat{\mathbb{E}}[F(z)\mathsf{A}_{jk}(z)v_k(t) \mid \mathcal{H}_{t-1}^{(i)}] = 0 \ \forall \textit{ strictly causal } F(z)\\
\iff &\langle F(z)\mathsf{A}_{jk}(z)v_k(t), \mathsf{A}_{ik}(z)v_k(t)\rangle = 0 \ \forall \textit{ strictly causal } F(z)\\
\iff &\tau_0(\mathsf{A}_{jk}) \geq \tau_\infty(\mathsf{A}_{ik})
\end{aligned}$$

Proof. The final equivalence follows immediately from Lemma 6. For the first equivalence we have

$$\begin{aligned}
&\hat{\mathbb{E}}[F(z)\mathsf{A}_{jk}(z)v_k(t) \mid \mathcal{H}_{t-1}^{(i)}] = 0 \ \forall \text{ strictly causal } F(z)\\
\iff &\langle F(z)A_{jk}(z)v_k(t), x_i(t-\tau)\rangle = 0 \ \forall \tau \geq 1, \text{ strictly causal } F(z),
\end{aligned}$$

which can be expanded by equation (9) to obtain (after cancelling all ancestors of i other than k)

$$\langle F(z)A_{jk}(z)v_k(t), \mathsf{A}_{ik}(z)v_k(t-\tau)\rangle = 0 \ \forall \tau \geq 1, \text{ strictly causal } F(z),$$

which by the Lemma is equivalent to $\tau_0(\mathsf{A}_{jk}) \geq \tau_\infty(\mathsf{A}_{ik})$ as stated. □

Proposition 5. *Suppose $\mathcal{G}$ is a strongly causal DAG and that $x(t)$ is persistent, then if there exists a k which confounds (i, j) we have $i \overset{PW}{\to} j$ and $j \overset{PW}{\to} i$.*

Proof. We will show that $j \overset{\text{PW}}{\to} i$, the other being symmetric. First note also that by proposition 3 we cannot have $i \in \mathcal{A}(j)$ or $j \in \mathcal{A}(i)$ and therefore every $k \in \mathcal{A}(i) \cap \mathcal{A}(j)$ will be a confounder.

It is sufficient to show that $\exists \psi \in \mathcal{H}_{t-1}^{(j)}$ such that

$$\langle \psi - \hat{\mathbb{E}}[\psi \mid \mathcal{H}_{t-1}^{(i)}], x_i(t) - \hat{\mathbb{E}}[x_i(t) \mid \mathcal{H}_{t-1}^{(i)}]\rangle \neq 0.$$

To this end, let $F(z)$ be an arbitrary but strictly causal linear filter. We apply equation (9) to $x_i(t)$ and $\psi \overset{\Delta}{=} F(z)x_j(t)$:

$$\begin{aligned}
&\langle \psi - \hat{\mathbb{E}}[\psi \mid \mathcal{H}_{t-1}^{(i)}], x_i(t) - \hat{\mathbb{E}}[x_i(t) \mid \mathcal{H}_{t-1}^{(i)}]\rangle \\
&\overset{(a)}{=} \langle \psi - \hat{\mathbb{E}}[\psi \mid \mathcal{H}_{t-1}^{(i)}], \mathsf{A}_{ii}(z)v_i(t) + \sum_{k \in \mathcal{A}(i)} \mathsf{A}_{ik}(z)v_k(t)\rangle \\
&\overset{(b)}{=} \langle \sum_{k \in \mathcal{A}(j)} \big(F(z)\mathsf{A}_{jk}(z)v_k(t) - \hat{\mathbb{E}}[F(z)\mathsf{A}_{jk}(z)v_k(t) \mid \mathcal{H}_{t-1}^{(i)}]\big), \mathsf{A}_{ii}(z)v_i(t) + \sum_{k \in \mathcal{A}(i)} \mathsf{A}_{ik}(z)v_k(t)\rangle \\
&\overset{(c)}{=} \sum_{k \in \mathcal{A}(i) \cap \mathcal{A}(j)} \Big(\langle F(z)\mathsf{A}_{jk}(z)v_k(t), \mathsf{A}_{ik}(z)v_k(t)\rangle - \langle \hat{\mathbb{E}}[F(z)\mathsf{A}_{jk}(z)v_k(t) \mid \mathcal{H}_{t-1}^{(i)}], \mathsf{A}_{ii}v_i(t)\rangle \\
&- \sum_{\ell \in \mathcal{A}(i)} \langle \hat{\mathbb{E}}[F(z)\mathsf{A}_{jk}(z)v_k(t) \mid \mathcal{H}_{t-1}^{(i)}], \mathsf{A}_{i\ell}(z)v_\ell(t)\rangle\Big)
\end{aligned}$$

where in (a) we have removed the $\hat{\mathbb{E}}[x_i(t) \mid \mathcal{H}_{t-1}^{(i)}]$ term via the orthogonality principle, in (b) there is no $F(z)\mathsf{A}_{jj}(z)v_j(t)$ term since due to $j \not\in \mathcal{A}(i)$ it is orthogonal to $\mathcal{H}_t^{(i)}$. Finally, (c) follows by applying orthogonality properties of $v(t)$, as well as the fact that $\hat{\mathbb{E}}[F(z)\mathsf{A}_{jk}(z)v_k(t) \mid \mathcal{H}_{t-1}^{(i)}] = 0$ for $k \not\in \mathcal{A}(i)$. Note that $\hat{\mathbb{E}}[F(z)\mathsf{A}_{jk}(z)v_k(t) \mid \mathcal{H}_{t-1}^{(i)}] \in \mathcal{H}_{t-1}^{(i)}$ and thence there is in general no cancellation in the final term above for $\ell \in \mathcal{A}(i)$.

This is clearly not immediately evident

This is 0 for every F if and only if for all F and $\forall k \in \mathcal{A}(i) \cap \mathcal{A}(j)$ we have

$$\hat{\mathbb{E}}[F(z)\mathsf{A}_{jk}(z)v_k(t) \mid \mathcal{H}_{t-1}^{(i)}] = 0$$

and

$$\langle F(z)\mathsf{A}_{jk}(z)v_k(t), \mathsf{A}_{ik}(z)v_k(t)\rangle = 0,$$

which by Corollary 5 occurs if and only if $\tau_0(\mathsf{A}_{jk}) \geq \tau_\infty(\mathsf{A}_{ik})$, which is impossible since by persistence $\tau_0(\mathsf{A}_{jk}) < \infty$ and $\tau_\infty(\mathsf{A}_{ik}) = \infty$. □

2.7 Recovering $\mathcal{G}$ via Pairwise Tests

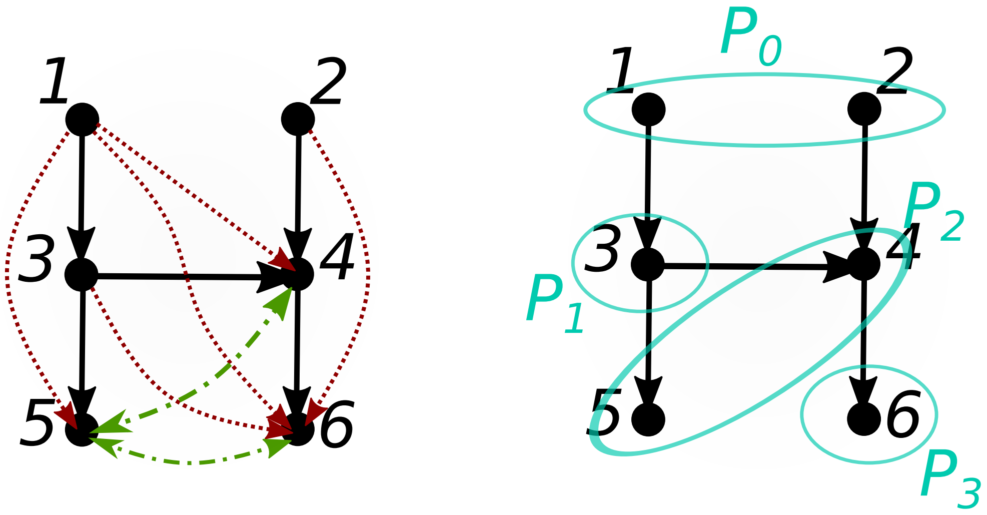
In this section we will show that if the $\mathcal{G}$ of a persistent process is a strongly causal DAG, then it is possible to recover $\mathcal{G}$ via pairwise tests alone. This is the main conclusion of the theoretical analysis in this paper.

Algorithm 1: Pairwise Graph Recovery

```
input     : Pairwise Granger-causality relations between a persistent process of
            dimension $n$ whose joint Granger-causality relations are known to form a
            strongly causal DAG $\mathcal{G}$.
output    : Edges $\mathcal{E} = \{(i, j) \in [n] \times [n] \mid i \overset{\text{GC}}{\to} j\}$ of the graph $\mathcal{G}$.
initialize: $S_0 = [n]$ # unprocessed nodes
            $E_0 = \emptyset$ # edges of $\mathcal{G}$
            $k = 1$ # a counter used only for notation

$W \leftarrow \{(i, j) \mid i \overset{\text{PW}}{\to} j, j \overset{\text{PW}}{\nrightarrow} i\}$ # candidate edges
$P_0 \leftarrow \{i \in S_0 \mid \forall s \in S_0\ (s, i) \notin W\}$ # parent-less nodes
while $S_{k-1} \neq \emptyset$ do
    $S_k \leftarrow S_{k-1} \setminus P_{k-1}$ # remove nodes with depth $k-1$
    $P_k \leftarrow \{i \in S_k \mid \forall s \in S_k\ (s, i) \notin W\}$ # candidate children of $P_{k-1}$

    $D_{k0} \leftarrow \emptyset$
    for $r = 1, \ldots, k$ do
        $Q \leftarrow E_{k-1} \cup \left(\bigcup_{\ell=0}^{r-1} D_{k\ell}\right)$ # currently known edges
        $D_{kr} \leftarrow \{(i, j) \in P_{k-r} \times P_k \mid (i, j) \in W,$ no $i \to \cdots \to j$ path in $Q\}$
    $E_k \leftarrow E_{k-1} \cup \left(\bigcup_{r=1}^{k} D_{kr}\right)$ # update $E_k$ with new edges

    $k \leftarrow k + 1$
return $E_{k-1}$
```

Figure 3: Example graph for Algorithm 1

Black arrows indicate true parent-child relations. Red dotted arrows indicate pairwise causality (due to non-parent relations), green dash-dotted arrows indicate bidirectional pairwise causality (due to the confounding node 1). Blue groupings indicate each P_k in Algorithm 1.



Theorem 2 (Pairwise Recovery)**.** *If the Granger-causality graph $\mathcal{G}$ for persistent process $x(t)$ is a strongly causal DAG then $\mathcal{G}$ can be inferred from pairwise causality tests. The procedure can be carried out, assuming we have an oracle for pairwise causality, via Algorithm (1).*

We will shortly prove the theorem by establishing the correctness of algorithm (1). The idea is to iteratively "peel away layers" of nodes by removing the nodes that have no parents remaining, which always exist since the graph is acyclic. The requirement of strong causality ensures that all actual edges of $\mathcal{G}$ manifest in some way as pairwise relations (by proposition 4), and the requirement of persistence allows confounding to be eliminated by removing bidirectional edges. Without persistence, each confounded pair would give rise to 4 possible pairwise topologies consistent with $\mathcal{G}$, one for each type of pairwise edge (no edge, unidirectional, bidirectional).

Example 6. The set W collects ancestor relations in $\mathcal{G}$ (see Lemma 7). In reference to figure 3, each of the solid black edges, as well as the dotted red edges will be included in W, but *not* the bidirectional green dash-dotted edges, which we are able to exclude by appealing to Corollary 3. The groupings $P_0, \ldots, P_3$ are also indicated in figure 3.

The algorithm proceeds first with the parentless nodes $1, 2$ on the initial iteration where the edge $(1, 3)$ is added to E. On the next iteration, the edges $(3, 4), (2, 4), (3, 5)$ are added, and the false edges $(1, 4), (1, 5)$ are excluded due to the paths $1 \rightarrow 3 \rightarrow 4$ and $1 \rightarrow 3 \rightarrow 5$ already being present. Finally, edge $(4, 6)$ is added, and the false $(1, 6), (3, 6), (2, 6)$ edges are similarly excluded due to the ordering of the inner loop.

That we need to proceed backwards through P_{k-r} as in the inner loop of on r can also be seen from this example, where if instead we simply added the set

$$D'_k = \{(i,j) \in \Big(\bigcup_{r=1}^{k} P_{k-r}\Big) \times P_k \mid i \overset{\mathrm{PW}}{\to} j\}$$

to E_k then we would infer the false positive edge 1 → 4. Moreover, the same example shows that simply using the set

$$D''_k = \{(i,j) \in P_{k-1} \times P_k \mid i \overset{\mathrm{PW}}{\to} j\},$$

causes the edge 1 → 3 to be missed.

Our proof proceeds in 5 steps stated formally as lemmas. Firstly, we characterize the sets W and P_k. Then we establish a correctness result for the inner loop on r, a correctness result for the outer loop on k, and finally that the algorithm terminates in a finite number of steps.

Lemma 7 (W Represents Ancestor Relations)**.** *In Algorithm 1 we have $(i,j) \in W$ if and only if $i \in \mathcal{A}(j)$. In particular, $W \subseteq \mathcal{E}$.*

Proof. Let $j \in [n]$ and suppose that $i \in \mathcal{A}(j)$. Then $i \overset{\mathrm{PW}}{\to} j$ by Proposition 4. Proposition 3 ensures that (i,j) are not confounded and Corollary 2 that $j \notin \mathcal{A}(i)$ so $j \overset{\mathrm{PW}}{\nrightarrow} i$ and thence by Proposition 2 $(i,j) \in W$.

Conversely, suppose $(i,j) \in W$. Then since $j \overset{\mathrm{PW}}{\nrightarrow} i$ Proposition 5 ensures that (j,i) are not confounded and so by Proposition 2 we must have $i \in \mathcal{A}(j)$. □

Definition 11 (Depth)**.** For our present purposes we will define the *depth* $d(j)$ of a node j in $\mathcal{G}$ to be the length of the *longest* path from a node in P_0 to j, where $d(j) = 0$ if $j \in P_0$. It is apparent that such a path will always exist. For example, in Figure 3 we have $d(3) = 1$ and $d(4) = 2$.

Lemma 8 (Depth Characterization of P_k)**.** *$i \in P_k \iff d(i) = k$ and $j \in S_k \iff d(j) \geq k$.*

Proof. We proceed by induction, noting that P_0 is non-empty since $\mathcal{G}$ is acyclic and therefore $\mathcal{G}$ contains nodes without parents. The base case $i \in P_0 \iff d(i) = 0$ is by definition, and $j \in S_0 \iff d(j) \geq 0$ is trivial since $S_0 = [n]$. So suppose that the lemma is true up to $k-1$.

($i \in P_k \implies d(i) = k$): Let $i \in P_k$. Suppose that $d(i) \geq k+1$, then $\exists j \in pa(i)$ such that $j \notin \cup_{r\geq 1} P_{k-r}$ (otherwise $d(i) \leq k$), this implies that $j \in S_k$ with $(j,i) \in W$ (by Lemma 7) which is not possible due to the construction of P_k and therefore $d(i) \leq k$. Moreover, $P_k \subseteq S_k \subseteq S_{k-1}$ implies that $d(i) \geq k-1$ by the induction hypothesis, but if $d(i) = k-1$ then $i \in P_{k-1}$ again by induction which is impossible since $i \in P_k$ and therefore $d(i) = k$.

($s \in S_k \implies d(s) \geq k$): Let $s \in S_k \subseteq S_{k-1}$. We have by induction that $d(s) \geq k-1$, but again by induction (this time on P_{k-1}) we have $d(s) \neq k-1$ since $S_k = S_{k-1} \setminus P_{k-1}$ and therefore $d(s) \geq k$.

($d(i) = k \implies i \in P_k$): Suppose $i \in [n]$ is such that $d(i) = k$. Then $i \in S_{k-1}$ by the hypothesis, but also $i \notin P_{k-1}$ so then $i \in S_k = S_{k-1} \setminus P_{k-1}$ and thus $d(i) \geq k$. Now, recalling the definition of P_k

$$P_k = \{i \in S_k \mid \forall s \in S_k \ (s, i) \notin W\},$$

if $s \in S_k$ is such that $(s, i) \in W$ then $s \overset{\text{PW}}{\to} i$ and $i \overset{\text{PW}}{\nrightarrow} s$ so that by persistence and Proposition 5 there cannot be a confounder of (s, i) (otherwise $i \overset{\text{PW}}{\to} s$) so then by Proposition 2 we have $s \in \mathcal{A}(i)$. We have shown that $s \in S_k \implies d(s) \geq k$ and so we must have $d(i) > k$, a contradiction, thence $s \notin \mathcal{A}(i)$, $s \overset{\text{PW}}{\nrightarrow} i$, $(s, i) \notin W$ and $i \in P_k$.

($d(j) \geq k \implies j \in S_k$): Let $j \in [n]$ such that $d(j) \geq k$, then by induction we have $j \in S_{k-1}$. This implies by the construction of S_k that $j \notin S_k$ only if $j \in P_{k-1}$, but we have shown that this only occurs when $d(j) = k - 1$, but $d(j) > k - 1$ so $j \in S_k$. □

Lemma 9 (Inner Loop). *Fix an integer $k \geq 1$ and suppose that $(i, j) \in E_{k-1}$ if and only if $(i, j) \in \mathcal{E}$ and $d(j) \leq k - 1$. Then, we have $(i, j) \in D_{kr}$ if and only if $(i, j) \in \mathcal{E}$, $d(j) = k$, and $d(i) = k - r$.*

Proof. We prove by induction on r, keeping in mind the results of Lemmas 7 and 8. For the base case, let $r = 1$ and suppose that $(i, j) \in \mathcal{E}$ with $d(j) = k$ and $d(i) = k - 1$. Then, $(i, j) \in W$ and by our assumptions on E_{k-1} there is no $i \to \cdots \to j$ path in E_{k-1} and therefore $(i, j) \in D_{k1}$. Conversely, suppose that $(i, j) \in D_{k1}$. Then, $d(i) = k - 1$ and $d(j) = k$ which, since $(i, j) \in W \implies i \in \mathcal{A}(j)$ implies that $i \in pa(j)$ and $(i, j) \in \mathcal{E}$.

Now, fix $r > 1$ and suppose that the result holds up to $r - 1$. Let $(i, j) \in \mathcal{E}$ with $d(j) = k$ and $d(i) = k - r$. Then, $(i, j) \in W$ and by induction and strong causality there cannot already be an $i \to \cdots \to j$ path in $E_{k-1} \cup \big(\bigcup_{\ell=0}^{r-1} D_{kr}\big)$, therefore $(i, j) \in D_{kr}$. Conversely, suppose $(i, j) \in D_{kr}$. Then we have $d(i) = k - r$, $d(j) = k$, and $i \in \mathcal{A}(j)$. Suppose by way of contradiction that $i \notin pa(j)$, then there must be some $u \in pa(j)$ such that $i \in \mathcal{A}(u)$. But, this implies that $d(i) < d(u)$ and by induction that $(u, j) \in \bigcup_{\ell=1}^{r-1} D_{k\ell}$. Moreover, since $d(u) < k$ (otherwise $d(j) > k$) each edge in the $i \to \cdots \to u$ path must already be in E_{k-1}, and so there must be an $i \to \cdots \to j$ path in $E_{k-1} \cup \big(\bigcup_{\ell=0}^{r-1} D_{kr}\big)$, which is a contradiction since we assumed $(i, j) \in D_{kr}$. Therefore $i \in pa(j)$ and $(i, j) \in \mathcal{E}$. □

Lemma 10 (Outer Loop). *We have $(i, j) \in E_k$ if and only if $(i, j) \in \mathcal{E}$ and $d(j) \leq k$. That is, at iteration k, E_k and $\mathcal{E}$ agree on the set of edges whose terminating node is at most k steps away from P_0.*

Proof. We will proceed by induction. The base case $E_0 = \emptyset$ is trivial, so fix some $k \geq 1$, and suppose that the lemma holds for all nodes of depth less than k.

Suppose that $(i, j) \in E_k = E_{k-1} \cup \big(\bigcup_{r=1}^{k} D_{rk}\big)$. Then clearly there is some $1 \leq r \leq k$ such that $(i, j) \in D_{kr}$ so that by Lemma 9 we have $(i, j) \in \mathcal{E}$ and $d(j) = k$.

Conversely, suppose that $(i, j) \in \mathcal{E}$ and $d(j) \leq k$. If $d(j) < k$ then by induction $(i, j) \in E_{k-1} \subseteq E_k$ so suppose further than $d(j) = k$. Since $i \in pa(j)$ we must have $d(i) < k$ (else $d(j) > k$) and again by Lemma 9 $(i, j) \in \bigcup_{r=1}^{k} D_{kr}$ which implies that $(i, j) \in E_k$. The result follows. □

Lemma 11 (Finite Termination). *Algorithm 1 terminates and returns $E_{k^\star-1} = \mathcal{E}$ for some $k^\star \leq n$.*

Proof. If $n = 1$, the algorithm is clearly correct, returning on the first iteration with $E_1 = \emptyset$. When $n > 1$ Lemma 10 ensures that E_k coincides with $\{(i, j) \in \mathcal{E} \mid d(j) \leq k\}$ and since $d(j) \leq n - 1$ for any $j \in [n]$ there is some $k^\star \leq n$ such that $E_{k^\star-1} = \mathcal{E}$. We must have $S_{k^\star} = \emptyset$ since $j \in S_{k^\star} \iff d(j) \geq k^\star$ (if $d(j) > k - 1$ then $E_{k^\star-1} \neq \mathcal{E}$) and therefore the algorithm terminates. □

Example 7. We close this section by noting that the conditions of persistence and strong causality are only sufficient conditions. For example, the complete directed graph with 2 nodes i.e.

$$B(1) = \begin{bmatrix} 1/2 & 1 \\ 1 & 1/2 \end{bmatrix}$$

contains a loop but is pairwise recoverable, though not by algorithm (1). Clearly, this example is somewhat artificial since when $n = 2$ there is no difference between pairwise Granger-causality and joint Granger-causality amongst all series – however, one can add any number of nodes having no parents or children to a graph containing a length 2 cycle, in which case the graph clearly remains pairwise recoverable.

3 Finite Sample Graph Recovery

In this section we provide a review of our methods for implementing Algorithm 1 given a *finite* sample of T data points. We apply the simplest reasonable methods in order to maintain a focus on our main contributions (i.e. Algorithm 1), more sophisticated schemes can only serve to improve the results. Textbook reviews of the following concepts are provided e.g. by [32], [33], and elsewhere.

In subsection 3.1 we define pairwise Granger-causality hypothesis tests, in subsection 3.2 a model order selection criteria, in subsection 3.3 an efficient estimation algorithm, in subsection 3.4 the method for choosing an hypothesis testing threshold, and finally in subsection 3.5 the unified finite sample algorithm.

3.1 Pairwise Hypothesis Testing

In performing pairwise checks for Granger-causality $x_j \overset{\text{PW}}{\rightarrow} x_i$ we follow the simple scheme of estimating the following two linear models:

$$H_0 : \widehat{x}_i^{(p)}(t) = \sum_{\tau=1}^{p} b_{ii}(\tau)x_i(t - \tau), \tag{24}$$

$$H_1 : \widehat{x}_i^{(p)}(t) = \sum_{\tau=1}^{p} b_{ii}(\tau)x_i(t - \tau) + \sum_{\tau=1}^{p} b_{ij}(\tau)x_j(t - \tau). \tag{25}$$

We formulate the statistic

$$F_{ij}(p) = \frac{T}{p}\Big(\frac{\xi_i(p)}{\xi_{ij}(p)} - 1\Big), \tag{26}$$

where $\xi_i(p)$ is the sample mean square of the residuals[5] $x_i(t) - \widehat{x}_i^{(p)}(t)$,

$$\xi_i(p) = \frac{1}{T-p}\sum_{t=p+1}^{T}(x_i(t) - \widehat{x}_i^{(p)}(t))^2,$$

and similarly for $\xi_{ij}(p)$. We test $F_{ij}(p)$ against a $\chi^2(p)$ distribution.

If the estimation procedure is consistent, we will have the following convergence (in $\mathbb{P}$ or a.s.):

$$F_{ij}(p) \to \begin{cases} 0; \ x_j \overset{\text{PW}}{\nrightarrow} x_i \\ \infty; \ x_j \overset{\text{PW}}{\rightarrow} x_i \end{cases} \text{as } T \to \infty. \tag{27}$$

In our finite sample implementation (see Algorithm 2) we add edges to $\widehat{\mathcal{G}}$ in order of the decreasing magnitude of F_{ij} instead of proceeding backwards through P_{k-r} in Algorithm 1. This makes greater use of the information provided by the test statistic F_{ij}, moreover, if $x_i \overset{\text{GC}}{\rightarrow} x_j$ and $x_j \overset{\text{GC}}{\rightarrow} x_k$, it is expected that $F_{kj} > F_{ki}$, thereby providing the same effect as proceeding backwards through P_{k-r}.

3.2 Model Order Selection

There are a variety of methods to choose the filter order p (see e.g. [34]), but we will focus in particular on the Bayesian Information Criteria (BIC). The BIC is substantially more conservative than the popular alternative Akaiake Information Criteria (the BIC is also asymptotically consistent), and since we are searching for *sparse graphs*, we therefore prefer the BIC, where we seek to *minimize* over p:

$$\begin{aligned} BIC_{\text{univariate}}(p) &= \mathsf{ln}\ \ \xi_i(p) + p\frac{\mathsf{ln}\ T}{T}, \\ BIC_{\text{bivariate}}(p) &= \mathsf{ln}\ \ \det\widehat{\Sigma}_{ij}(p) + 4p\frac{\mathsf{ln}\ T}{T}, \end{aligned} \tag{28}$$

where $\widehat{\Sigma}_{ij}(p)$ is the 2×2 residual covariance matrix for the $\mathsf{VAR}(p)$ model of $(x_i(t), x_j(t))$. The bivariate errors $\xi_{ij}(p)$ and $\xi_{ji}(p)$ are the diagonal entries of $\widehat{\Sigma}_{ij}(p)$.

We carry this out by a simple direct search on each model order between 0 and some prescribed $p_{\max}$, resulting in a collection p_{ij} of model order estimates. In practice, it is sufficient to pick $p_{\max}$ ad-hoc or via some simple heuristic e.g. plotting the sequence $BIC(p)$ over p, though it is not technically possible to guarantee that the optimal p is less than the chosen $p_{\max}$ (since there can in general be arbitrarily long lags from one variable to another).

[5] This quantity is often denoted $\widehat{\sigma}$, but we maintain notation from Definition 2.

3.3 Efficient Model Estimation

In practice, the vast majority of computational effort involved in implementing our estimation algorithm is spent calculating the error estimates $\xi_i(p_i)$ and $\xi_{ij}(p_{ij})$. This requires fitting a total of $n^2 p_{\max}$ autoregressive models, where the most naive algorithm (e.g. solving a least squares problem for each model) for this task will consume $O(n^2 p_{\max}^4 T)$ time, it is possible to carry out this task in a much more modest $O(n^2 p_{\max}^2) + O(n^2 p_{\max} T)$ time via the autocorrelation method [35] which substitutes the following autocovariance estimates in the Yule-Walker equations:[6]

$$\widehat{R}_x(\tau) = \frac{1}{T} \sum_{t=\tau+1}^{T} x(t)x(t-\tau)^{\mathsf{T}};\ \tau = 0, \ldots, p_{\max}, \tag{29}$$

It is imperative that the first index in the summation is $\tau + 1$, as opposed perhaps to $p_{\max}$ and that the normalization is $1/T$, as opposed perhaps to $1/(T - p_{\max})$, in order to guarantee that $\widehat{R}_x(\tau)$ forms a valid (i.e. positive definite) covariance sequence. This results in some bias, however the dramatic computational speedup is worth it for our purposes.

These covariance estimates constitute the $O(n^2 p_{\max} T)$ operation. Given these particular estimates, the variances $\xi_i(p)$ for $p = 1, \ldots, p_{\max}$ can be evaluated in $O(p_{\max}^2)$ time each by applying the Levinson-Durbin recursion to $\widehat{R}_{ii}(\tau)$, which effectively estimates a sequence of AR models, producing $\xi_i(p)$ as a side-effect (see [35] and [36]).

Similarly, the variance estimates $\widehat{\Sigma}_{ij}(p)$ (which include ξ_{ij} and ξ_{ji}) can be obtained by estimating $\frac{(n+1)n}{2}$ bivariate AR models, again in $O(p_{\max}^2)$ time via Whittle's generalized Levinson-Durbin recursion[7] [37].

3.4 Edge Probabilities and Error Rate Controls

Denote F_{ij} the Granger-causality statistic of equation 26 with model orders chosen by the methods of Section 3.2. We assume that this statistic is asymptotically $\chi^2(p_{ij})$ distributed (the disturbances are Gaussian), and denote by G the cumulative distribution function thereof. We will define the matrix

$$P_{ij} = G(F_{ij}), \tag{30}$$

to be the matrix of pairwise edge inclusion P-values. This is motivated by the hypothesis test where the hypothesis H_0 will be rejected (and thence we will conclude that $x_j \overset{\text{PW}}{\rightarrow} x_i$) if $P_{ij} > 1 - \delta$.

The value δ can be chosen by a variety of methods, in our case we apply the Benjamini Hochberg criteria [38] [32] to control the false discovery rate of pairwise edges to a level α (where we generally take $\alpha = 0.05$).

[6] The particular indexing and normalization given in equation 29 is critical to ensure $\widehat{R}$ is positive semidefinite. The estimate can be viewed as calculating the covariance sequence of a signal multiplied by a rectangular window.

[7] We have made use of standalone tailor made implementations of these algorithms, available at github.com/RJTK/Levinson-Durbin-Recursion.

3.5 Finite Sample Recovery Algorithm

After the graph topology $\widehat{\mathcal{G}}$ has been estimated via Algorithm 2, we refit the entire model with the specified sparsity pattern directly via ordinary least squares.

We note that producing graph estimates which are not strongly causal can potentially be achieved by performing sequential estimates $\widehat{x}_1(t), \widehat{x}_2(t), \ldots$ estimating a strongly causal graph with the residuals of the previous model as input, and then refitting on the combined sparsity pattern. We experiment with this heuristic in our example application of Section 5, but reserve theoretical analysis for future work.

4 Empirical Evaluation

We have implemented our empirical experiments in Python [39], in particular we leverage the LASSO implementation from `sklearn` [40] and the random graph generators from `networkx` [41]. We run experiments using two separate graph topologies having n nodes: a strongly causal graph (SCG) and a directed acyclic graph (DAG). These are generated respectively by drawing a random tree and a random Erdos Renyi graph (with edge probability $q = \frac{2}{n}$ resulting in approximately the same number of edges for the SCG as for the DAG), then creating a directed graph by directing edges from lower numbered nodes to higher numbered nodes.

We populate each of the edges (including self loops) with random linear filters constructed by placing 5 transfer function poles (i.e. $p = 5$) uniformly at random in a disc of radius 3/4 (which guarantees stability for acyclic graphs). The resulting system is driven by i.i.d. Gaussian random noise, each component having random variance $\sigma_i^2 = 1/2 + r_i$ where $r_i \sim \exp(1/2)$. We set $p_{\max} = 15$. Results and representative graphs are collected in Figures 4, 5, 6, 7, 8.

We compare our results against the adaptive LASSO [25], which outperformed substantially both the LASSO and the grouped LASSO. Motivated by scaling, we split the squared error term into separate terms, one for each group of incident edges on a node, and estimate the collection of n incident filters $\big\{\mathsf{B}_{ij}(z)\big\}_{j=1}^{n}$ that minimizes ξ_i^{LASSO} in the following:

$$\begin{aligned}\xi_i^{\text{LASSO}}(\lambda) &= \min_{B} \frac{1}{T} \sum_{t=p+1}^{T} \big(x_i(t) - \sum_{\tau=1}^{p}\sum_{j=1}^{n} B_{i,j}(\tau) x(t-\tau)\big)^2 + \lambda \sum_{\tau=1}^{p}\sum_{j=1}^{n} |B_{ij}(\tau)| \\ \xi_i^{\text{LASSO}} &= \min_{\lambda \geq 0} \xi_i^{\text{LASSO}}(\lambda) + \mathsf{BIC}\big(B_i^{\text{LASSO}}(\lambda)\big)\end{aligned} \tag{31}$$

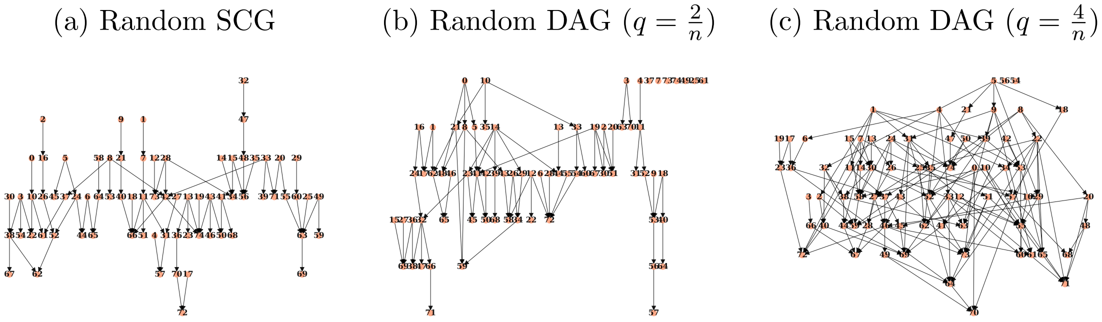
where we are choosing λ, the regularization parameters, via the BIC.

Remark 5 (Graph Topologies)**.** We depict in Figure 4 the topologies of random graphs used in our empirical evaluation. For values of q close to $\frac{2}{n}$, the resulting random graphs tend to have a topology which is, at least qualitatively, close to the SCG. As the value of q increases, the random graphs deviate farther from the SCG topology, and we therefore expect the LASSO to outperform PWGC for larger values of q. This can be observed in Figure 8, at least where performance is measured by the support recovery (i.e. via the MCC).

Algorithm 2: Finite Sample Pairwise Graph Recovery (PWGC)

```
input     : Estimates of pairwise Granger-causality statistics $F_{ij}$ (eqn 26). Matrix of
            edge probabilities $P_{ij}$ (eqn 30). Hypothesis testing threshold $\delta$ chosen via
            the Benjamini-Hochberg criterion (Section 3.4)
output    : A strongly causal graph $\widehat{\mathcal{G}}$
initialize: $S = [n]$ # unprocessed nodes
            $E = \emptyset$ # edges of $\widehat{\mathcal{G}}$
            $k = 1$ # a counter used only for notation
$W_\delta \leftarrow \{(i, j) \mid P_{ji} > 1 - \delta, F_{ji} > F_{ij}\}$ # candidate edges
$\mathcal{I}_0 \leftarrow \big(\sum_{j \in S : (j,i) \in W_\delta} P_{ij},\ \text{for } i \in S\big)$ # total node incident probability
$P_0 \leftarrow \{i \in S \mid \mathcal{I}_0(i) < \lceil \min(\mathcal{I}_0) \rceil\}$ # Nodes with fewest incident edges
if $P_0 = \emptyset$ then
    $P_0 \leftarrow \{i \in S \mid \mathcal{I}_0(i) \leq \lceil \min(\mathcal{I}_0) \rceil\}$ # Ensure non-empty

while $S \neq \emptyset$ do
    $S \leftarrow S \setminus P_{k-1}$ # remove processed nodes
    $\mathcal{I}_k \leftarrow \big(\sum_{j \in S : (j,i) \in W_\delta} P_{ij},\ \text{for } i \in S\big)$
    $P_k \leftarrow \{i \in S \mid \mathcal{I}_k(i) < \lceil \min(\mathcal{I}_k) \rceil\}$
    if $P_k = \emptyset$ then
        $P_k \leftarrow \{i \in S \mid \mathcal{I}_k(i) \leq \lceil \min(\mathcal{I}_k) \rceil\}$

    # add strongest edges, maintaining strong causality
    $U_k \leftarrow \bigcup_{r=1}^{k} P_{k-r}$ # Include all forward edges
    for $(i, j) \in \text{sort}\Big(\{(i, j) \in U_k \times P_k \mid (i, j) \in W_\delta\} \text{ by descending } F_{ji}\Big)$ do
        if is_strongly_causal($E \cup \{(i, j)\}$) then
            # is_strongly_causal can be implemented by keeping
            # track of ancestor / descendant relationships
            $E \leftarrow E \cup \{(i, j)\}$
    $k \leftarrow k + 1$
return $([n], E)$
```

Figure 4: Representative Random Graph Topologies



(a) Random SCG (b) Random DAG ($q = \frac{2}{n}$) (c) Random DAG ($q = \frac{4}{n}$)

Remark 6 (MCC as a Support Recovery Measurement)**.** We apply Matthew's Correlation Coefficient (MCC) [42] as a statistic for measuring support recovery performance. This statistic synthesizes the confusion matrix into a single score appropriate for unbalanced labels and is calibrated to fall into the range $[-1, 1]$ with 1 being perfect performance, 0 being the performance of random guessing, and -1 being perfectly opposed.

Remark 7 (Error Measurement)**.** We estimate the 1-step ahead prediction error by forming the variance matrix estimate

$$\widehat{\Sigma}_v \triangleq \frac{1}{T_{\text{out}}} \sum_{t=1}^{T_{\text{out}}} (x(t) - \widehat{x}(t))(x(t) - \widehat{x}(t))^{\mathsf{T}}$$

on a long stream of out-of-sample data. We then report the quantity

$$\frac{\ln \operatorname{tr} \widehat{\Sigma}_v}{\ln \operatorname{tr} \Sigma_v}$$

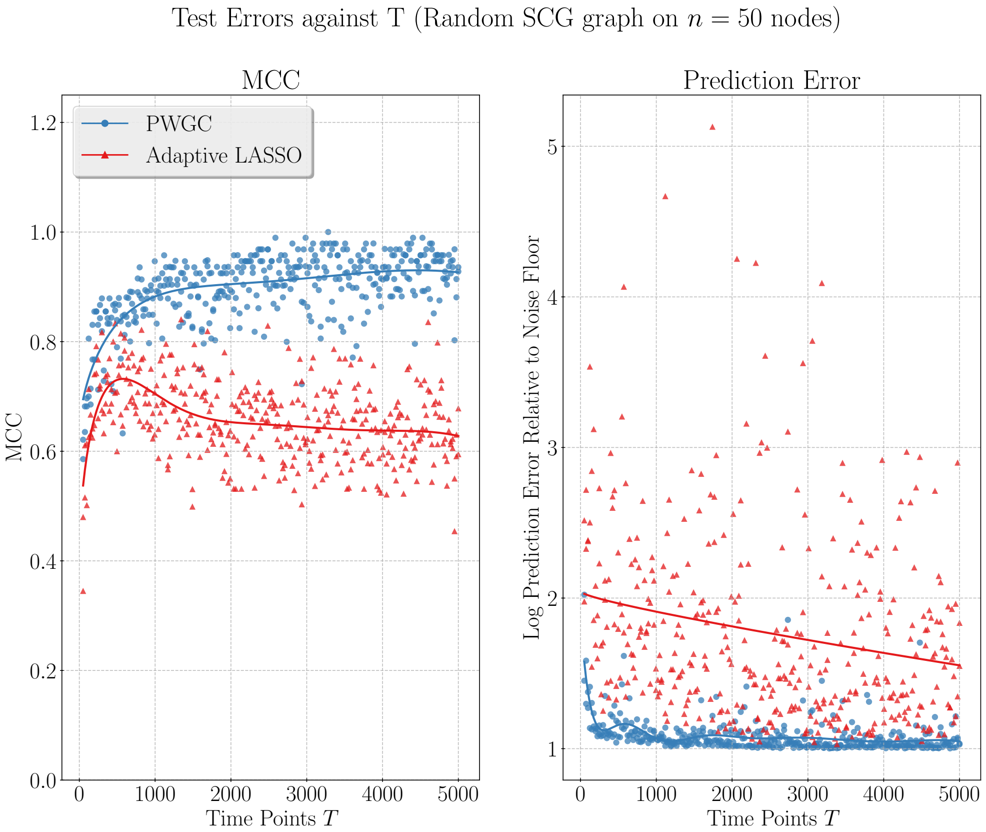
where $\widehat{\Sigma}_v = \Sigma_v$ is the best possible performance.

In reference to figure 5 it should not be overly surprising that our PWGC algorithm performs better than the LASSO for the case of a strongly causal graph, since in this case the assumptions which guarantee the correctness of Algorithm 2 hold. However, the performance is still markedly superior in the case of a more general DAG. We would conjecture that a DAG having a similar degree of sparsity as an SCG is likely to be "close" to an SCG. Figure 8 illustrates the severe (expected) degradation in performance as the number of edges increases while the number of data samples T remains fixed. For larger values q in this plot, the number of edges in the graph is comparable to the number of data samples.

We have also paid close attention to the performance of PWGC in the very small sample ($T \leq 100$) regime (see Figure 7b), as this is the regime many applications must contend with.

In regards scalability, we have observed that performing the $O(n^2)$ pairwise Granger-causality calculations consumes the vast majority ($> 90\%$) of the computation time. Since this step is trivially parallelizable, our algorithm also scales well with multiple

Figure 5: PWGC Compared Against AdaLASSO [25] (SCG)



Comparison of PWGC and LASSO for $\mathsf{VAR}(p)$ model estimation. We make comparisons against both the MCC and the relative log mean-squared prediction error $\frac{\ln \operatorname{tr}\widehat{\Sigma}_v}{\ln \operatorname{tr}\Sigma_v}$. Results in Figure 5 are for systems guaranteed to satisfy the assumptions required for Theorem 2.

Figure 6: PWGC vs adaLASSO (DAG, $q = \frac{2}{n}$)

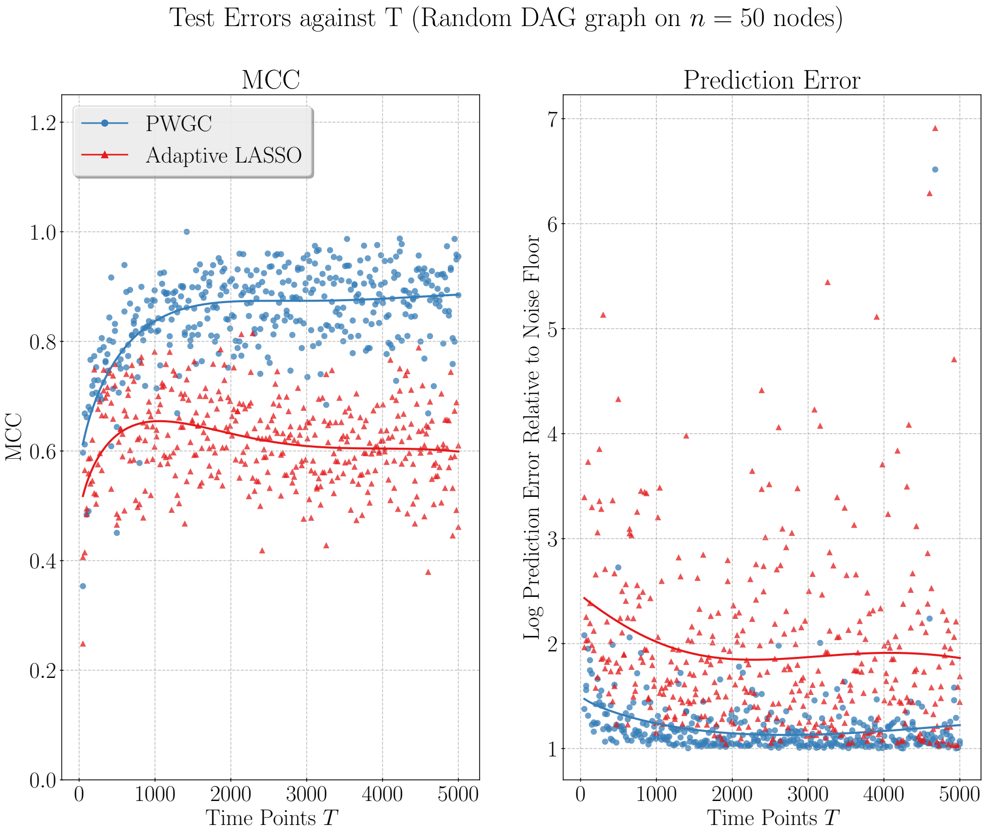
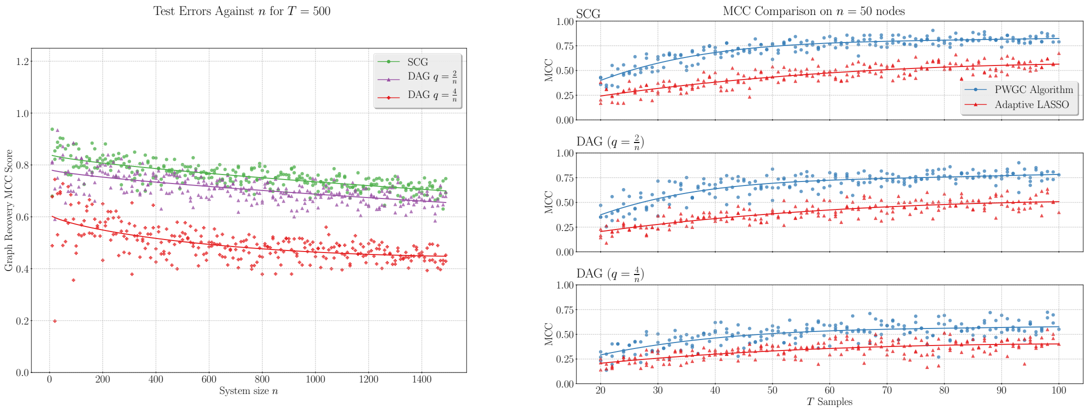


Figure 6 provides results for systems which do not guarantee the assumptions of Theorem 2, though the graph has a similar level of sparsity.

Figure 7: PWGC Scaling and Small Sample Performance



(a) Fixed T, increasing n (SCG)

(b) MCC Comparison for $T \leq 100$

Figure 7a measures support recovery performance as the number of nodes n increases, and the edge proportion as well as the number of samples T is held fixed. Remarkably, the degradation as n increases is limited, it is primarily the graph topology (SCG or non-SCG) as well as the level of sparsity (measured by q) which are the determining factors for support recovery performance.

Figure 7b provides a support recovery comparison for very small values of, T typical for many applications.

cores or multiple machines. Figure 7a is a demonstration of this scalability, where we are able to estimate graphs having over 1500 nodes (over 2.25×10^6 possible edges) using only $T = 500$ data points, granted, an SCG on this many nodes is extremely sparse.

5 Application

In this section we apply our methods to a real set of EEG data obtained from the "EEG Database Data Set"[8] [43] on the UCI machine learning repository [44]. This dataset contains 1 second long measurements of (64 channel) EEG signals from patients who are given visual stimuli. The subjects in the study are labeled as being either "control" or "alcoholic", however, we will ignore this label and instead focus on *distinguishing between subjects* based on the Granger-causality graphs inferred from the subject's trials. Our reasoning is to focus on the underlying question: "Does our PWGC algorithm uncover meaningful Granger-causal connections from EEG data?". Focusing only on the subject label allows us to answer this question without simultaneously grappeling with the physiological question of whether alcoholic subjects as a group have discernable differences in their EEG readings[9], which is a stronger requirement than simply that there are meaningful distinctions between the EEG readings of subjects generally.

The dataset consists of $\mathcal{D} = \{\big(x^{(i)}(t)\big)_{t=1}^{T}, y^{(i)}\}_{i=1}^{N}$ where $T = 256$, $x^{(i)}(t) \in \mathbb{R}^{64}$, $y^{(i)} \in [N_{subjects}]$ (with $N_{subjects} = 119$), and $N = 10723$. There are on average 90 trials

[8] http://archive.ics.uci.edu/ml/datasets/EEG+Database

[9] Some exploratory analysis actually suggests that alcoholic Granger-causality graphs are not substantively different

Figure 8: Fixed T, n, increasing edges q (DAG)

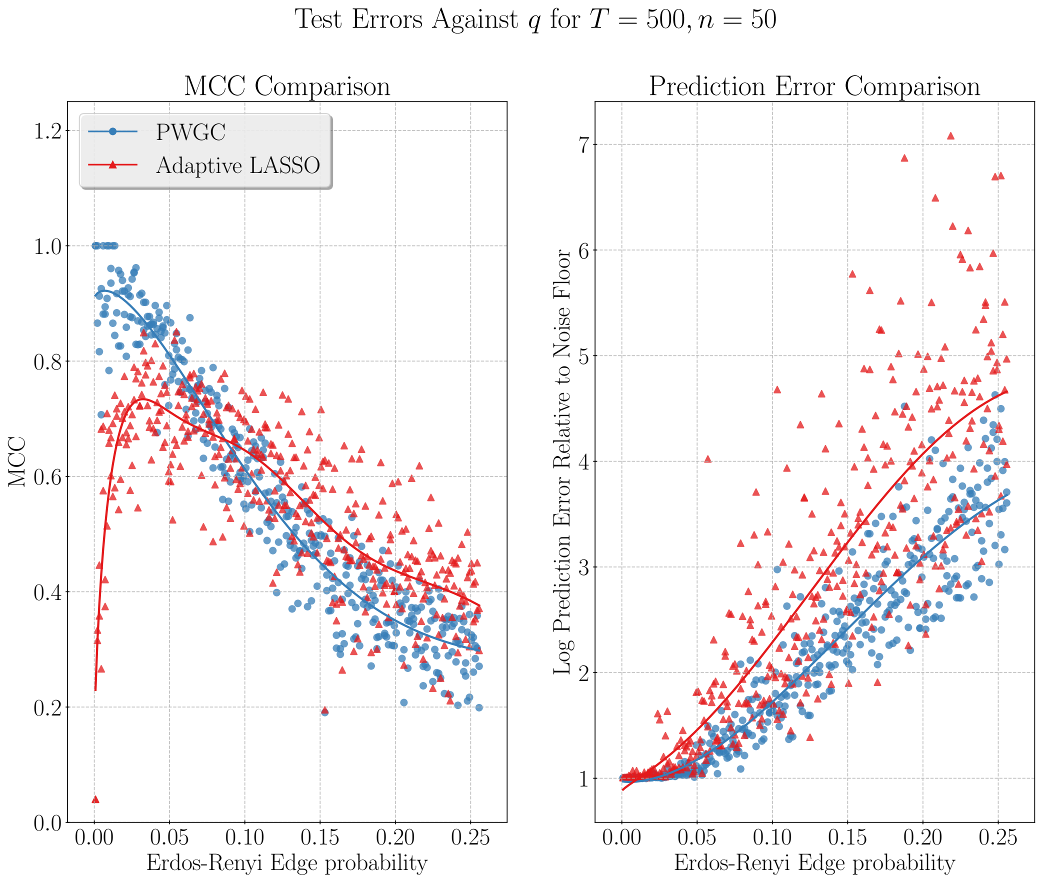


Figure 8 provides a comparison between PWGC and AdaLASSO as the density of graph edges (as measured by q) increases. For reference, $\frac{2}{n} = 0.04$ has approximately the same level of sparsity as the SCGs we simulated. As q increases, the AdaLASSO outperforms PWGC as measured by the MCC. However, PWGC maintains superior performance for 1-step-ahead prediction. We speculate that this is a result of fitting the sparsity pattern recovered by PWGC via OLS which directly seeks to optimize this metric, whereas the LASSO is encumbered by the sparsity inducing penalty.

for each subject, ranging between 30 and 119.

We have constructed a simple pipeline for discriminating between subjects by first applying an iterative PWGC algorithm (see Algorithm 3) directly to the EEG data $x^{(i)}$ to obtain a Granger-causality graph $\widehat{G}^{(i)}$. We then feed the vectorized adjacency matrix of $\widehat{G}^{(i)}$ through a polynomial Kernel multinomial logistic regression model with parameters fit by cross validation.

Remark 8. The iterated application of PWGC is an example of a heuristic by which graph estimates that aren't constrained to being strongly causal can be obtained. We defer to future work the theoretical analysis of more sophisticated heuristics, or algorithms appropriate for different topological assumptions.

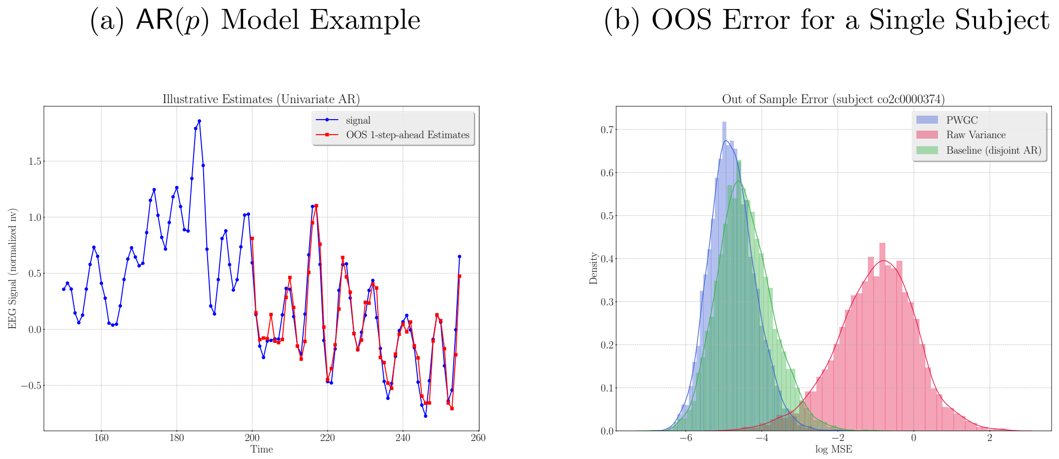
Algorithm 3: Iterated PWGC Heuristic

```
input     : Data $x(t)$, maximum number of iterations $N$, threshold $r \in (0, 1)$.
output    : Granger-causality graph estimate $\widehat{G}$ obtaind from iterated application of
            PWGC Algorithm 2
initialize: $i = 1$, $\widehat{x}_0(t) \leftarrow x(t)$, $\sigma_0 \leftarrow Var\ x(t)$
while $i \leq N$ do
    Estimate VAR model via PWGC on $x_{i-1}(t)$ to obtain $\widehat{x}_{i-1}(t)$ and $\widehat{G}_{i-1}$
    $\epsilon_i(t) \leftarrow x_{i-1}(t) - \widehat{x}_{i-1}(t)$ # Compute residuals
    $\sigma_i^2 \leftarrow \text{MSE}\big(\epsilon_i(t)\big)$ # Compute MSE
    if $\sigma_i^2 > r\sigma_{i-1}^2$ then
        break # Insignificant Improvement
    $i \leftarrow i + 1$
$\widehat{G} \leftarrow \bigcup_i \widehat{G}_i$ # Combine PWGC graph estimates
return $\widehat{G}$
```

While it is almost certainly possible to achieve much greater classification accuracy on this dataset by constructing a classifier to act directly on $x^{(i)}(t)$, our purpose is to demonstrate that relevant information is being captured by the graph estimates $\widehat{G}^{(i)}$. That is, that the latent structure $\widehat{G}^{(i)}$ may provide scientifically relevant insights, as opposed to simply being a feature for classification tasks.

Figures 9a and 9b illustrate the appropriateness of modelling $x(t)$ with VAR models, as opposed simply to a collection of n unidimensional autoregressiive models. The conclusion being that the Granger-causality graph estimates are not purely spurious.

We provide the final results in Figure 10b where the classification accuracy is quantified by the (multiclass) MCC of ≈ 0.20 on a held out validation set consisting of 20% of the data. This MCC score exceeds by a modest margin the performance of randomly guessing, lending strong evidence to the assertion that PWGC successfully recovers some meaningful differences between the Granger-causality graphs of different subjects, particularly if we recall that there are generally fewer examples for each subject than there are subjects overall. Moreover, distinguishing only between two


(a) AR(p) Model Example

(b) OOS Error for a Single Subject

Figure 9a provides an example of an autoregressive model fir via the Levinson-Durbin algorithm (choosing the system order via the BIC) on a single EEG channel. Linear autoregressive models appear qualitatively to be adequate for this data. Figure 9b demonstrates the improvement of a unified VAR model over independent AR models, providing evidence that intra-node edges inferred by PWGC are not simply spurious.

Out-of-Sample estimates are performed on data that were not used in fitting the models.

particular subjects (as opposed to distinguishing between one subject and 118 others) is substantially easier – an illustration is provided in Figure 10a where we embed the data into the plane through supervised dimensionality reduction. In general, fitting a simple classifier between two subjects achieves an MCC in excess of 0.6 and up to 0.9, even when the same hyperparameters are carried over for different pairs of subjects.

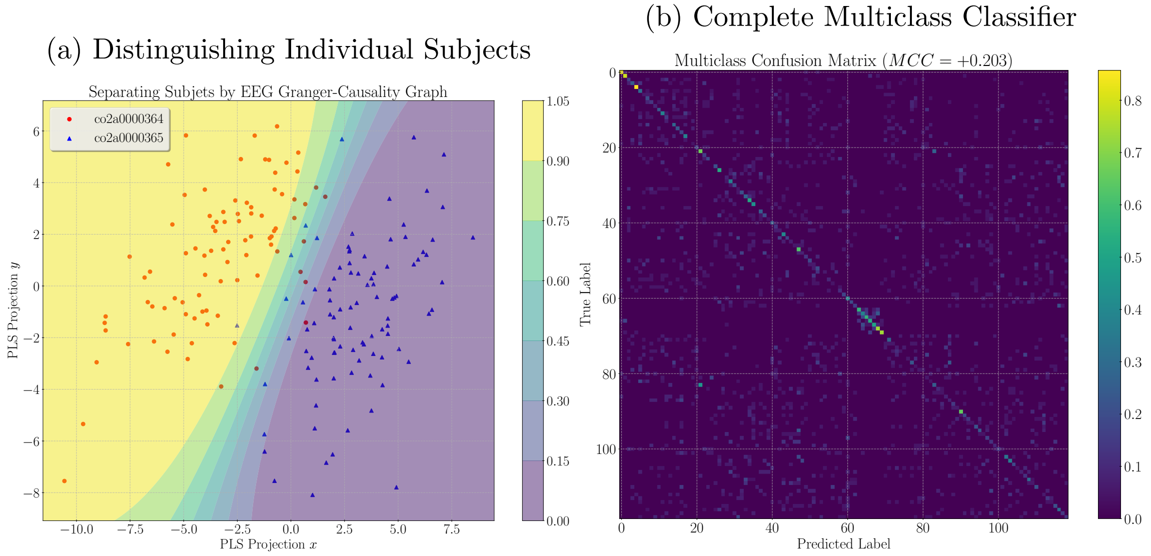
6 Conclusion

In this paper we have argued that considering particular topological properties of Granger-causality networks can provide substantial theoretical insights as well as computational benefits. In particular, the notion of a strongly-causal graph has been exploited to establish conditions under which pairwise causality testing alone is sufficient for recovering a complete Granger-causality graph. Moreover, examples from the literature suggest that such topological assumptions may be reasonable in some applications and secondly, even when the strong-causality assumption is not met, we have provided simulation evidence to suggest that our pairwise testing algorithm PWGC can still outperform the LASSO and adaLASSO, both of which are commonly employed in applications. Application to inference of more general graphs may also be within reach via an iterative graph building process as suggested in Section 5.

Our application in Section 5 provides strong evidence that PWGC is capable of uncovering meaningful features from networks of time series data by constructing a classifier which accurately discriminates between subjects in an EEG experiment based purely on the Granger-causality graph topology inferred by the iterated application of PWGC.

We emphasize that the causality graph topology is one of the key defining features of time series analysis in comparison to standard multivariate regression and therefore

Figure 10: Subject Classification from Granger-causality Graphs



(a) Distinguishing Individual Subjects

(b) Complete Multiclass Classifier

Figure 10a illustrates discriminating between two particular subjects based on their EEG Granger-causality graphs. In this case $MCC = 0.75$ on held out data. Visualization is constructed by supervised dimensionality reduction and is purely illustrative.

Figure 10b provides the row-normalized confusion matrix (computed on held out validation data) of a multiclass logistic regression classifier used to classify subjects based on their EEG Granger-causality graphs. $MCC = 0.20$

advocate for further study of how different topological assumptions may impact the recovery of causality graphs. For example, are there provable guarantees on the error rate of PWGC when applied to non strongly-causal graphs? Computationally, can constraint systems or cunning adaptive weighting schemes impose useful prior knowledge about graph topology for the LASSO algorithm? In another direction, it is known that the more general notion of transfer entropy reduces to Granger-causality in the case of Gaussian data [28], do our results extend to causality networks defined via transfer entropy? Finally, the work of [45] has established the superiority of Granger-causality testing in state space models (as opposed to pure autoregressions) in many cases. Combining this work with our PWGC algorithm (by modifying the approach described in Section 3.1 to instead utilize state-space Granger-causality testing) therefore is likely to enable application to very large networks of time series data which are not well approximated by finite $\mathsf{VAR}(p)$ models.

References

[1] C. W. J. Granger. “Investigating Causal Relations by Econometric Models and Cross-spectral Methods”. In: *Econometrica* 37.3 (1969), pp. 424–438. ISSN: 00129682, 14680262. URL: http://www.jstor.org/stable/1912791.

[2] C.W.J. Granger. “Testing for causality: A personal viewpoint”. In: *Journal of Economic Dynamics and Control* 2 (1980), pp. 329 –352. ISSN: 0165-1889. DOI: http://dx.doi.org/10.1016/0165-1889(80)90069-X. URL: http://www.sciencedirect.com/science/article/pii/016518898090069X.

[3] Steven L Bressler and Anil K Seth. “Wiener–Granger causality: a well established methodology”. In: *Neuroimage* 58.2 (2011), pp. 323–329.

[4] Anna Korzeniewska et al. “Dynamics of event-related causality in brain electrical activity”. In: *Human Brain Mapping* 29.10 (2008), pp. 1170–1192. ISSN: 1097-0193. DOI: 10.1002/hbm.20458. URL: http://dx.doi.org/10.1002/hbm.20458.

[5] Olivier David et al. “Identifying neural drivers with functional MRI: an electrophysiological validation”. In: *PLoS Biol* 6.12 (2008), e315. URL: http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0060315.

[6] Monica Billio et al. *Econometric Measures of Systemic Risk in the Finance and Insurance Sectors*. Working Paper 16223. National Bureau of Economic Research, 2010. DOI: 10.3386/w16223. URL: http://www.nber.org/papers/w16223.

[7] André Fujita et al. “Modeling gene expression regulatory networks with the sparse vector autoregressive model”. In: *BMC Systems Biology* 1.1 (2007), p. 39. ISSN: 1752-0509. DOI: 10.1186/1752-0509-1-39. URL: http://dx.doi.org/10.1186/1752-0509-1-39.

[8] Phan Nguyen. “Methods for Inferring Gene Regulatory Networks from Time Series Expression Data”. EN. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2019-05-11. PhD thesis. 2019, p. 175. ISBN: 9781392028162. URL: http://search.proquest.com.proxy.lib.uwaterloo.ca/docview/2205046162?accountid=14906.

[9] Aurélie C Lozano et al. “Grouped graphical Granger modeling for gene expression regulatory networks discovery”. In: *Bioinformatics* 25.12 (2009), pp. i110–i118.

[10] Ali Shojaie and George Michailidis. "Discovering graphical Granger causality using the truncating lasso penalty". In: *Bioinformatics* 26.18 (Sept. 2010), pp. i517–i523. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq377. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/26/18/i517/536841/btq377.pdf. URL: https://doi.org/10.1093/bioinformatics/btq377.

[11] Michail Misyrlis et al. "Sparse Causal Temporal Modeling to Inform Power System Defense". In: *Procedia Computer Science* 95 (2016). Complex Adaptive Systems Los Angeles, {CA} November 2-4, 2016, pp. 450 –456. ISSN: 1877-0509. DOI: http://dx.doi.org/10.1016/j.procs.2016.09.316. URL: //www.sciencedirect.com/science/article/pii/S1877050916324899.

[12] Tao Yuan and S Joe Qin. "Root cause diagnosis of plant-wide oscillations using Granger causality". In: *Journal of Process Control* 24.2 (2014), pp. 450–459.

[13] Dimitris Bertsimas, Angela King, Rahul Mazumder, et al. "Best subset selection via a modern optimization lens". In: *The annals of statistics* 44.2 (2016), pp. 813–852.

[14] Trevor Hastie, Robert Tibshirani, and Ryan J Tibshirani. "Extended comparisons of best subset selection, forward stepwise selection, and the lasso". In: *arXiv preprint arXiv:1707.08692* (2017).

[15] Francis R Bach and Michael I Jordan. "Learning graphical models for stationary time series". In: *IEEE transactions on signal processing* 52.8 (2004), pp. 2189–2199.

[16] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), pp. 267–288.

[17] Martin J Wainwright. "Sharp thresholds for High-Dimensional and noisy sparsity recovery using l1-Constrained Quadratic Programming (Lasso)". In: *IEEE transactions on information theory* 55.5 (2009), pp. 2183–2202.

[18] Sumanta Basu and George Michailidis. "Regularized estimation in sparse high-dimensional time series models". In: *Ann. Statist.* 43.4 (Aug. 2015), pp. 1535–1567. DOI: 10.1214/15-AOS1315. URL: https://doi.org/10.1214/15-AOS1315.

[19] Kam Chung Wong, Zifan Li, and Ambuj Tewari. "Lasso Guarantees for Time Series Estimation Under Subgaussian Tails beta-Mixing". In: *arXiv preprint arXiv:1602.04265* (2016).

[20] Y. Nardi and A. Rinaldo. “Autoregressive process modeling via the Lasso procedure”. In: *Journal of Multivariate Analysis* 102.3 (2011), pp. 528 – 549. ISSN: 0047-259X. DOI: https://doi.org/10.1016/j.jmva.2010.10.012. URL: http://www.sciencedirect.com/science/article/pii/S0047259X10002186.

[21] David Hallac et al. “Network Inference via the Time-Varying Graphical Lasso”. In: *CoRR* abs/1703.01958 (2017). arXiv: 1703.01958. URL: http://arxiv.org/abs/1703.01958.

[22] Stefan Haufe et al. “Sparse causal discovery in multivariate time series”. In: *Proceedings of the 2008th International Conference on Causality: Objectives and Assessment-Volume 6*. JMLR. org. 2008, pp. 97–106.

[23] Andrew Bolstad, Barry D Van Veen, and Robert Nowak. “Causal network inference via group sparse regularization”. In: *IEEE transactions on signal processing* 59.6 (2011), pp. 2628–2641.

[24] Yuejia He, Yiyuan She, and Dapeng Wu. “Stationary-sparse causality network learning.” In: *Journal of Machine Learning Research* 14.1 (2013), pp. 3073–3104.

[25] Hui Zou. “The adaptive lasso and its oracle properties”. In: *Journal of the American statistical association* 101.476 (2006), pp. 1418–1429.

[26] PE Caines, CW Chan, and Port Sunlight. “Feedback free processes filtering and system identification”. In: *Proc. Conf. Inform. Sci. Syst.* 1975.

[27] Anders Lindquist and Giorgio Picci. *Linear stochastic systems: A geometric approach to modeling, estimation and identification*. Vol. 1. Springer, 2015.

[28] Lionel Barnett, Adam B Barrett, and Anil K Seth. “Granger causality and transfer entropy are equivalent for Gaussian variables”. In: *Physical review letters* 103.23 (2009), p. 238701.

[29] Judea Pearl. *Causality*. Cambridge university press, 2009.

[30] Sisi Ma et al. “De-Novo Learning of Genome-Scale Regulatory Networks in S. cerevisiae”. In: *PLOS ONE* 9.9 (Sept. 2014), pp. 1–20. DOI: 10.1371/journal.pone.0106479. URL: https://doi.org/10.1371/journal.pone.0106479.

[31] Judea Pearl. “The art and science of cause and effect”. In: *Causality: models, reasoning and inference* (2000), pp. 331–358.

[32] Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

[33] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012. ISBN: 0262018020, 9780262018029.

[34] Helmut Lütkepohl. *New introduction to multiple time series analysis.* Springer Science & Business Media, 2005.

[35] Monson H Hayes. *Statistical digital signal processing and modeling.* John Wiley & Sons, 2009.

[36] James Durbin. "The fitting of time-series models". In: *Revue de l'Institut International de Statistique* (1960), pp. 233–244.

[37] P. WHITTLE. "On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix". In: *Biometrika* 50.1-2 (June 1963), pp. 129–134. ISSN: 0006-3444. DOI: 10.1093/biomet/50.1-2.129. eprint: http://oup.prod.sis.lan/biomet/article-pdf/50/1-2/129/803509/50-1-2-129.pdf. URL: https://doi.org/10.1093/biomet/50.1-2.129.

[38] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[39] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python.* 2001–. URL: http://www.scipy.org/.

[40] Fabian Pedregosa et al. "Scikit-learn: Machine learning in Python". In: *Journal of machine learning research* 12.Oct (2011), pp. 2825–2830.

[41] Aric Hagberg, Pieter Swart, and Daniel S Chult. *Exploring network structure, dynamics, and function using NetworkX.* Tech. rep. Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.

[42] Brian W Matthews. "Comparison of the predicted and observed secondary structure of T4 phage lysozyme". In: *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405.2 (1975), pp. 442–451.

[43] X L Zhang et al. "Event related potentials during object recognition tasks". In: *Brain Research Bulletin* 38 (1995), pp. 531–538.

[44] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository.* 2017. URL: http://archive.ics.uci.edu/ml.

[45] Lionel Barnett and Anil K Seth. "Granger causality for state-space models". In: *Physical Review E* 91.4 (2015), p. 040101.