Network Science **11** (3): 431–457, 2023
doi:10.1017/nws.2023.11

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

Exact recovery of Granger causality graphs with unconditional pairwise tests

R. J. Kinnear and R. R. Mazumdar

Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada
Corresponding author: R. R. Mazumdar; Email: mazum@uwaterloo.ca

Action Editor: Paolo Pin

Abstract

We study Granger Causality in the context of wide-sense stationary time series. The focus of the analysis is to understand how the underlying topological structure of the causality graph affects graph recovery by means of the pairwise testing heuristic. Our main theoretical result establishes a sufficient condition (in particular, the graph must satisfy a polytree assumption we refer to as *strong causality*) under which the graph can be recovered by means of unconditional and *binary* pairwise causality testing. Examples from the gene regulatory network literature are provided which establish that graphs which are strongly causal, or very nearly so, can be expected to arise in practice. We implement finite sample heuristics derived from our theory, and use simulation to compare our pairwise testing heuristic against LASSO-based methods. These simulations show that, for graphs which are strongly causal (or small perturbations thereof) the pairwise testing heuristic is able to more accurately recover the underlying graph. We show that the algorithm is scalable to graphs with thousands of nodes, and that, as long as structural assumptions are met, exhibits similar high-dimensional scaling properties as the LASSO. That is, performance degrades slowly while the system size increases and the number of available samples is held fixed. Finally, a proof-of-concept application example shows, by attempting to classify alcoholic individuals using only Granger causality graphs inferred from EEG measurements, that the inferred Granger causality graph topology carries identifiable features.

Keywords: Granger-causality; LASSO regression; sparsity; autoregressive models; time series analysis; electroencephalogram

1. Introduction

We study the use of Granger causality (Granger, 1969, 1980; Caines, 1976) as a means of uncovering an underlying causal structure in multivariate time series. There are two basic ideas behind Granger's notion of causality. First, that causes should precede their effects. Second, that the cause of an event must be external to the event itself. Granger's methodology uses these ideas to test a *necessary* condition of causation: whether forecasting a process x_c (the purported *caused* process) using another process x_e (the external, or purported *causing* process), in addition to its own history, is more accurate than using the past data of x_c alone. If x_e is in fact a cause of x_c then it must be that the combined estimator is more accurate. If our data of interest can be appropriately modeled as a multivariate process in which Granger's test for causation is also sufficient, then Granger causality tests provide a means of defining a network of causal connections. The nodes of this network are individual stochastic processes, and the directed edges between nodes indicate Granger causal connections. This causality graph is the principal object of our study.

In practice, the underlying causality graph cannot be observed directly, and its presence is inferred as a latent structure among observed time series data. This notion has been used in a variety of applications, for example, in Neuroscience as a means of recovering interactions amongst brain regions (Bressler & Seth, 2011; Korzeniewska et al., 2008; David et al., 2008; Seth et al., 2015); in the study of the dependence and connectedness of financial institutions (Billio et al., 2010); gene expression networks (Fujita et al., 2007; Nguyen, 2019; Lozano et al., 2009; Shojaie & Michailidis, 2010; Ma et al., 2014); and power system design (Michail et al., 2016; Yuan & Qin, 2014). For additional context in applications (Michailidis & d'Alché-Buc, 2013) reviews some methods (not only Granger causality) for the general problem of gene expression network inference, and He et al. (2019) for the case of EEG brain connectivity.

Granger's original discussions (Granger, 1969, 1980) did not fundamentally assume the use of linear time series models, but their use in practice is so widespread that the idea of Granger causality has become nearly synonymous with sparse estimation of vector autoregressive (VAR) models. The problem has also been formulated for the more general linear state space setting as well (Solo, 2015; Barnett & Seth, 2015). This focus is in part due to the tractability of linear models (Granger's original argument in their favor) but clear generalizations beyond the linear case in fact exhibit some unexpected pathologies (James et al., 2016), and it is challenging to apply the notion to rigorously *define* a causal network. Thus, it is henceforth to be understood that we are referring to Granger causation in the context of linear time series models.

1.1 Background and overview

Our primary goal is to estimate the Granger causal relations amongst processes in a multivariate time series, that is, to recover the Granger Causality Graph (GCG). One formulation of the problem of GCG recovery, for finite data samples, is to search for the best (according to a chosen metric) graph structure or sparsity pattern consistent with observed data. This becomes a difficult discrete optimization problem (best subset selection) and is only tractable for relatively small systems (Hastie et al., 2017), particularly because the number of possible edges scales quadratically in the number of nodes.

An early heuristic approach to the problem in the context of large systems is provided by Bach & Jordan (2004), where the authors apply a local search heuristic to the Whittle likelihood with an AIC (Akaike Information Criterion) penalization. The local search heuristic is a common approach to combinatorial optimization due to its simplicity, but is liable to get stuck in shallow local minima.

Another successful and widely applied method is the LASSO regularizer (Tibshirani, 1996; Tibshirani et al., 2015), which can be understood as a natural convex relaxation to penalizing the count of non-zero edges in a graph recovery optimization problem. There are theoretical guarantees regarding the LASSO performance in linear regression (Wainwright, 2009), which remain at least partially applicable to the case of stationary time series data, as long as there is a sufficiently fast rate of dependence decay (Basu & Michailidis, 2015; Wong et al., 2016; Nardi & Rinaldo, 2011). There is by now a vast literature on sparse machine learning and numerous variations on the basic LASSO regularization procedure have been studied including grouping (Yuan & Lin, 2006), adaptation (Zou, 2006), covariance selection (graph lasso) (Friedman et al., 2008), post-selection methods (Lee et al., 2016), bootstrap-like procedures (stability selection) (Bach, 2008; Meinshausen & Bühlmann, 2010), and Bayesian methods (Ahelegbey et al., 2016).

This variety of LASSO adaptations has found natural application in Granger causality analysis. For a sample of this literature consider: Haufe et al.(2008) and Bolstad et al. (2011) which applies the group LASSO for edge-wise joint sparsity of VAR model coefficients; Hallac et al. (2017) applies grouping to a non-stationary dataset; He et al. (2013) applies Bootstrapping, a modified LASSO penalty, and a stationarity constraint to fundamental economic data; Lozano et al. (2009) considers both the adaptive and grouped variations for gene expression networks; Haury et al.

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

(2012) applies stability selection, also to gene expression networks; Shojaie & Michailidis (2010) develops another tailored LASSO-like penalty to the problem of model order selection in p order VAR models; and Basu et al. (2019) includes a penalty which also encourages low-rank solutions (via the nuclear norm).

In the context of time series data, sparsity assumptions remain important but there is significant additional structure induced by the *directionality* of causation. This directionality induces graphical structure, that is, the *graph topological aspects*, are specific to the time series case. Directly combining topological assumptions with the concept of Granger causality is a radically different approach than the use of LASSO regularization. Indeed, the LASSO is particularly designed for *static* linear regression, where there is no immediate notion of directed causation. Taking directed graphical structure into account is therefore a major avenue of possible exploration. A famous early consideration of network topologies (though not directly in the context of time series) is due to Chow & Liu (1968) where a tree structured graph is used to approximate a joint probability distribution. The work is further generalized to polytree structured graphs in Rebane & Pearl (1987). Recent work in the case of Granger causality includes (Józsa, 2019) with a detailed analysis of Granger causality in star-structured graphs, and Datta-Gupta & Mazumdar (2013) for acyclic graphs.

Part of the impetus of this approach follows since it is now known that many real world networks indeed exhibit special topological structure. For instance, "small-world" graphs (Newman & Watts, 1999; Watts & Strogatz, 1998) are pervasive—they are characterized by being highly clustered yet having short connections between any two nodes. Indeed, analysis of network topology is now an important field of neuroscience (Bassett & Sporns, 2017), where Granger causality has long been frequently applied.

1.2 Contributions and organization

The principal contributions of this paper are as follows: First, in Section 2 we study *pairwise* Granger causality relations, providing theorems connecting the structure of the causality graph to the "pairwise causal flow" in the system, that is, if there is a path from x_i to x_j, when can that connection be detected by observing only those two variables? These results can be seen as a continuation of the related work of Datta-Gupta & Mazumdar (2013), from where we have drawn inspiration for our approach. The contributions made here are firstly technical: we are careful to account for some pathological cases not recognized in this earlier work. Moreover, we are able to derive results which are able to *fully* recover the underlying graph structure (sec 2.7) with pairwise testing alone, which was not possible in Datta-Gupta & Mazumdar (2013).

This work also builds upon Materassi & Innocenti (2010) and Tam et al. (2013) that have pioneered the use of pairwise modeling as a heuristic. The work of Innocenti & Materassi (2008) and Materassi & Innocenti (2010) studies similar questions and provides similar results to ours, namely, they prove elegant results on graph recovery using minimum spanning trees (where distance is a measure of coherence between nodes) which recovers the graph whenever it is, in fact, a tree (or more generally, a polytree or forest). By contrast, our theoretical results show that complete graph recovery of polytrees (which we have referred to as strongly causal graphs) is possible with only *binary* measurements of directed Granger causality. That is, while the final conclusion is similar (a causality graph is recovered), we are able to do so with *weaker* pairwise tests.

We present simulation results in Section 3. These results establish that considering topological information provides significant potential for improvement over existing methods. In particular, a simple implementation of our pairwise testing and graph recovery algorithm outperforms the adaptive LASSO (adaLASSO) (Zou, 2006) in relevant regimes and exhibits superior computational scaling. Indeed, one of our simulations was carried out on graphs with up to 1500 nodes and a VAR lag length of 5 (this system has over 2 million possible directed edges, and over 10 million variables) and shows statistical scaling behavior similar to that of the LASSO itself where

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

performance degrades slowly as the number of variables increases but the amount of data remains fixed. Moreover, while our algorithm is asymptotically exact only when the underlying graph has a special "strongly causal" structure, simulation shows that performance can still exceed that of the adaLASSO on sparse acyclic graphs that do not have this property, even though the adaLASSO is asymptotically exact on *any* graph.

Following our simulation results, a small proof-of-concept application is described in Section 4 where Granger causality graphs are inferred from the EEG readings of multiple subjects, and these causality graphs are used to train a classifier that indicates whether or not the subject has alcoholism.

2. Graph topological aspects

2.1 Formal setting

Consider the usual Hilbert space, $L_2(\Omega)$, of finite variance random variables over a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ 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 $x(n)$ (with $n \in \mathbb{Z}$) where the N elements take values in L_2. We suppose that $x(n)$ has zero mean, $\mathbb{E}x(n) = 0$, and has absolutely summable matrix valued covariance sequence $R(m) \stackrel{\Delta}{=} \mathbb{E}x(n)x(n-m)^{\mathsf{T}}$, and hence an absolutely continuous spectral density.

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

$$\begin{aligned} \mathscr{H}_n^{(x)} &= \mathsf{cl}\left\{\sum_{m=0}^{p} A(m)^{\mathsf{T}} x(n-m),\ A(m) \in \mathbb{R}^N, p \in \mathbb{N}\right\} \subseteq L_2(\Omega) \\ H_n^{(x)} &= \{ax(n) \mid a \in \mathbb{R}\} \subseteq L_2(\Omega) \end{aligned} \tag{1}$$

where the closure is in mean-square. We will often omit the superscript (x), which should be clear from context. 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, for example, $\mathscr{H}_n^{(i,j)}$, $\mathscr{H}_n^{(i)}$ or by all but a particular component as $\mathscr{H}_n^{(-j)}$.

From the Wold decomposition theorem (Lindquist & Picci, 2015), every WSS sequence has the moving average $\mathsf{MA}(\infty)$ representation

$$x(n) = c(n) + \sum_{m=0}^{\infty} A(m)v(n-m) \tag{2}$$

where $c(n)$ is a purely deterministic sequence, $v(n)$ is an uncorrelated sequence and $A(0) = I$. We will assume that $c(n) = 0\ \forall n$, which makes $x(n)$ a *regular non-deterministic sequence*. This representation can be inverted (Lindquist & Picci, 2015) to yield the *Vector Autoregressive* $\mathsf{VAR}(\infty)$ form

$$x(n) = \sum_{m=1}^{\infty} B(m)x(n-m) + v(n) \tag{3}$$

The Equations (2) and (3) can be represented as $x(n) = \mathsf{A}(z)v(n) = \mathsf{B}(z)x(n) + v(n)$ via the action of the operators

$$\mathsf{A}(z) := \sum_{m=0}^{\infty} A(m)z^{-m}$$

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

and

$$\mathsf{B}(z) := \sum_{m=1}^{\infty} B(m) z^{-m}$$

where the operator z^{-1} is the backshift operator acting on $\ell_2^N(\Omega)$, the space of square summable sequences of N-vectors having elements in $L_2(\Omega)$ that is:

$$\mathsf{B}_{ij}(z) x_j(n) := \sum_{m=1}^{\infty} B_{ij}(m) x_j(n-m) \tag{4}$$

The operator $\mathsf{A}(z)$ contains only non-positive powers of z and is therefore called *causal*, and if $A(0) = 0$ then it is *strictly causal*. An operator containing only non-negative powers of z, for example, $\mathsf{C}(z) = c_0 + c_1 z + c_2 z^2 + \cdots$ is *anti-causal*. These operators are often referred to as *filters* in signal processing, terminology which we often adopt as well.

Finally, we have the well known inversion formula

$$\mathsf{A}(z) = (I - \mathsf{B}(z))^{-1} = \sum_{k=0}^{\infty} \mathsf{B}(z)^k \tag{5}$$

The above assumptions are quite weak, and are essentially the minimal requirements needed to have a non-degenerate model. The strongest assumption we require is finally that Σ_v is a *diagonal* and positive-definite matrix. Equivalently, the Wold decomposition could be written with $A(0) \neq I$ and $\Sigma_v = I$, and we would be assuming that $A(0)$ is diagonal. This precludes the possibility of *instantaneous causality* (Granger, 1969) or of processes being only weakly (as opposed to strongly) feedback free, in the terminology of Caines (1976). We formally state our setup as a definition:

Definition 2.1 (Basic Setup). *The process $x(n)$ is an N dimensional, zero mean, wide-sense stationary process having invertible $\mathsf{VAR}(\infty)$ representation (3), where $v(n)$ is sequentially uncorrelated and has a diagonal positive-definite covariance matrix. The $\mathsf{MA}(\infty)$ representation of Equation (2) has $c(n) = 0\ \forall n$ and $A(0) = I$.*

2.2 Granger causality

Definition 2.2 (Granger Causality). *We will say that component x_j Granger-causes component x_i (with respect to x) and write $x_j \overset{\text{GC}}{\to} x_i$ if*

$$\xi[x_i(n) \mid \mathscr{H}_{n-1}] < \xi[x_i(n) \mid \mathscr{H}_{n-1}^{(-j)}] \tag{6}$$

where $\xi[x \mid \mathscr{H}] := \mathbb{E}(x - \hat{\mathbb{E}}[x \mid \mathscr{H}])^2$ is the linear mean-squared estimation error and

$$\hat{\mathbb{E}}[x \mid \mathscr{H}] := \text{proj}_{\mathscr{H}}(x)$$

denotes the (unique) projection onto the Hilbert space $\mathscr{H}$.

This notion captures the idea that the process x_j provides information about x_i that is not available from elsewhere. The notion is closely related to the information theoretic measure of transfer entropy, indeed, if the distribution of $v(n)$ is known to be Gaussian then they are equivalent (Barnett et al., 2009).

The notion of conditional orthogonality is the essence of Granger causality, and enables us to obtain results for a fairly general class of WSS processes, rather than simply $\mathsf{VAR}(p)$ models.

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

Definition 2.3 (Conditional Orthogonality). *Consider three closed subspaces of a Hilbert space $\mathscr{A}$, $\mathscr{B}$, $\mathscr{X}$. We say that $\mathscr{A}$ is conditionally orthogonal to $\mathscr{B}$ given $\mathscr{X}$ and write $\mathscr{A} \perp \mathscr{B} \mid \mathscr{X}$ if*

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

An equivalent condition is that (see Lindquist & Picci (2015) Proposition 2.4.2)

$$\hat{\mathbb{E}}[\beta \mid \mathscr{A} \vee \mathscr{X}] = \hat{\mathbb{E}}[\beta \mid \mathscr{X}] \; \forall \beta \in \mathscr{B}$$

The following theorem captures a number of well known equivalent definitions or characterizations of Granger causality:

Theorem 2.1 (Granger Causality Equivalences). *The following are equivalent:*

(a) $x_j \overset{\text{GC}}{\nrightarrow} x_i$

(b) $\forall m \in \mathbb{N}_+ \; B_{ij}(m) = 0$, that is, $\mathsf{B}_{ij}(z) = 0$

(c) $H_n^{(i)} \perp \mathscr{H}_{n-1}^{(j)} \mid \mathscr{H}_{n-1}^{(-j)}$

(d) $\hat{\mathbb{E}}[x_i(n) \mid \mathscr{H}_{n-1}^{(-j)}] = \hat{\mathbb{E}}[x_i(n) \mid \mathscr{H}_{n-1}]$

Proof. These facts are quite well known. $(a) \Rightarrow (b)$ follows as a result of the uniqueness of orthogonal projection (i.e.,, the best estimate is necessarily the coefficients of the model). $(b) \Rightarrow (c)$ follows since in computing $(y - \hat{\mathbb{E}}[y \mid \mathscr{H}_{n-1}^{(-j)}])$ for $y \in H_n^{(i)}$ it is sufficient to consider $y = x_i(n)$ by linearity, then since $H_{n-1}^{(i)} \subseteq \mathscr{H}_{n-1}^{(-j)}$ we have $(x_i(n) - \hat{\mathbb{E}}[x_i(n) \mid \mathscr{H}_{t-1}^{(-j)}]) = v_i(n)$ since $\mathsf{B}_{ij}(z) = 0$. The result follows since $v_i(n) \perp \mathscr{H}_{n-1}$. $(c) \iff (d)$ is a result of the equivalence in Definition 2.3. And, $(d) \implies (a)$ follows directly from the Definition. □

2.3 Granger causality graphs

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

Definition 2.4 (Graph Theory Review). *A graph $\mathscr{G} = (V, \mathscr{E})$ is a tuple of sets respectively called nodes and edges. Throughout this paper, we have in all cases $V = [N] := \{1, 2, \ldots, N\}$. We will also focus solely on directed graphs, where the edges $\mathscr{E} \subseteq V \times V$ are ordered pairs.*

A (directed) path (of length r) from node i to node j, denoted $i \to \cdots \to j$, is a sequence $a_0, a_1, \ldots, a_{r-1}, a_r$ with $a_0 = i$ and $a_r = j$ such that $\forall \, 0 \leq k \leq r \; (a_k, a_{k+1}) \in \mathscr{E}$, 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 $(i, i) \in \mathscr{E}$ (which we do not consider to be a cycle) is referred to as a loop.

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

Definition 2.5 (Parents, Grandparents, Ancestors). *A node j is a parent of node i if $(j, i) \in \mathscr{E}$. 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 $(i, i) \in \mathscr{E}$.*

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 $\mathscr{G}$ from j to i. Clearly, $pa(i) = gp_1(i)$.

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

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

Recall that we do not allow a node to be its own parent, but unless $\mathscr{G}$ is a DAG, a node can be its own ancestor. We will occasionally need to explicitly exclude i from $\mathscr{A}(i)$, in which case we will write $\mathscr{A}(i) \setminus \{i\}$.

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

Definition 2.6 (Causality graph). *We define the Granger causality graph $\mathscr{G} = ([N], \mathscr{E})$ to be the directed graph formed on N vertices where an edge $(j, i) \in \mathscr{E}$ if and only if x_j Granger-causes x_i (with respect to x). That is,*

$$(j, i) \in \mathscr{E} \iff j \in pa(i) \iff x_j \overset{\text{GC}}{\to} x_i$$

The edges of the Granger causality graph $\mathscr{G}$ can be given a general notion of "weight" by associating an edge (j, i) with the *strictly causal* operator $\mathsf{B}_{ij}(z)$ (see Equation (4)). Hence, the matrix $\mathsf{B}(z)$ is analogous to a *weighted adjacency matrix*[1] for the graph $\mathscr{G}$. 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 (5).

From the VAR representation of $x(n)$ there is clearly a close relationship between each node and its parent nodes, the relationship is quantified through the sparsity pattern of $\mathsf{B}(z)$. The work of Caines & Chan (1975) *starts* by defining causality (including between groups of processes) via the notion of *feedback free* processes in terms of $\mathsf{A}(z)$. The following proposition provides a similar interpretation of the sparsity pattern of $A(z)$ in terms of the causality graph $\mathscr{G}$.

Proposition 2.1 (Ancestor Expansion). *The component $x_i(n)$ of $x(n)$ can be represented in terms of its parents in $\mathscr{G}$:*

$$x_i(n) = v_i(n) + \mathsf{B}_{ii}(z)x_i(n) + \sum_{k \in pa(i)} \mathsf{B}_{ik}(z)x_k(n) \tag{7}$$

Moreover, $x_i(\,\cdot\,)$ can be expanded in terms of its ancestor's $v(n)$ components only:

$$x_i(n) = \mathsf{A}_{ii}(z)v_i(n) + \sum_{k \in \mathscr{A}(i) \setminus \{i\}} \mathsf{A}_{ik}(z)v_k(n) \tag{8}$$

where $\mathsf{A}(z) = \sum_{m=0}^{\infty} A(m)z^{-m}$ is the filter from the Wold decomposition representation of $x(n)$, Equation (2).

Proof. Equation (7) is immediate from the VAR(∞) representation of Equation (3) and Theorem 2.1, we are left to demonstrate (8).

From Equation (3), which we are assuming throughout the paper to be invertible, we can write

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

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

$$(I - \mathsf{B}(z))^{-1} = \sum_{k=0}^{\infty} \mathsf{B}(z)^k \tag{9}$$

Invoking the Cayley-Hamilton theorem allows writing the infinite sum of (9) 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 D.1 S is the transpose of the adjacency matrix for $\mathscr{G}$ and hence $(S^k)_{ij}$ is non-zero if and

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

only if $j \in gp_k(i)$, and therefore $\mathsf{B}(z)^k_{ij} = 0$ if $j \notin gp_k(i)$. Finally, since $\mathscr{A}(i) = \bigcup_{k=1}^{N} gp_k(i)$ we see that $\mathsf{A}_{ij}(z)$ is zero if $j \notin \mathscr{A}(i)$.

Therefore

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

This completes the proof. □

Proposition 2.1 is ultimately about the *sparsity pattern* in the Wold decomposition matrices $A(m)$ since $x_i(n) = \sum_{m=0}^{\infty} \sum_{j=1}^{N} A_{ij}(m) v_j(n-m)$. The proposition states that if $j \notin \mathscr{A}(i)$, then $\mathsf{A}_{ij}(z) = 0$.

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 fully conditional relations of $\mathscr{G}$.

Definition 2.7 (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 $j \in \mathscr{A}(i)$ we expect that $j \overset{\text{PW}}{\to} i$. It may also be useful for reasoning about the conditions under which *unobserved* components of $x(n)$ may or may not interfere with inference in the actually observed components. Finally, motivated from a practical standpoint to analyze causation in large systems, practical estimation procedures based purely on pairwise causality tests are of interest since the computation of such pairwise relations is substantially easier than the full conditional relations.

The following propositions are essentially lemmas used for the proof of the upcoming Proposition 2.4, but remain relevant for providing intuitive insight into the problems at hand.

Proposition 2.2 (Fully Disjointed Nodes; Proof in Section D.2). *Consider distinct nodes i, j in a Granger causality graph $\mathscr{G}$. If*

(a) $j \notin \mathscr{A}(i)$ *and* $i \notin \mathscr{A}(j)$
(b) $\mathscr{A}(i) \cap \mathscr{A}(j) = \emptyset$

then $\mathscr{H}_n^{(i)} \perp \mathscr{H}_n^{(j)}$, that is, $\forall l, m \in \mathbb{Z}_+$ $\mathbb{E}[x_i(n-l)x_j(n-m)] = 0$. Moreover, this means that $j \overset{\text{PW}}{\nrightarrow} i$ and $\hat{\mathbb{E}}[x_j(n) \mid \mathscr{H}_n^{(i)}] = 0$.

Remark 2.1. It is possible for components of $x(n)$ to be correlated at some time lags without resulting in pairwise causality. For instance, the conclusion $j \overset{\text{PW}}{\nrightarrow} i$ of Proposition 2.2 will still hold

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

even if $i \in \mathscr{A}(j)$, since j cannot provide any information about i that is not available from observing i itself.

Proposition 2.3 (Not an Ancestor, No Common Cause; Proof in Appendix D.2). *Consider distinct nodes i, j in a Granger causality graph $\mathscr{G}$. If*

(a) $j \notin \mathscr{A}(i)$
(b) $\mathscr{A}(i) \cap \mathscr{A}(j) = \emptyset$

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

The previous result can still be strengthened significantly; notice that it is possible to have some $k \in \mathscr{A}(i) \cap \mathscr{A}(j)$ where still $j \overset{\text{PW}}{\nrightarrow} i$, an example is furnished by the three node graph $k \to i \to j$ where clearly $k \in \mathscr{A}(i) \cap \mathscr{A}(j)$ 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 2.8 (Confounder). *A node k will be referred to as a confounder of nodes i, j (neither of which are equal to k) if $k \in \mathscr{A}(i) \cap \mathscr{A}(j)$ 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 $i \leftarrow k \to j$.*

A similar proposition as the following also appears as [Datta-Gupta, (2014), Prop. 5.3.2]:

Proposition 2.4 (Necessary Conditions). *If in a Granger causality graph $\mathscr{G}$ where $j \overset{\text{PW}}{\to} i$ then $j \in \mathscr{A}(i)$ or $\exists k \in \mathscr{A}(i) \cap \mathscr{A}(j)$ which is a confounder of (i, j).*

Proof. The proof is by way of contradiction. To this end, suppose that j is a node such that:

(a) $j \notin \mathscr{A}(i)$
(b) every $k \in \mathscr{A}(i) \cap \mathscr{A}(j)$ 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 $u \in \mathscr{A}(i)$ then $u \in \mathscr{A}(i) \cap \mathscr{A}(j)$ so by our assumption every $u \to \cdots \to j$ path must contain i, but $u \in pa(j)$ so $u \to j$ is a path that doesn't contain i, and therefore $u \notin \mathscr{A}(i)$; moreover, if we consider $w \in \mathscr{A}(i) \cap \mathscr{A}(u)$ then we also have $w \in \mathscr{A}(i) \cap \mathscr{A}(j)$ so the assumption implies that every $w \to \cdots \to j$ path must contain i. These properties therefore extend inductively to every $u \in \big(\mathscr{A}(j) \setminus \{i\}\big)$.

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

$$C_0(u) = \{k \in pa(u) \mid i \notin \mathscr{A}(k), \mathscr{A}(i) \cap \mathscr{A}(k) = \emptyset, k \neq i\}$$
$$C_1(u) = \{k \in pa(u) \mid i \in \mathscr{A}(k) \text{ or } k = i\}$$
$$C_2(u) = \{k \in pa(u) \mid i \notin \mathscr{A}(k), \mathscr{A}(i) \cap \mathscr{A}(k) \neq \emptyset, k \neq i\}$$

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 $\exists w \in \mathscr{A}(i) \cap \mathscr{A}(k)$ (and $w \in \mathscr{A}(i) \cap \mathscr{A}(u)$, since $k \in pa(u)$) such that $i \notin \mathscr{A}(k)$ and therefore there must be a path $w \to \cdots \to k \to u$ which does not contain i, contradicting property (b). Moreover, for any $u \in \big(\mathscr{A}(j) \setminus \{i\}\big)$ and $k \in C_0(u)$, Proposition 2.3 shows that $H_n^{(i)} \perp \mathscr{H}_{n-1}^{(j)} | \mathscr{H}_{n-1}^{(i)}$.

In order to establish $j \overset{\text{PW}}{\nrightarrow} i$, choose an arbitrary element of $\mathscr{H}_{t-1}^{(j)}$ and represent it via the action of a strictly causal filter $\Phi(z)$, that is, $\Phi(z)x_j(n) \in \mathscr{H}_{n-1}^{(j)}$, by Theorem 2.1 it suffices to show that

$$\langle x_i(n), \Phi(z)x_j(n) - \hat{\mathbb{E}}[\Phi(z)x_j(n) \mid \mathscr{H}_{t-1}^{(i)}] \rangle = 0 \tag{10}$$

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

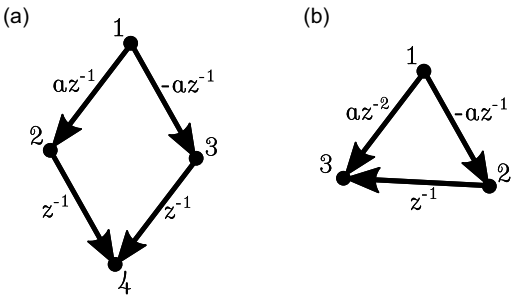

Figure 1. Examples illustrating the difficulty of obtaining a converse to Proposition 2.4. (a) Path cancelation: $j \in \mathscr{A}(i) \not\Rightarrow j \overset{\text{PW}}{\to} i$. (b) Cancellation from time lag: $j \in pa(i) \not\Rightarrow j \overset{\text{PW}}{\to} i$.

Denote $e_j := e_j^{S(j)}$, we can write $x_j(n) = e_j^{\mathsf{T}} x_{S(j)}(n)$, and therefore from Equation (D3) there exist strictly causal filters $\Gamma_s(z)$ and $\Lambda_{sk}(z)$ (defined for ease of notation) such that

$$x_j(n) = \sum_{s \in S(j)} \Gamma_s(z) v_s(n) + \sum_{\substack{s \in S(j) \\ k \in pa(s) \cap S(j)^{\mathsf{c}}}} \Lambda_{sk}(z) x_k(n)$$

When we substitute this expression into the left hand side of Equation (10), we may cancel each term involving v_s by Lemma D.2, and each $k \in C_0(s)$ by our earlier argument, leaving us with

$$\sum_{\substack{s \in S(j) \\ k \in C_1(s) \cap S(j)^{\mathsf{c}}}} \langle x_i(n), \Phi(z)\Lambda_{sk}(z)x_k(n) - \hat{\mathbb{E}}[\Phi(z)\Lambda_{sk}(z)x_k(n) \mid \mathscr{H}_{t-1}^{(i)}] \rangle$$

Since each $k \in C_1(s)$ with $k \neq i$ inherits properties (a) and (b) above, we can recursively expand each x_k of the above summation until reaching $k = i$ (which is guaranteed to terminate due to the definition of $C_1(u)$) which leaves us with some strictly causal filter $F(z)$ such that the left hand side of Equation (10) is equal to

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

and this is 0 since $\Phi(z)F(z)x_i(n) \in \mathscr{H}_{n-1}^{(i)}$. □

Remark 2.2. 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 ($j \in \mathscr{A}(i)$) or there must be a confounder k through which common information is received.

An interesting corollary is the following:

Corollary 2.1. *If the graph $\mathscr{G}$ is a DAG then $j \overset{\text{PW}}{\to} i, i \overset{\text{PW}}{\to} j \implies \exists k \in \mathscr{A}(i) \cap \mathscr{A}(j)$ confounding (i, j).*

It seems reasonable to expect a converse of Proposition 2.4 to hold, that is, $j \in \mathscr{A}(i) \Rightarrow j \overset{\text{PW}}{\to} i$. Unfortunately, this is not the case in general, as different paths through $\mathscr{G}$ can lead to cancelation (see Figure 1(a)). In fact, we do not even have $j \in pa(i) \Rightarrow j \overset{\text{PW}}{\to} i$ (see Figure 1(b)).

2.5 Strongly causal graphs

In this section and the next we will seek to understand when converse statements of Proposition 2.4 *do* hold. One possibility is to restrict the coefficients of the system matrix, for example, by requiring that $B_{ij}(m) \geq 0$. Instead, we think it more meaningful to focus on the

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

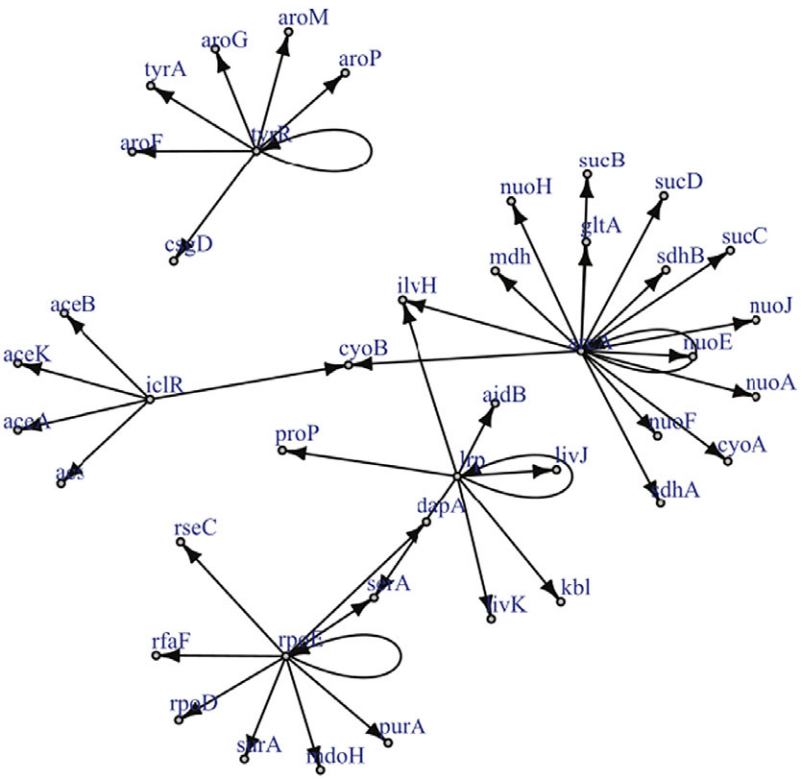

Figure 2. Reproduced under the creative commons attribution non-commercial license (http://creativecommons.org/licenses/by-nc/2.5) is a known gene regulatory network for a certain strain of E.Coli. The graph has only a single edge violating strong causality.

defining feature of time series networks, that is, the topology of $\mathscr{G}$. To this end, we introduce the following notion of a *strongly causal graph* (SCG).

Definition 2.9 (Strongly Causal). *We will say that a Granger causality graph $\mathscr{G}$ is strongly causal if there is at most one directed path between any two nodes. That is, if the graph is a polytree. Strongly Causal Graphs will be referred to as SCGs.*

Examples of strongly causal graphs include directed trees (or forests), the diamond shaped graph containing $i \to j \leftarrow k$ and $i \to j' \leftarrow k$, and Figure 4 of this paper. Importantly, a maximally connected bipartite graph with $2N$ nodes is also strongly causal, demonstrating both 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 2.1. 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 Shojaie & Michailidis (2010) cite an example of the so called "transcription regulatory network of E.coli" (reproduced in Figure 2), and Ma et al. (2014) study a much larger regulatory network of Saccharomyces cerevisiae. These networks, which we reproduce in Figure 3, appear to have at most a small number of edges which violate the strong causality condition.

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

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

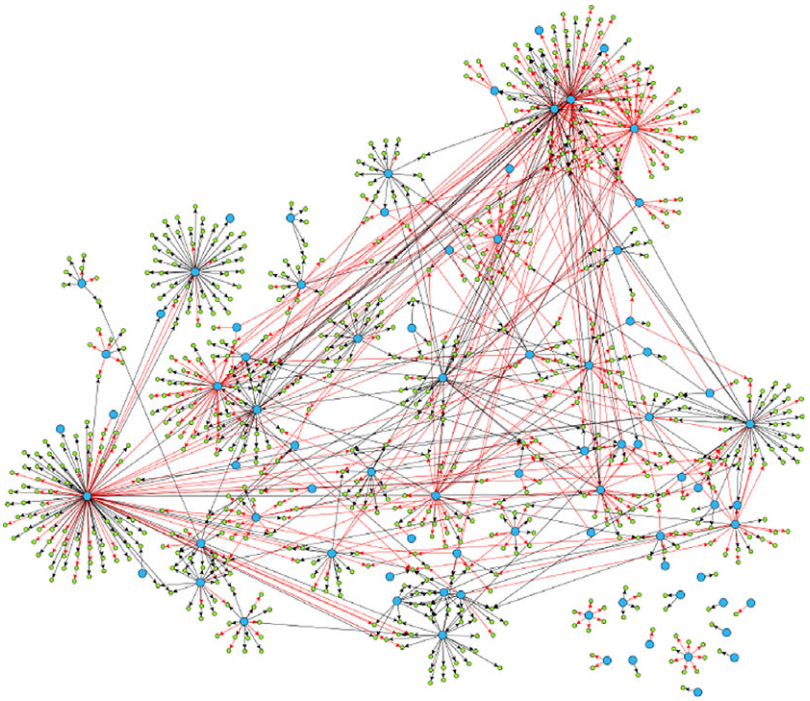

Figure 3. Reproduced under the creative commons attribution license (https://creativecommons.org/licenses/by/4.0/) is a much larger gene regulatory network. It exhibits similar qualitative structure (a network of hub nodes) as does the much smaller network of Figure 2.

following important property essentially strengthens Proposition 2.4 for the case of strongly causal graphs.

Proposition 2.5. *In a strongly causal graph, if* $j \in \mathscr{A}(i)$ *then any* $k \in \mathscr{A}(i) \cap \mathscr{A}(j)$ *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 $j \in \mathscr{A}(i)$) which contradicts the assumption of strong causality. □

Corollary 2.2. *If* $\mathscr{G}$ *is a strongly causal DAG then* $i \overset{\text{PW}}{\to} j$ *and* $j \in \mathscr{A}(i)$ *are alternatives, that is* $i \overset{\text{PW}}{\to} j \Rightarrow j \notin \mathscr{A}(i)$.

Proof. Suppose that $i \overset{\text{PW}}{\to} j$ and $j \in \mathscr{A}(i)$. Then since $\mathscr{G}$ is acyclic $i \notin \mathscr{A}(j)$, and by Proposition 2.4 there is some $k \in \mathscr{A}(i) \cap \mathscr{A}(j)$ which is a confounder. However, by Proposition 2.5 k cannot be a confounder, a contradiction. □

Corollary 2.3. *If* $\mathscr{G}$ *is a strongly causal DAG such that* $i \overset{\text{PW}}{\to} j$ *and* $j \overset{\text{PW}}{\to} i$, *then* $i \notin \mathscr{A}(j)$ *and* $j \notin \mathscr{A}(i)$. *In particular, a pairwise bidirectional edge indicates the absence of any edge in* $\mathscr{G}$.

Proof. This follows directly from applying Corollary 2.2 to $i \overset{\text{PW}}{\to} j$ and $j \overset{\text{PW}}{\to} i$. □

In light of Proposition 2.5, the following provides a partial converse to Proposition 2.4, and supports the intuition of "causal flow" through paths in $\mathscr{G}$.

Proposition 2.6 (Pairwise Causal Flow). *If* $\mathscr{G}$ *is a strongly causal DAG, then* $j \in \mathscr{A}(i) \Rightarrow j \overset{\text{PW}}{\to} i$.

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

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

$$\langle \psi - \hat{\mathbb{E}}[\psi \mid \mathscr{H}_{n-1}^{(i)}], x_i(n) - \hat{\mathbb{E}}[x_i(n) \mid \mathscr{H}_{n-1}^{(i)}] \rangle \neq 0 \tag{11}$$

and therefore that $H_n^{(i)} \not\perp \mathscr{H}_{n-1}^{(j)} \mid \mathscr{H}_{n-1}^{(i)}$, which by Theorem (2.1) is enough to establish that $j \overset{\text{PW}}{\to} i$.

Firstly, we will establish a representation of $x_i(n)$ that involves $x_j(n)$. Denote by $a_{r+1} \to a_r \to \cdots \to a_1 \to a_0$ with $a_{r+1} := j$ and $a_0 := i$ the *unique* $j \to \cdots \to i$ path in $\mathscr{G}$, we will expand the representation of Equation (7) backwards along this path:

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

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

Using this representation to expand Equation (11), we obtain the following cumbersome expression:

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

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

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

Now since $F_{r+1}(z) \neq 0$ by Theorem 2.1, and $F_{r+1}(z)x_j(n) \notin \mathscr{H}_{n-1}^{(i)}$, we have by the Cauchy-Schwarz inequality that this expression is equal to 0 if and only if

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

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

or by rearranging and applying the representation for $x_i(n)$ obtained earlier, if and only if

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

But, this is impossible since $x_i(n) \notin \mathscr{H}_{n-1}^{(i)}$. □

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

Corollary 2.4. *If $\mathscr{G}$ is a strongly causal DAG then $j \overset{\text{GC}}{\to} i \Rightarrow j \overset{\text{PW}}{\to} i$.*

Example 2.2. As a final remark of this subsection we note that a complete converse to Proposition 2.4 is not possible without additional conditions. Consider the "fork" system on 3 nodes (i.e., $2 \leftarrow 1 \rightarrow 3$) defined by

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

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

If we were to augment this system by simply adding an autoregressive component (i.e., some "memory") to $x_1(n)$, for example, $x_1(n) = v_1(n) + bx_1(n-1)$ then we would have $2 \overset{\text{PW}}{\to} 3$ since then $x_3(n) = v_3(n) + av_1(n-1) - bv_2(n-1) + bx_2(n-1)$. 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 2.4 via the notion of a strongly causal graph topology (see Proposition 2.6). In this section, we study conditions under which a converse to part (*b*) will hold.

Definition 2.10 (Lag Function). *Given a causal filter $\mathsf{B}(z) = \sum_{m=0}^{\infty} b(m)z^{-m}$ define*

$$m_0(\mathsf{B}) = \inf\{m \in \mathbb{Z}_+ \mid b(m) \neq 0\} \tag{12}$$

$$m_\infty(\mathsf{B}) = \sup\{m \in \mathbb{Z}_+ \mid b(m) \neq 0\} \tag{13}$$

that is, *the "first" and "last" coefficients of the filter $\mathsf{B}(z)$, where $m_\infty(\mathsf{B}) := \infty$ if the filter has an infinite length, and $m_0(\mathsf{B}) := \infty$ if $\mathsf{B}(z) = 0$.*

Definition 2.11 (Persistent). *We will say that the process $x(n)$ with Granger causality graph $\mathscr{G}$ is persistent if for every $i \in [N]$ and every $k \in \mathscr{A}(i)$ we have $m_0(\mathsf{A}_{ik}) < \infty$ and $m_\infty(\mathsf{A}_{ik}) = \infty$.*

Remark 2.3. 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, see Example 2.3 for a more formal statement.

Moreover, persistence is not the weakest condition necessary for the results of this section: by inspecting our proofs, the condition that for each i, j there is some $k \in \mathscr{A}(i) \cap \mathscr{A}(j)$ such that $m_0(\mathsf{A}_{jk}) < m_\infty(\mathsf{A}_{ik})$ can be seen to be sufficient. The intuition being that nodes i and j are not receiving temporally disjoint information from k.

The etymology for the persistence condition can be explained by supposing that the two nodes i, j each have a loop (i.e., $\mathsf{B}_{ii}(z) \neq 0, \mathsf{B}_{jj}(z) \neq 0$) then this autoregressive component acts as "memory," and so the influence from the confounder k persists in $x_i(n)$, and $m_\infty(\mathsf{A}_{ik}) = \infty$ for each confounder can be expected.

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

Example 2.3 (Ubiquity of Persistent Systems; Proof in Section D.2). Consider a process $x(n)$ generated by the VAR(1) model[2] having $\mathsf{B}(z) = Bz^{-1}$. If B is diagonalizable, and has at least 2 distinct eigenvalues, then $x(n)$ is persistent. A proof of this fact is provided in the Appendix.

This example shows that the collection of finite VAR(p) systems which are not persistent are pathological, in the sense that their system matrices have zero measure.

In order to eliminate the possibility of a particular sort of cancelation, an ad hoc assumption is required. Strictly speaking, the persistence condition is not a necessary or sufficient condition for the following, but cases where the following fails to hold, and persistence *does* hold, are unavoidable pathologies.

Assumption 2.1 (no-cancellation). *Fix $i, j \in [N]$ and let $H_i(z)$ be the strictly causal filter such that*

$$H_i(z)x_i(n) = \hat{\mathbb{E}}[x_i(n) \mid \mathscr{H}_{t-1}^{(i)}]$$

and similarly for $H_j(z)$. Then define

$$T_{ij}(z) = \sum_{k \in \mathscr{A}(i) \cap \mathscr{A}(j)} \sigma_k^2 \mathsf{A}_{ik}(z^{-1})(1 - H_i(z^{-1}))(1 - H_j(z))\mathsf{A}_{jk}(z) \tag{14}$$

where $\sigma_k^2 = \mathbb{E}v_k(n)^2$.

*We will say that Assumption 2.1 is satisfied if for every $i, j \in [N]$, $T_{ij}(z)$ is either constant over z (*i.e., *each z^k coefficient for $k \in \mathbb{Z} \setminus \{0\}$ is 0), or is neither causal (*i.e., *containing only z^{-k} terms, for $k \geq 0$) or anti-causal (*i.e., *containing only z^k terms, for $k \geq 0$). Put succinctly, $T_{ij}(z)$ must be two-sided.* □

Remark 2.4. Under the condition of persistence, the only way for Assumption 2.1 to fail is through cancelation in the terms defining $T_{ij}(z)$. For example, the condition is assured if $x(n)$ is persistent, and there is only a single confounder. Unfortunately, some pathological behavior resulting from confounding nodes seems to be unavoidable without some assumptions (regarding sets of zero measure) about the parameters of the MA(∞) system defining $x(n)$.

Proposition 2.7 (Pairwise Causation and Confounders). *Fix $i, j \in [N]$ and suppose $\exists k \in \mathscr{A}(i) \cap \mathscr{A}(j)$ which confounds i, j. Then, if $T_{ij}(z)$ is not causal we have $j \overset{\text{PW}}{\to} i$, and if $T_{ij}(z)$ is not anti-causal we have $i \overset{\text{PW}}{\to} j$. Moreover, if Assumption 2.1 is satisfied, then $j \overset{\text{PW}}{\to} i \iff i \overset{\text{PW}}{\to} j$.*

Proof. Recalling Theorem 2.1, consider some $\psi \in \mathscr{H}_{n-1}^{(j)}$ and represent it as $\psi(n) = F(z)x_j(n)$ for some strictly causal filter $F(z)$. Then

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

where (a) applies the orthogonality principle, (b) expands with Equation (8) with $H_i(z)x_i(n) = \hat{\mathbb{E}}[x_i(n) \mid \mathscr{H}_{t-1}^{(i)}]$, and (c) follows by performing cancelations of $v_k(n) \perp v_\ell(n)$ and noting that by the contra-positive of Proposition 2.5 we cannot have $i \in \mathscr{A}(j)$ or $j \in \mathscr{A}(i)$.

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

Through symmetric calculation, we can obtain the expression relevant to the determination of $i \overset{\text{PW}}{\to} j$ for $\phi \in \mathscr{H}_{n-1}^{(i)}$ represented by the strictly causal filter $G(z) : \phi(n) = G(z)x_i(n)$

$$\begin{aligned}&\langle \phi(n) - \hat{\mathbb{E}}[\phi(n) \mid \mathscr{H}_{t-1}^{(j)}], x_j(n) - \hat{\mathbb{E}}[x_j(n) \mid \mathscr{H}_{t-1}^{(j)}] \rangle \\ &= \sum_{k \in \mathscr{A}(i) \cap \mathscr{A}(j)} \langle G(z)\mathsf{A}_{ik}(z)v_k(n), (1 - H_j(z))\mathsf{A}_{jk}(z)v_k(n) \rangle\end{aligned}$$

where $H_j(z)x_j(n) = \hat{\mathbb{E}}[x_j(n) \mid \mathscr{H}_{t-1}^{(j)}]$.

We have therefore

$$(j \overset{\text{PW}}{\to} i) : \exists F(z) \text{ s.t. } \sum_{k \in \mathscr{A}(i) \cap \mathscr{A}(j)} \langle F(z)\mathsf{A}_{jk}(z)v_k(n), (1 - H_i(z))\mathsf{A}_{ik}(z)v_k(n) \rangle \neq 0 \tag{15}$$

$$(i \overset{\text{PW}}{\to} j) : \exists G(z) \text{ s.t. } \sum_{k \in \mathscr{A}(i) \cap \mathscr{A}(j)} \langle G(z)\mathsf{A}_{ik}(z)v_k(n), (1 - H_j(z))\mathsf{A}_{jk}(z)v_k(n) \rangle \neq 0 \tag{16}$$

The persistence condition, by Corollary D.1, ensures that for each $k \in \mathscr{A}(i) \cap \mathscr{A}(j)$ there is some $F(z)$ and some $G(z)$ such that at least one of the above terms constituting the sum over k is non-zero. It remains to eliminate the possibility of cancelation in the sum.

The adjoint of a linear filter $C(z)$ is simply $C(z^{-1})$, which recall is strictly anti-causal if $C(z)$ is strictly causal. Using this, we can write

$$\begin{aligned}&\sum_{k \in \mathscr{A}(i) \cap \mathscr{A}(j)} \langle F(z)\mathsf{A}_{jk}(z)v_k(n), (1 - H_i(z))\mathsf{A}_{ik}(z)v_k(n) \rangle \\ &= \sum_{k \in \mathscr{A}(i) \cap \mathscr{A}(j)} \langle \mathsf{A}_{ik}(z^{-1})(1 - H_i(z^{-1}))F(z)\mathsf{A}_{jk}(z)v_k(n), v_k(n) \rangle\end{aligned}$$

Moreover, it is sufficient to find some strictly causal $F(z)$ of the form $F(z)(1 - H_j(z))$ (abusing notation) since $1 - H_j(z)$ is causal. Similarly for $G(z)$, this leads to symmetric expressions for $j \overset{\text{PW}}{\to} i$ and $i \overset{\text{PW}}{\to} j$ respectively:

$$\sum_{k \in \mathscr{A}(i) \cap \mathscr{A}(j)} \langle \mathsf{A}_{ik}(z^{-1})(1 - H_i(z^{-1}))F(z)(1 - H_j(z))\mathsf{A}_{jk}(z)v_k(n), v_k(n) \rangle \tag{17}$$

$$\sum_{k \in \mathscr{A}(i) \cap \mathscr{A}(j)} \langle \mathsf{A}_{ik}(z^{-1})(1 - H_i(z^{-1}))G(z^{-1})(1 - H_j(z))\mathsf{A}_{jk}(z)v_k(n), v_k(n) \rangle \tag{18}$$

Recall the filter from Assumption 2.1

$$T_{ij}(z) = \sum_{k \in \mathscr{A}(i) \cap \mathscr{A}(j)} \sigma_k^2 \mathsf{A}_{ik}(z^{-1})(1 - H_i(z^{-1}))(1 - H_j(z))\mathsf{A}_{jk}(z) \tag{19}$$

Since each $v_k(n)$ is uncorrelated through time, $\langle T_{ij}(z)v_k(n), v_k(n) \rangle = \sigma_k^2 T_{ij}(0)$, and therefore we have $j \overset{\text{PW}}{\to} i$ if $T_{ij}(z)$ is *not* causal and $i \overset{\text{PW}}{\to} j$ if $T_{ij}(z)$ it *not* anti-causal. Moreover, we have $i \overset{\text{PW}}{\nrightarrow} j$ *and* $j \overset{\text{PW}}{\to} i$ if $T_{ij}(z)$ is a constant. Therefore, under Assumption 2.1 $j \overset{\text{PW}}{\to} i \iff i \overset{\text{PW}}{\to} j$.

This follows since if $T_{ij}(z)$ is not causal then $\exists k > 0$ such that the z^k coefficient of $T_{ij}(z)$ is non-zero, and we can choose strictly causal $F(z) = z^{-k}$ such that (D11) is non-zero and therefore $j \overset{\text{PW}}{\to} i$.

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

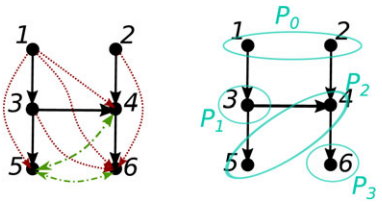

Figure 4. Black arrows indicate true parent-child relations. Red dotted arrows indicate pairwise causality (due to non-parent relations), green dash-dotted arrows indicate bidirectional pairwise causality (due to the confounding node 1). Blue groupings indicate each P_k in Algorithm 1.

Similarly, if $T_{ij}(z)$ is not anti-causal, then $\exists k > 0$ such that the z^{-k} coefficient of $T_{ij}(z)$ is non-zero, and we can choose strictly causal $G(z)$ so that $G(z^{-1}) = z^k$, and then (D12) is non-zero and therefore $i \overset{\text{PW}}{\to} j$. □

Remark 2.5. The importance of this result is that when $i \overset{\text{PW}}{\to} j$ is a result of a confounder k, then $i \overset{\text{PW}}{\to} j \iff j \overset{\text{PW}}{\to} i$. This implies that in a strongly causal graph every bidirectional pairwise causality relation must be the result of a confounder. Therefore, in a strongly causal graph, pairwise causality analysis is immune to confounding (since we can safely remove all bidirectional edges).

2.7 Recovering $\mathcal{G}$ via pairwise tests

We arrive at the main conclusion of the theoretical analysis in this paper.

Theorem 2.2 (Pairwise Recovery; Proof in Section D.3). *If the Granger causality graph $\mathcal{G}$ for the process $x(n)$ is a strongly causal DAG and Assumption 2.1 holds, 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).*

Example 2.4. The set W of Algorithm 1 collects ancestor relations in $\mathcal{G}$ (see Lemma D.5). In reference to Figure 4, 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 as results of confounding. The groupings $P_0, \ldots, P_3$ are also indicated in Figure 4.

The algorithm proceeds first with the parent-less 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 $1 \to 3 \to 4$ and $1 \to 3 \to 5$ 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 on r can also be seen from this example, where if instead we simply added the set

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

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

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

causes the edge $1 \to 3$ to be missed.

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

Algorithm 1: Pairwise Granger Causality Algorithm

Algorithm: 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 $\mathscr{G}$.

output : Edges $\mathscr{E} = \{(i,j) \in [N] \times [N] \mid i \overset{\text{GC}}{\to} j\}$ of the graph $\mathscr{G}$.

initialize: $S_0 = [N]$ `# unprocessed nodes`
$E_0 = \emptyset$ `# edges of` $\mathscr{G}$
$k = 1$ `# a counter`

$W \leftarrow \{(i,j) \mid i \overset{\text{PW}}{\to} j, j \overset{\text{PW}}{\nrightarrow} i\}$
`# parentless nodes`
$P_0 \leftarrow \{i \in S_0 \mid \forall s \in S_0\ (s,i) \notin W\}$
while $S_{k-1} \neq \emptyset$ **do**
  `# remove nodes with depth` $k-1$
  $S_k \leftarrow S_{k-1} \setminus P_{k-1}$
  `# candidate children`
  $P_k \leftarrow \{i \in S_k \mid \forall s \in S_k\ (s,i) \notin W\}$
  $D_{k0} \leftarrow \emptyset$
  for $r = 1, \ldots, k$ **do**
    `# currently known edges`
    $Q \leftarrow E_{k-1} \cup \left(\bigcup_{\ell=0}^{r-1} D_{k\ell}\right)$
    $D_{kr} \leftarrow \{(i,j) \in P_{k-r} \times P_k \mid$
     $(i,j) \in W,$ no $i \to j$ path in $Q\}$
  `# update` E_k `with new edges`
  $E_k \leftarrow E_{k-1} \cup \left(\bigcup_{r=1}^{k} D_{kr}\right)$
  $k \leftarrow k+1$
return E_{k-1}

3. Simulations

We implement an heuristic inspired by Algorithm 1 by replacing the population statistics with finite sample tests, the details of which can be found in the supplementary material Section B.5 (see Algorithm 2). The heuristic is essentially controlling the false discovery rate substantially below what it would be with a threshold based pairwise scheme. The methods are easily parallelizable, and can scale to graphs with thousands of nodes on simple personal computers. By contrast, scaling the LASSO to this large of a network (millions of variables) is not trivial.

We first ran experiments using two separate graph topologies having $N = 50$ nodes, and with a total of T samples drawn from the resulting system. The graphs we consider are a strongly causal graph (SCG) and a directed acyclic graph (DAG), and the filters along each edges are constructed by uniformly at random placing $p = 5$ poles in a disc of radius $3/4$. We compare our results against the adaptive LASSO (adaLASSO) (Zou, 2006), which, since it enjoys the *oracle property* should be expected to, asymptotically, fully recover the underlying causality graph. Indeed, we found early on that this LASSO variant drastically out-performed alternative LASSO-based procedures including the ordinary LASSO, grouped LASSO, and elastic net by wide margins.

The usual methodology of applying the LASSO to Granger causality is to split the square one-step-ahead prediction error into separate terms, one for each group of incident edges on a

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

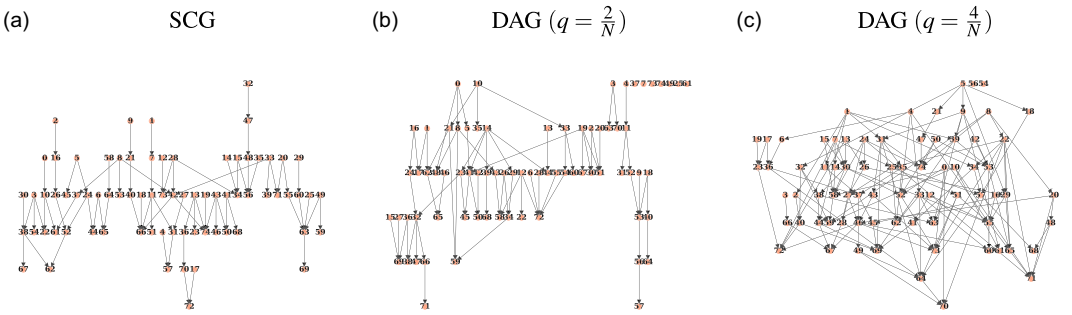

Figure 5. Representative random graph topologies on $N = 50$ nodes.

node, and estimating the collection of N incident filters $\{\mathsf{B}_{ij}(z)\}_{j=1}^{N}$ that minimizes ξ_i^{LASSO} in the following:

$$\begin{aligned} e(n) &= x_i(n) - \sum_{m=1}^{p_{\max}} \sum_{j=1}^{N} B_{i,j}(m) x(n-m) \\ \xi_i^{\text{LASSO}}(\lambda) &= \min_{B} \frac{1}{T} \sum_{n=p_{\max}+1}^{T} e(n)^2 + \lambda \sum_{m=1}^{p_{\max}} \sum_{j=1}^{N} |B_{ij}(m)| \\ \xi_i^{\text{LASSO}} &= \min_{\lambda \geq 0} \left[\xi_i^{\text{LASSO}}(\lambda) + \mathsf{BIC}\big(B_i^{\text{LASSO}}(\lambda)\big) \right] \end{aligned} \tag{20}$$

where we are choosing λ, the regularization parameters, via the Bayesian Information Criteria (see, i.e., Claeskens et al. (2008)). This is similar to the work of Arnold et al. (2007), except that we have replaced the LASSO with the better performing adaLASSO.

Remark 3.1 (Graph Topologies). We depict in Figure 5 some representative graph topologies used in our empirical evaluation. To construct these graphs we first use simple Erdos-Renyi graphs $\mathscr{G}(N, q)$ with N nodes and edge probability q, and then remove edges until the graph is acyclic. 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.

Remark 3.2 (MCC as a Support Recovery Measurement). We apply Matthew's Correlation Coefficient (MCC) (Matthews, 1975; Chicco & Jurman, 2020) as a statistic for measuring support recovery performance (see also Chicco (2017) tip # 8). 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 selection, and -1 being perfectly opposed.

Remark 3.3 (Error Measurement). We estimate the out of sample 1-step ahead prediction error by forming the variance matrix estimate

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

on a long stream of T_{out} samples of out-of-sample data. We then report the Log-Relative-Error:

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

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

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

Table 1. Simulation results—PWGC vs AdaLASSO

Metric		LRE		MCC	
Algorithm		alasso	pwgc	alasso	pwgc
T	*q*				
50					
	SCG	1.71	**1.55**	0.46	**0.55**
	0.04	1.97	**1.77**	0.41	**0.53**
	0.08	2.95	**2.72**	0.36	**0.39**
	0.32	9.02	**8.17**	**0.14**	0.10
250					
	SCG	1.30	**1.18**	0.70	**0.81**
	0.04	1.40	**1.31**	0.68	**0.76**
	0.08	2.49	**2.21**	0.55	**0.57**
	0.32	8.67	**7.62**	**0.18**	0.15
1250					
	SCG	1.20	**1.11**	0.68	**0.88**
	0.04	1.28	**1.20**	0.64	**0.84**
	0.08	2.12	2.05	0.60	**0.64**
	0.32	7.78	**7.39**	**0.21**	0.18

1. Results of Monte Carlo simulations comparing PWGC and AdaLASSO ($N = 50, p = 5, p_{\max} = 10$) for small samples and when the SCG assumption doesn't hold. The superior result is bolded when the difference is statistically significant, as measured by `scipy.stats.ttest_rel`. 100 iterations are run for each set of parameters.
2. LRE: Log-Relative-Error, that is, the log sum of squared errors at each node relative to the strength of the driving noise $\frac{\ln \mathrm{tr}\hat{\Sigma}_\nu}{\ln \mathrm{tr}\Sigma_\nu}$. MCC: Matthew's Correlation Coefficient.
3. Values of q (edge probability) range between $2/N$, $4/N$, $16/N$ where $2/N$ has the property that the random graphs have on average the same number of edges as the SCG.

3.1 Results

Our first simulation results are summarized in Table 1. It is clear that the superior performance of PWGC in comparison to AdaLASSO is as a result of limiting the false discovery rate. It is unsurprising that PWGC exhibits superior performance when the graph is an SCG, but even in the case of more general DAGs, the PWGC heuristic is still able to more reliably uncover the graph structure for small values of q. We would conjecture that for small q, random graphs are "likely" to be "close" to SCGs in some appropriate sense. As q increases, there are simply not enough edges allowed by the SCG topology for it to be possible to accurately recover $\mathscr{G}$.

Figures 6 and 7 provide an empirical analysis of how the algorithm scales with increasing N, and for a small number of training samples T. It is surprising that even for general DAGs (which don't necessarily have the SCG property) performance degrades extremely slowly with respect to N, and that performance is markedly better than random even when $N > T$. These observations are reminiscent of results for the LASSO algorithm applied to linear regression (Tibshirani et al., 2015). For example, under certain conditions, it holds that

$$||\hat{\beta}_{LASSO} - \beta^\star||_2^2 = \mathscr{O}\left(\frac{s}{T} \ln N\right)$$

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

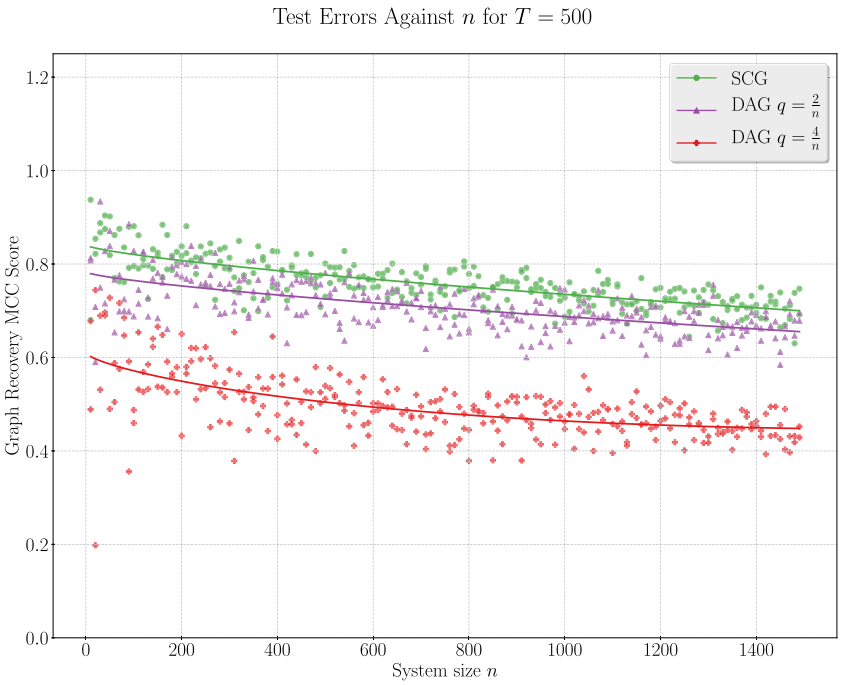

Figure 6. Measures support recovery performance as the number of nodes *N* increases, and the edge proportion as well as the number of samples *T* is held fixed. Remarkably, the degradation as *N* increases is limited, it is primarily the graph topology (SCG or non-SCG) as well as the level of sparsity (measured by *q*) which are the determining factors for support recovery performance.

where s is the number of non-zero parameters of $\beta^\star$ in the ground-truth linear regression model (Wainwright, 2009).

Additional experimental results, and a detailed description of the heuristic algorithm based on the theory of Section 2.7 is available in Appendix C.

4. Application example: Alcoholism classification from EEG networks

As was reviewed in the introduction to this paper, Granger causality analysis is frequently utilized for the analysis of brain connectivity networks in Neuroscience. This application is by now well established (Bressler & Seth, 2011), but is not without criticism. In particular, the critical paper of Haufe et al. (2013) states that "*Many problems in neuroimaging (such as brain connectivity analysis) are inherently unsupervised, which means that the 'ground truth' cannot be retrieved. In these cases, simulations are the only way to benchmark a method's ability to solve the task if theoretical results are not available, while a neuroscientific finding on real data that matches prior expectations should not be mistaken for a proof-of-concept of the method.*" This critical point has clear merit. Indeed, the motivation behind the simulation results of Section 3 is to apply our methodology to real data in a situation where we *can* identify the ground truth (since we constructed it), and therefore make certain types of comparisons between algorithms.

In the present application example, we challenge this dichotomy. We apply the Granger causality inference heuristic derived in this paper to construct estimates of signal space EEG causality networks for a large number of patients. A proportion of the patients in the data set have alcoholism, and we will show that, using only the inferred causality network, a logistic regression classifier can be constructed which discriminates between networks calculated from control subjects, and alcoholic subjects. The fact that it is possible to successfully perform this classification,

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

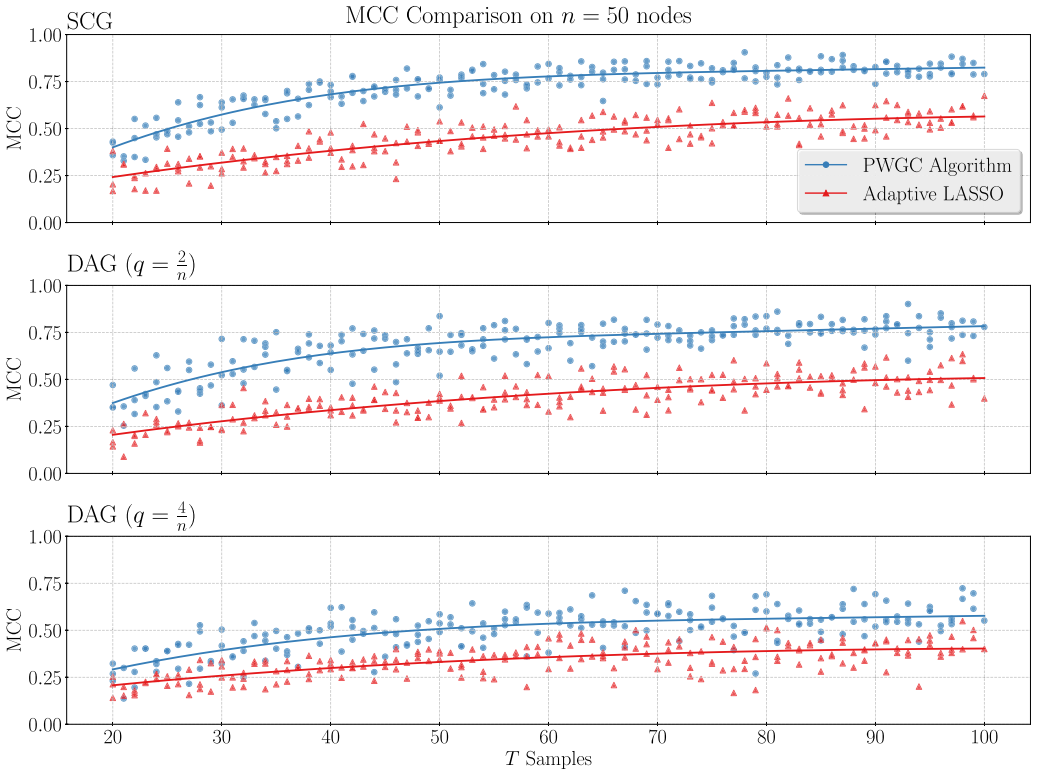

Figure 7. Provides a support recovery comparison for very small values of, T typical for many applications. The system has a fixed number $N = 50$ nodes.

at a rate better than random, is evidence that the Granger causality analysis detects and recovers consistent differences in the GCG networks of control and alcoholic EEG readings. This mode of analyzing the validity of a causality inference algorithm is a third method, conceptually distinct from the theory-or-simulation dichotomy identified by Haufe et al. (2013).

The data set we make use of was obtained from the "EEG Database Data Set"[3] (Zhang et al., 1995) from the UCI machine learning repository (Dua & Graff, 2017). This data set contains 1 s long measurements of 64 channel EEG signals from patients who are given visual stimuli. The original data set contributors have filtered the frequencies of the recorded signals to lie between 0.02 and 50 Hz, and removed samples having excessive artifacts. As well as various important pre-processing steps particular to EEG data.

Precisely, the data set is described by

$$\mathscr{D} = \{\big(x^{(i)}(n)\big)_{n=1}^{T}, z^{(i)}\}_{i=1}^{N}$$

where $T = 256$, $x^{(i)}(n) \in \mathbb{R}^{64}$, $z^{(i)} \in \{\mathsf{control}, \mathsf{alcoholic}\}$, and N = 10,723. That is, $x^{(i)}(n)$ is the EEG recordings of example i, and $z^{(i)}$ is a binary marker indicating the presence or absence of alcoholism in the subject. There are on average 90 trials for each subject, ranging between 30 and 119.

Applying a Granger causality inference algorithm to this data set results in a derivative data set $\widehat{\mathscr{D}} = \{A^{(i)}, z^{(i)}\}_{i=1}^{N}$, where $A^{(i)} \in \{0, 1\}^{64\times 64}$ is the adjacency matrix of the estimated causality network $\widehat{\mathscr{G}}^{(i)}$ from example i.

By fitting a classifier $\mathsf{clf} : \mathscr{G} \to \{\mathsf{control}, \mathsf{alcoholic}\}$, which uses *only* the information available in the estimated causality network, we are able to make inference about whether or not the estimated causality graphs $\widehat{\mathscr{G}}^{(i)}$ are uncovering relevant factors.

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

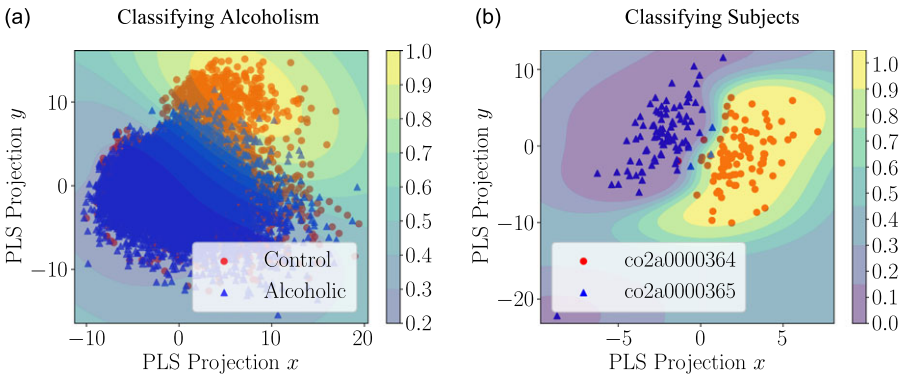

Figure 8. Is a low-dimensional embedding representing the classification of alcoholic and non-alcoholic subjects based entirely on the Granger causality graphs constructed from EEG signals. Figure 8b further illustrates a classification task using EEG causality graphs. In this case, one particular subject can be almost perfectly separated from another particular subject with a binary classifier. These results indicate the causality graph carries consistent and identifiable cross-subject patterns.

4.1 Classifier methodology and evaluation

The classifier we apply is an off-the-shelf polynomial kernel logistic regression model: `sklearn.linear_model.LogisticRegression` (Pedregosa et al., 2011), using a Nystroem kernel approximation (with 240 components): `sklearn.kernel_approximation.Nystroem(kernel = 'poly')`. Hyperparameters are chosen via 10-fold cross validation with Bayesian Optimization (Kumar & Head, 2017), using 80% of the data. The remaining 20% of the data is held out for a final validation step. A polynomial kernel is natural for this task since the input features are binary (the adjacency matrix); a kernel degree of, say, 3, indicates that graphical substructures no more complicated than triangles can be used by the classifier. Similarly to our simulation results, we use the MCC as the metric for evaluation.

As a final verification, a label permutation test ("Test 1" in Ojala & Garriga (2010)) is applied to verify that $A^{(i)}$ and $(z^{(i)})$ are not independent, and to alleviate the fear that the classifier may simply be picking up on spurious patterns in the estimated causality graphs. This test consists of refit our classifier on randomly relabeled data and comparing the results those of the classifier fit on the correctly labeled data. An inability of the model to accurately discriminate between the labels in an incorrectly labeled data set is further evidence that accurate classification is not simply a result of overfitting or of picking up on spurious variation in the data $A^{(i)}$.

4.2 Results

The final value of the MCC for the alcoholism classification task, computed on a 10-fold cross validation is $MCC = 0.305$, with the individual folds falling into the interval $[0.229, 0.396]$. The MCC obtained on the final held-out validation set (i.e., data which was not used for picking *any* model parameters) was 0.29. The p-value estimated from the label permutation test satisfies $p < 0.01$ and where the average MCC obtained on the validation data, by the 100 classifiers fit on data with randomly permuted labels, was zero through two decimal places. While the performance of the classifier is not exemplary, these results establish clearly that the performance is statistically significantly better than random.

The EEG alcoholism data set has other potential classification labels. One natural example is to classify the *subject* label. That is, to identify the subject to which the EEG example belongs (each subject had multiple EEG recordings taken). We do not report the details as the task is considerably easier and the methodology is similar to that described in Section 4.1. However, a classifier trained on any two specific subjects is able to discern out-of-sample causality graphs

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

between the two subjects with almost perfect accuracy. Specifically, we obtained an MCC of 0.99 and the p-value estimated by a permutation test satisfied $p < 0.001$.

An illustration of the alcoholism classifier, as well as a particular example from the subject-subject classifier is provided in Figure 8. The visualization is constructed by projecting onto two basis dimensions by partial least squares (a supervised dimensionality reduction technique). These visualizations are provided entirely for *qualitative context*, but drawing substantive conclusions from them is not appropriate. In particular, the low-dimensional visualization does not necessarily accurately capture the high-dimensional decision surface utilized by the classifiers.

5. Conclusion

We have argued that considering particular topological properties of Granger causality networks can provide substantial insights into the structure of causality graphs with potential for providing improvements to causality graph estimation when structural assumptions are met. 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.

In addition to this simulation study, we have constructed an application example using EEG data. In this example, we use off-the-shelf machine learning classifiers to detect the presence or absence of alcoholism based entirely upon the Granger causality graph topology. While this alcoholism classification task is quite challenging, we demonstrate that classifying between distinct subjects in the study, again based only upon the inferred Granger causality graphs, is easy: such classifiers can obtain nearly 100% accuracy. This evidence demonstrates that consistent and meaningful features are being extracted by the Granger causality analysis.

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 advocate for further study of how different topological assumptions may impact the recovery of causality graphs.

Competing interests. None.

Supplementary materials. For supplementary material for this article, please visit http://doi.org/10.1017/nws.2023.11

Notes

1 We are using the convention that $\mathsf{B}_{ij}(z)$ is a filter with input $x_j(\cdot)$ and output $x_i(\cdot)$ so as to write the action of the system as $\mathsf{B}(z)x(n)$ with $x(n)$ 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.

2 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.

3 http://archive.ics.uci.edu/ml/datasets/EEG+Database

References

Ahelegbey, D. F., Billio, M., & Casarin, R. (2016). Bayesian graphical models for structural vector autoregressive processes. *Journal of Applied Econometrics*, *31*(2), 357–386.

Arnold, A., Liu, Y., & Abe, N. (2007). Temporal causal modeling with graphical granger methods. In Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 66-75). ACM

Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap.In Proceedings of the 25th international conference on machine learning (pp. 33-40). ACM

Bach, F. R., & Jordan, M. I. (2004). Learning graphical models for stationary time series. *IEEE Transactions on Signal Processing*, *52*(8), 2189–2199.
Barnett, L., & Seth, A. K. (2015). Granger causality for state-space models. *Physical ReviewE, 91*(4), 040101.
Barnett, L., Barrett, A. B., & Seth, A. K. (2009). Granger causality and transfer entropy are equivalent for gaussian variables. *Physical Review*Letters, *103*(23), 238701.
Bassett, D. S., & Sporns, O. (2017). Network neuroscience. *Nature Neuroscience*, *20*(3), 353–364.
Basu, S., & Michailidis, G. (2015). Regularized estimation in sparse high-dimensional time series models. *The Annals of Statistics*, *43*(4), 1535–1567.
Basu, S., Li, X., & Michailidis, G. (2019). Low rank and structured modeling of high-dimensional vector autoregressions. *IEEE Transactions on Signal Processing*, *67*(5), 1207–1222.
Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.
Billio, M., Getmansky, M., Lo, A. W., & Pelizzon, L. (2010 (7)). Econometric measures of systemic risk in the finance and insurance sectors. *Working Paper 16223*, National Bureau of Economic Research.
Bolstad, A., Van Veen, B. D., & Nowak, R. (2011). Causal network inference via group sparse regularization. *IEEE Transactions on Signal Processing*, *59*(6), 2628–2641.
Bressler, S. L., & Seth, A. K. (2011). Wiener–granger causality: a well established methodology. *Neuroimage*, *58*(2), 323–329.
Caines, P. (1976). Weak and strong feedback free processes. *IEEE Transactions on Automatic Control*, *21*(5), 737–739.
Caines, P., & Chan, C. (1975). Feedback between stationary stochastic processes. *IEEE Transactions on Automatic Control*, *20*(4), 498–508.
Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *Biodata Mining*, *10*(1), 35.
Chicco, D., & Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13.
Chow, C., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, *14*(3), 462–467.
Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*, vol. *330*. Cambridge: Cambridge University Press.
Datta-Gupta, S., & Mazumdar, R. R. (2013). Inferring causality in networks of wss processes by pairwise estimation methods. In *2013 information theory and applications workshop (ITA)*(pp. 1-9). IEEE.
Datta Gupta, S. (2014). On mmse approximations of stationary time series, Ph.D. thesis. University of Waterloo.
David, O., Guillemain, I., Saillet, S., Reyt, S., Deransart, C., Segebarth, C. . . . Depaulis, A. (2008). Identifying neural drivers with functional mri: an electrophysiological validation. *Plos Biol*, *6*(12), e315.
Dua, D., & Graff, C. (2017). UCI machine learning repository.
Durbin, J. (1960). The fitting of time-series models. *Revue de l'institut international de statistique*, *28*(3), 233–244.
Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441.
Fujita, A., Sato, J. R., Garay-Malpartida, H. M., Yamaguchi, R., Miyano, S., Sogayar, M. C. . . . Ferreira, C. E. (2007). Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology*, *1*(1), 39.
Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, *37*(3), 424–438.
Granger, C. W. J. (1980). Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and Control*, *2*, 329–352.
Hagberg, A., Swart, P., & Schult, D. (2008). Exploring network structure, dynamics, and function using networkx. Tech. rept. Los Alamos, NM (United States): Los Alamos National Lab.(LANL).
Hallac, D., Park, Y., Boyd, S. P., & Leskovec, J. (2017). *Network inference via the time-varying graphical lasso*. Corr, **abs/1703.01958**
Hastie, T., Tibshirani, R., & Tibshirani, R. J. (2017). Extended comparisons of best subset selection, forward stepwise selection, and the lasso, arxiv preprint arxiv: 1707.08692.
Haufe, S., Müller, K.-R., Nolte, G., & Krämer, N. (2008). Sparse causal discovery in multivariate time series. In Proceedings of the 2008th international conference on causality: Objectives and assessment-volume 6 (pp. 97-106), JMLR. org.
Haufe, S., Nikulin, V. V., Müller, K.-R., & Nolte, G. (2013). A critical assessment of connectivity measures for EEG data: A simulation study. *Neuroimage*, *64*, 120–133.
Haury, A.-C., Mordelet, F., Vera-Licona, P., & Vert, J.-P. (2012). Tigress: trustful inference of gene regulation using stability selection. *BMC Systems Biology*, *6*(1), 145.
Hayes, M. H. (2009). *Statistical digital signal processing and modeling*. John Wiley & Sons.
He, B., Astolfi, L., Valdés-Sosa, P. A., Marinazzo, D., Palva, S. O., Bénar, C.-G. . . . Koenig, T. (2019). Electrophysiological brain connectivity: theory and implementation. *IEEE Transactions on Biomedical Engineering*, *66*(7), 2115–2137.
He, Y., She, Y., & Wu, D. (2013). Stationary-sparse causality network learning. *Journal of Machine Learning Research*, *14*(1), 3073–3104.
Innocenti, G., & Materassi, D. (2008). A modeling approach to multivariate analysis and clusterization theory. *Journal of Physics A: Mathematical and theoretical*, *41*(20), 205101.

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

James, R. G., Barnett, N., & Crutchfield, J. P. (2016). Information flows? a critique of transfer entropies. *Physical Review Letters*, *116*(23), 238701.
Jones, E., Oliphant, T., & Peterson, P. (2001). SciPy: Open source scientific tools for Python.
Korzeniewska, A., Crainiceanu, C. M., Kuś, R., Franaszczuk, P. J., & Crone, N. E. (2008). Dynamics of event-related causality in brain electrical activity. *Human Brain Mapping*, *29*(10), 1170–1192.
Kumar, G. L. M., & Head, T. (2017). Scikit-optimize. *Tim head and contributors*.
Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, *44*(3), 907–927.
Lindquist, A., & Picci, G. (2015). *Linear stochastic systems: A geometric approach to modeling, estimation and identification*, vol. *1*. Springer.
Lozano, A. C., Abe, N., Liu, Y., & Rosset, S. (2009). Grouped graphical granger modeling for gene expression regulatory networks discovery. *Bioinformatics*, *25*(12), i110–i118.
Lütkepohl, H. (2005). *New introduction to multiple time series analysis*. Springer Science & Business Media.
Michail, M., Kannan, R., Chelmis, C., & Prasanna, V. K. (2016). Sparse causal temporal modeling to inform power system defense. *Procedia Computer Science*, *95*, 450–456. Complex Adaptive Systems Los Angeles, CA November 2-4, 2016.
Ma, S., Kemmeren, P., Gresham, D., & Statnikov, A. (2014). De-novo learning of genome-scale regulatory networks in S. cerevisiae. *Plos One*, *9*(9), 1–20.
Materassi, D., & Innocenti, G. (2010). Topological identification in networks of dynamical systems. *IEEE Transactions on Automatic Control*, *55*(8), 1860–1871.
Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et biophysica acta (bba)-Protein Structure*, *405*(2), 442–451.
Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(4), 417–473.
Michailidis, G., & d'Alché Buc, F. (2013). Autoregressive models for gene regulatory network inference: sparsity, stability and causality issues. *Mathematical Biosciences*, *246*(2), 326–334.
MJózsa. (2019). Relationship between granger non-causality and network graph of state-space representations, Ph.D. thesis. University of Groningen.
Nardi, Y., & Rinaldo, A. (2011). Autoregressive process modeling via the lasso procedure. *Journal of Multivariate Analysis*, *102*(3), 528–549.
Newman, M., & Watts, D. J. (1999). Renormalization group analysis of the small-world network model. *Physics LettersA, 263*(4-6), 341–346.
Nguyen, P. (2019). Methods for inferring gene regulatory networks from time series expression data. Ph.D. thesis. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2019-05-11.
Ojala, M., & Garriga, G. C. (2010). Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, *11*(Jun), 1833–1863.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.. . .Dubourg, V. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct), 2825–2830.
Rebane, G., & Pearl, J. (1987). The recovery of causal poly-trees from statistical data. In UAI '87: Proceedings of the third annual conference on uncertainty in artificial intelligence, Seattle, WA, USA: Elsevier, July 10-12, *1987*.
Seth, A. K., Barrett, A. B., & Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, *35*(8), 3293–3297.
Shojaie, A., & Michailidis, G. (2010). Discovering graphical Granger causality using the truncating lasso penalty. *Bioinformatics*, *26*(18), i517–i523.
Solo, V. (2015). State space methods for granger-geweke causality measures, arxiv preprint arxiv: 1501.04663.
Tam, G. H. F., Chang, C., & Hung, Y. S. (2013). Gene regulatory network discovery using pairwise granger causality. *IET Systems Biology*, *7*(5), 195–204.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.
Tibshirani, R., Wainwright, M. J., & Hastie, T. (2015). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.
Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, *55*(5), 2183–2202.
Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.
Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of small-worldnetworks. *Nature*, *393*(6684), 440–442.
Whittle, P. (1963). On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika*, *50*(1-2), 129–134.
Wong, K. C., Li, Z., & Tewari, A. (2016). Lasso guarantees for time series estimation under subgaussian tails beta-mixing, arxiv preprint arxiv: 1602.04265.

https://doi.org/10.1017/nws.2023.11 Published online by Cambridge University Press

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.
Yuan, T., & Qin, S. J. (2014). Root cause diagnosis of plant-wide oscillations using granger causality. *Journal of Process Control*, *24*(2), 450–459.
Zhang, X. L., Begleiter, H., Porjesz, B., Wang, W., & Litke, A. (1995). Event related potentials during object recognition tasks. *Brain Research Bulletin*, *38*(6), 531–538.
Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

Cite this article: Kinnear R. J. and Mazumdar R. R. (2023). Exact recovery of Granger causality graphs with unconditional pairwise tests. *Network Science* **11**, 431–457. https://doi.org/10.1017/nws.2023.11