A Overview

We detail the methods used for our simulations and finite sample implementation in Section B and provide additional simulation results in Section C. All of the proofs of the paper's results that were not given in the main text can be found in Section D.

Code will be made available at `https://github.com/RJTK/granger_causality`.

B Pairwise Recovery Methods with Finite Datasets

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 [67, 35], and elsewhere.

In subsection B.1 we define pairwise Granger causality hypothesis tests, in subsection B.2 a model order selection criteria, in subsection B.3 an efficient estimation algorithm, in subsection B.4 the method for choosing an hypothesis testing threshold, and finally in subsection B.5 the unified finite sample algorithm.

B.1 Pairwise Hypothesis Testing

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

$$H_0 : \widehat{x}_i^{(p)}(n) = \sum_{l=1}^{p} b_{ii}(l) x_i(n-l), \tag{B 1}$$

$$H_1 : \widehat{x}_{i|j}^{(p)}(n) = \sum_{l=1}^{p} b_{ii}(l) x_i(n-l) + \sum_{l=1}^{p} b_{ij}(l) x_j(n-l). \tag{B 2}$$

We formulate the statistic

$$F_{ij}(p) = \frac{T}{p}\Big(\frac{\xi_i(p)}{\xi_{ij}(p)} - 1\Big), \tag{B 3}$$

where $\xi_i(p)$ is the sample mean square of the residuals[4] $x_i(n) - \widehat{x}_i^{(p)}(n)$,

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

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.):

$$F_{ij}(p) \to \begin{cases} 0; \; x_j \overset{\text{PW}}{\nrightarrow} x_i \\ \infty; \; x_j \overset{\text{PW}}{\to} x_i \end{cases} \quad \text{as } T \to \infty. \tag{B 4}$$

In our finite sample implementation (see Algorithm 2) we add edges to $\widehat{\mathscr{G}}$ in order of the decreasing magnitude of F_{ij} instead of proceeding backwards through P_{k-r} in Algorithm

[4] This quantity is often denoted $\widehat{\sigma}$, but we maintain notation from Definition 2.2.

1. This makes greater use of the information provided by the test statistic F_{ij}, moreover, if $x_i \overset{\text{GC}}{\to} x_j$ and $x_j \overset{\text{GC}}{\to} x_k$, it is expected that $F_{kj} > F_{ki}$, thereby providing the same effect as proceeding backwards through P_{k-r}.

B.2 Model Order Selection

There are a variety of methods to choose the filter order p (see *e.g.,* [46]), 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:

$$\begin{aligned} BIC_{\text{univariate}}(p) &= \ln\ \xi_i(p) + p\frac{\ln T}{T}, \\ BIC_{\text{bivariate}}(p) &= \ln\ \det\widehat{\Sigma}_{ij}(p) + 4p\frac{\ln T}{T}, \end{aligned} \tag{B 5}$$

where $\widehat{\Sigma}_{ij}(p)$ is the 2×2 residual covariance matrix for the $\mathsf{VAR}(p)$ model of $(x_i(n), x_j(n))$. 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).

B.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:[5]

$$\widehat{R}_x(m) = \frac{1}{T}\sum_{t=m+1}^{T} x(n)x(n-m)^{\mathsf{T}};\ m = 0,\ldots,p_{\max}, \tag{B 6}$$

It is imperative that the first index in the summation is $m+1$, 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 $\widehat{R}_x(m)$ 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.

[5] The particular indexing and normalization given in Equation (B 6) 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.

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 $\widehat{R}_{ii}(m)$, which effectively estimates a sequence of *AR* models, producing $\xi_i(p)$ as a side-effect (see [35] and [24]).

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[6] [69].

B.4 Edge Probabilities and Error Rate Controls

Denote F_{ij} the Granger causality statistic of Equation (B 3) with model orders chosen by the methods of Section B.2. We assume that this statistic is asymptotically $\chi^2(p_{ij})$ distributed (the expected result if the disturbances are Gaussian), and denote by G the cumulative distribution function thereof. We will define the matrix

$$P_{ij} = G(F_{ij}), \tag{B 7}$$

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 (in which case we will conclude that $x_j \overset{\text{PW}}{\to} x_i$) 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 [10] [67] to control the false discovery rate of pairwise edges to a level α (our experiments use the conventional value $\alpha = 0.05$).

B.5 Finite Sample Recovery Algorithm

After the graph topology $\widehat{\mathscr{G}}$ 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 stacking models. That is, performing sequential estimates $\widehat{x}_1(n),\widehat{x}_2(n),\ldots$ estimating a strongly causal graph with the residuals of the previous model as input, and then refitting on the combined sparsity pattern.

C Simulation

We have implemented our empirical experiments in Python [40], in particular we leverage the LASSO implementation from `sklearn` [58] and the random graph generators from `networkx` [29]. Our first experiments use two6 separate graph topologies having $N = 50$ nodes. These are generated respectively by drawing a random tree and a random Erdos Renyi graph then creating a directed graph by directing edges from lower numbered nodes to higher numbered nodes.

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

Algorithm 2: Finite Sample Pairwise Graph Recovery (PWGC)

input : Estimates of pairwise Granger causality statistics F_{ij} (eqn. B 3). Matrix of edge probabilities P_{ij} (eqn. B 7). Hypothesis testing threshold δ chosen via the Benjamini-Hochberg criterion (Section B.4)

output : A strongly causal graph $\widehat{\mathscr{G}}$

initialize: $S = [N]$ `# unprocessed nodes`
$E = \emptyset$ `# edges of` $\widehat{\mathscr{G}}$
$k = 1$ `# a counter used only for notation`

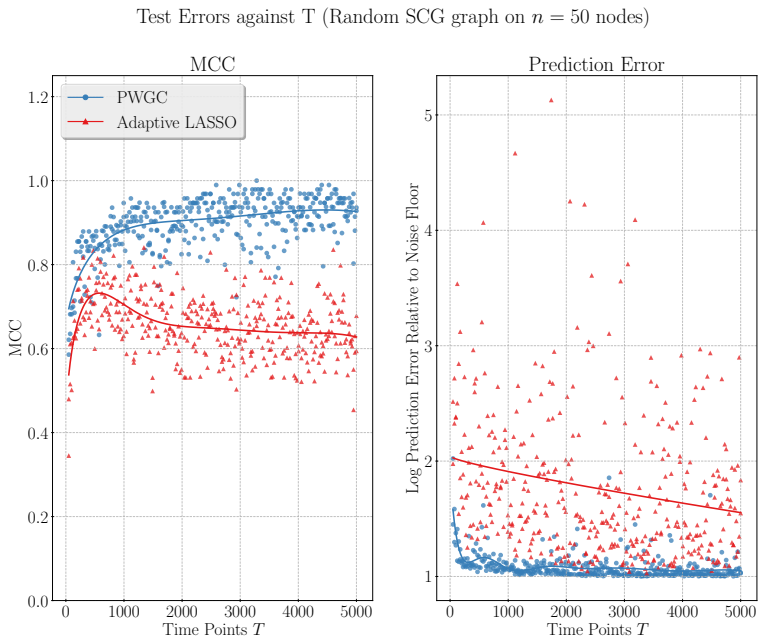
$W_\delta \leftarrow \{(i,j) \mid P_{ji} > 1-\delta, F_{ji} > F_{ij}\}$ `# candidate edges`
$\mathscr{I}_0 \leftarrow \left(\sum_{j \in S:(j,i)\in W_\delta} P_{ij}, \text{ for } i \in S\right)$ `# total node incident probability`
$P_0 \leftarrow \{i \in S \mid \mathscr{I}_0(i) < \lceil \min(\mathscr{I}_0) \rceil\}$ `# Nodes with fewest incident edges`
if $P_0 = \emptyset$ **then**
  $P_0 \leftarrow \{i \in S \mid \mathscr{I}_0(i) \leq \lceil \min(\mathscr{I}_0) \rceil\}$ `# Ensure non-empty`
while $S \neq \emptyset$ **do**
  $S \leftarrow S \setminus P_{k-1}$ `# remove processed nodes`
  $\mathscr{I}_k \leftarrow \left(\sum_{j \in S:(j,i)\in W_\delta} P_{ij}, \text{ for } i \in S\right)$
  $P_k \leftarrow \{i \in S \mid \mathscr{I}_k(i) < \lceil \min(\mathscr{I}_k) \rceil\}$
  if $P_k = \emptyset$ **then**
    $P_k \leftarrow \{i \in S \mid \mathscr{I}_k(i) \leq \lceil \min(\mathscr{I}_k) \rceil\}$
  `# add strongest edges, maintaining strong causality`
  $U_k \leftarrow \bigcup_{r=1}^{k} P_{k-r}$ `# Include all forward edges`
  for $(i,j) \in \text{sort}\left(\{(i,j) \in U_k \times P_k \mid (i,j) \in W_\delta\} \text{ by descending } F_{ji}\right)$ **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 $([N], E)$

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 $r_i \sim \exp(1/2)$. To ensure we are generating data from a stationary system, we first discard samples during a long burn in period.

For both PWGC and adaLASSO we set the maximum lag length $p_{\max} = 10$. Results are collected in Figures C 1, C 2, **??**, C 3.

In reference to Figure C 1 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

Fig. C 1: PWGC Compared Against AdaLASSO [74] (SCG)



Comparison of PWGC and LASSO for 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 C 1 are for systems guaranteed to satisfy the assumptions required for Theorem 2.2.

the theory from which our heuristic derives is valid. 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, in some appropriate sense.

Figure C 3 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 7), as this is the regime many applications must contend with.

In regards to computational 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 cores or multiple machines. Figure 6 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.

Fig. C 2: PWGC vs adaLASSO (DAG, $q = \frac{2}{N}$)

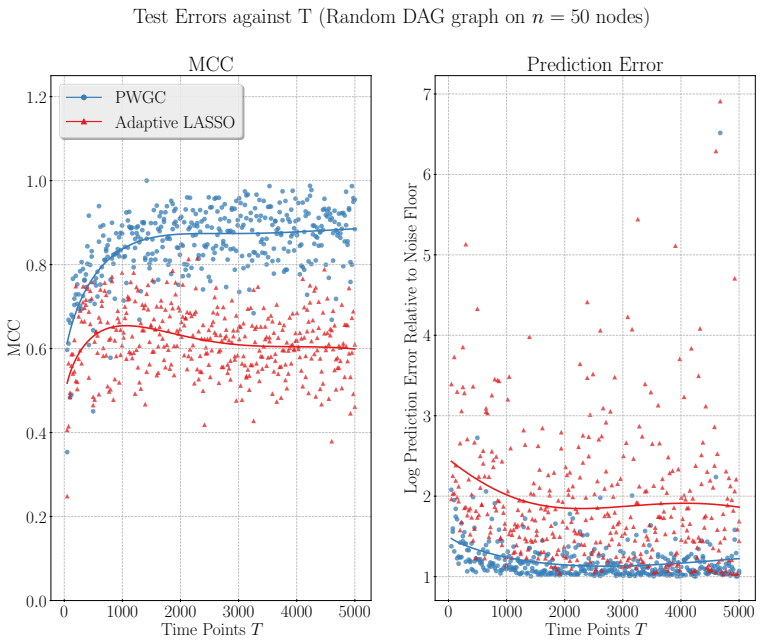


Figure C 2 provides results for systems which do not guarantee the assumptions of Theorem 2.2, though the graph has a similar level of sparsity.

D Proofs

This appendix is devoted to proofs of some of the intermediate results in the paper as well as a proof of the man result.

We begin with some basic lemmas in Section D.1. Section D.2 proves some of the more interesting intermediate results about pairwise causation and finally, Section D.3 establishes the main theorem. An analysis of Example 2.3, which suggests that persistent systems are likely ubiquitous in practice, is provided in Section D.4.

D.1 Lemmas and Preparation

This section states a number of lemmas that are used in later proofs, but which either did not fit into the main body of the paper.

Lemma D.1 (Adjacency Matrix Powers)**.** *Let S be the transposed adjacency matrix[7] of the Granger causality graph $\mathscr{G}$. 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 $j \notin \mathscr{A}(i)$.*

[7] 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.

Fig. C 3: Fixed T, N, increasing edges q (DAG)

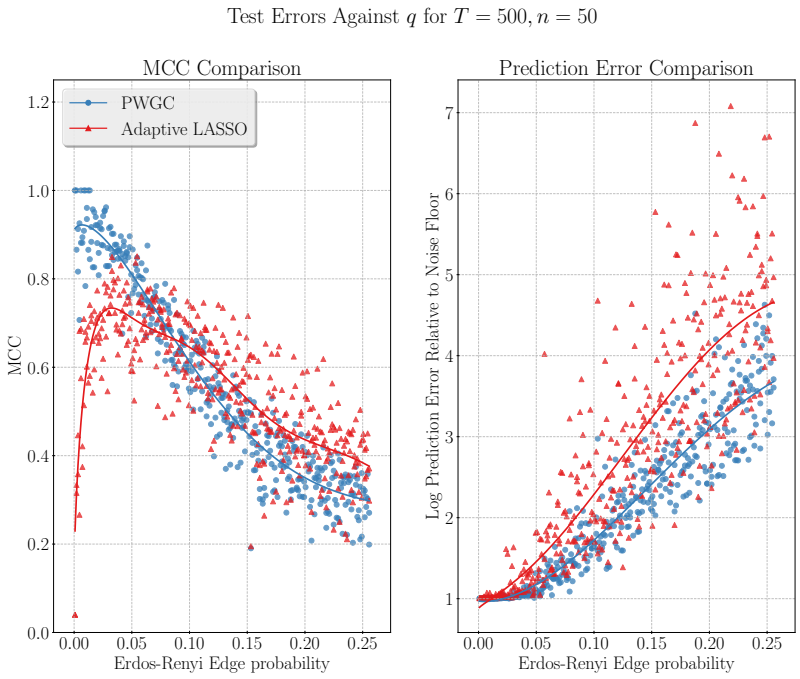


Figure C 3 provides a comparison between PWGC and AdaLASSO as the density of graph edges (as measured by q) increases. For reference, $\frac{2}{N} = 0.04$ has approximately the same level of sparsity as the SCGs we simulated. As q increases, the AdaLASSO outperforms PWGC as measured by the MCC. However, 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.

Proof
This is a well known theorem, proof follows by induction.

Lemma D.2. *Consider distinct nodes i, j in a Granger causality graph $\mathscr{G}$. If $j \notin \mathscr{A}(i)$, then $\mathscr{H}_n^{(v_j)} \perp \mathscr{H}_n^{(i)}$, and therefore for any causal filter $\Phi(z)$ we have*

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

Proof
Fix $m, l \geq 0$, then by expanding with Equation (8)

$$\mathbb{E}[x_i(n-m)v_j(n-l)] = \mathbb{E}[\big(\mathsf{A}_{ii}(z)v_i(n) + \sum_{k\in\mathscr{A}(i)} \mathsf{A}_{ik}(z)v_k(n)\big)v_j(n-l)]$$
$$= 0.$$

This follows since $i \neq j$ and $j \notin \mathscr{A}(i)$ and $v(n)$ is isotrophic and uncorrelated. □

Remark D.1. In order to prove Proposition 2.4 we require some additional notation, as well as another representation theorem. The difficulty addressed by the following Definition D.1 and Lemma D.3 is that in the representation of $x_j(t)$ in terms of its parents (i.e., Equation (7))

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

the filter $\mathsf{B}_{ii}(z)$ need not be stable. That is, the inverse filter $(1-\mathsf{B}_{ii}(z))^{-1}$ need not exist. An example of this issue is furnished by

$$\mathsf{B}(z) = \begin{bmatrix} \rho & -a \\ a & 0 \end{bmatrix} z^{-1},$$

for which, depending on the value of a, may still be stable even if $|\rho| > 1$. This implies that it is not always possible to represent $x_i(n)$ in terms of $v_i(n)$ and $x_k(n), k \in pa(i)$ alone, that is, as

$$x_i(n) = (1-\mathsf{B}_{ii}(z))^{-1}\big(v_i(n) + \sum_{k\in pa(i)} \mathsf{B}_{ik}(z)x_k(n)\big).$$

The difficulty presented by the non-existence of such a representation may become apparent upon studying the proof of Proposition 2.4.

Definition D.1 (Strongly Connected Components)**.** In a graph $\mathscr{G}$, the *ordered* (by the natural ordering on $\mathbb{N}$) subset $S \subseteq [N]$ is *strongly connected* if $\forall i, j \in S:\ i \in \mathscr{A}(j), j \in \mathscr{A}(i)$. We will denote by $S(j)$ (which may be the singleton (j)) the largest strongly connected component (SCC) containing j. We will denote $x_{S(j)}(n)$ to be the ordered vector of processes

$$x_{S(j)}(n) = \big(x_s(n) \mid s \in S(j)\big),$$

whose indices are given the same (natural) ordering as $S(j)$. Similarly, the sub-filter of $\mathsf{B}(z)$ acting on $x_{S(j)}(n)$ will be denoted $\mathsf{B}_{S(j)}(z)$.

Lemma D.3 (Expansion in SCCs)**.** *Given some $j \in [N]$, the process $x_{S(j)}(n)$ can be represented by*

$$x_{S(j)}(n) = v_{S(j)}(n) + \mathsf{B}_{S(j)}(z)x_{S(j)}(n) + \sum_{\substack{s\in S(j)\\ k\in pa(s)\cap S(j)^{\mathsf{c}}}} B_{sk}(z)x_k(n)e_s^{S(j)}, \qquad \text{(D 1)}$$

where $e_s^{S(j)}$ denotes the length $|S(j)|$ canonical basis vector with a 1 in the component corresponding to x_s in the vector $x_{S(j)}$, and the summation is a double sum on s and k.

Moreover, the filter $\mathsf{B}(z)$ is stable with $I - \mathsf{B}_{S(j)}(z)$ invertible:

$$(I - \mathsf{B}_{S(j)}(z))^{-1} = \sum_{k=0}^{\infty} \mathsf{B}_{S(j)}(z)^k, \qquad \text{(D 2)}$$

therefore

$$x_{S(j)}(n) = (I - \mathsf{B}_{S(j)}(z))^{-1}\big(v_{S(j)}(n) + \sum_{\substack{s\in S(j)\\ k\in pa(s)\cap S(j)^c}} B_{sk}(z)x_k(n)e_s^{S(j)}\big). \tag{D 3}$$

Proof

The representation follows directly from the VAR representation of $x(n)$ (i.e., Equation (3))

$$x(n) = \mathsf{B}(z)x(n) + v(n),$$

which, when rearranged appropriately, can be written as

$$\begin{bmatrix} x_{S(j)}(n) \\ x_{S(j)^c}(n) \end{bmatrix} = \mathbf{B}_{S(j)}(z) \begin{bmatrix} x_{S(j)}(n) \\ x_{S(j)^c}(n) \end{bmatrix} + \begin{bmatrix} v_{S(j)}(n) \\ v_{S(j)^c}(n) \end{bmatrix}..$$

where:

$$\mathbf{B}_{S(j)}(z) = \begin{bmatrix} \mathsf{B}_{S(j)}(z) & \mathsf{B}_{S(j),S(j)^c}(z) \\ \mathsf{B}_{S(j)^c,S(j)}(z) & \mathsf{B}_{S(j)^c}(z) \end{bmatrix}$$

Theorem 2.1 is invoked in order to restrict the summation to $k \in pa(s)$ (since other elements are 0).

Now, we can partition $\mathscr{G}$ into its maximal SCCs $S_1,\ldots,S_K$, (one of which is $S(j)$) and then consider the DAG formed on N nodes with edges $I \to J$ included on the condition that $\exists j \in S_J, i \in S_I$ s.t. $i \in \mathscr{A}(j)$. By topologically sorting this DAG, we obtain an ordering σ of $[n]$ such that $\mathsf{B}_\sigma(z)$ is block upper triangluar, with one of its diagonal blocks consisting of the (possibly reordered) matrix $\mathsf{B}_{S(j)}(z)$. So we have

$$\begin{aligned} \forall\, |z^{-1}| \le 1 : \det \mathsf{B}(z) &= \prod_{i=1}^{N} \det \mathsf{B}_{S_i}(z) \neq 0 \\ &\implies \forall\, |z^{-1}| \le 1 : \det \mathsf{B}_{S(j)}(z) \neq 0, \end{aligned}$$

and therefore $\mathsf{B}_{S(j)}(z)$ is stable, invertible, and Equation (D 2) holds. □

Lemma D.4 (Time Lag Cancellation)**.** *Suppose $v(n)$ is a scalar process with unit variance and zero autocorrelation and let $\mathsf{A}(z),\mathsf{B}(z)$ be nonzero and strictly causal (i.e., $1 \le m_0(\mathsf{A}) < \infty$, $1 \le m_0(\mathsf{B}) < \infty$) linear filters. Then,*

$$\langle F(z)\mathsf{A}(z)v(n), \mathsf{B}(z)v(n)\rangle = 0 \;\forall \textit{ strictly causal filters } F(z) \tag{D 4}$$

if and only if $m_0(\mathsf{A}) \ge m_\infty(\mathsf{B})$.

Proof

We have

$$\langle \mathsf{A}(z)v(n), \mathsf{B}(z)v(n)\rangle = \sum_{l=1}^{\infty}\sum_{m=1}^{\infty} a(l)b(m)\mathbb{E}[v(n-m)v(n-l)] \tag{D 5}$$

$$= \sum_{l=\max(m_0(\mathsf{A}),m_0(\mathsf{B}))}^{\min(m_\infty(\mathsf{A}),m_\infty(\mathsf{B}))} a(l)b(l), \tag{D 6}$$

$$\tag{D 7}$$

since $\mathbb{E}[v(n-m)v(n-l)] = \delta_{m-l}$. This expression is 0 if and only if $m_0(\mathsf{A}) \geq 1 + m_\infty(\mathsf{B})$, or $m_0(\mathsf{B}) \geq 1 + m_\infty(\mathsf{A})$, or the coefficients are orthogonal along the common support.

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

$$\begin{aligned}\forall F\ m_0(F\mathsf{A}) \geq 1 + m_\infty(\mathsf{B}) &\overset{(a)}{\iff} m_0(\mathsf{A}) \geq 1 + m_\infty(\mathsf{B}) - \min_F m_0(F) \\ &\overset{(b)}{\iff} m_0(\mathsf{A}) \geq m_\infty(\mathsf{B}),\end{aligned}$$

where (a) follows since $m_0(F\mathsf{A}) = m_0(F) + m_0(\mathsf{A})$, and (b) since $\min_F m_0(F) = 1$. □

Corollary D.1. *For $k \in \mathscr{A}(i) \cap \mathscr{A}(j)$, for all strictly causal $F(z)$ we have,*

$$\begin{aligned}&\hat{\mathbb{E}}[F(z)\mathsf{A}_{jk}(z)v_k(n) \mid \mathscr{H}^{(i)}_{n-1}] = 0 \\ &\iff \langle F(z)\mathsf{A}_{jk}(z)v_k(n), \mathsf{A}_{ik}(z)v_k(n)\rangle = 0 \\ &\iff m_0(\mathsf{A}_{jk}) \geq m_\infty(\mathsf{A}_{ik})\end{aligned}$$

Proof

For the first equivalence we have for all causal $F(z)$,

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

which can be expanded by Equation (8) to obtain (after cancelling all ancestors of i other than k)

$$\langle F(z)A_{jk}(z)v_k(n), \mathsf{A}_{ik}(z)v_k(n-l)\rangle = 0\ \forall l \geq 1,$$

which by the Lemma is equivalent to $m_0(\mathsf{A}_{jk}) \geq m_\infty(\mathsf{A}_{ik})$ as stated.

The final equivalence follows immediately from Lemma D.4.

D.2 Basic Results

This section restates and proves all of the basic results concerning the properties of pairwise Granger causality and the representations of $x_i(t)$ in terms of $\mathscr{A}(i)$ and $pa(i)$. These results are fundamental in proving the main theorem, and may be of some independent interest.

Proposition D.1 (Fully Unconnected Nodes Proposition 2.2)**.** *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* $j \overset{PW}{\nrightarrow} i$ *and* $\hat{\mathbb{E}}[x_j(n) \mid \mathscr{H}_n^{(i)}] = 0$*.*

Proof
We show directly that $\forall l, m \in \mathbb{Z}_+ \; \mathbb{E}[x_i(n-l)x_j(n-m)] = 0$. To this end, fix $l, m \geq 0$, then by expanding with Equation (8) we have

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

Keeping in mind that $v(n)$ 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 $j \notin \mathscr{A}(i)$ and $i \notin \mathscr{A}(j)$ and finally the fourth since $\mathscr{A}(i) \cap \mathscr{A}(j) = \emptyset$. □

Proposition D.2 (Not an Ancestor, No Common Cause; Proposition 2.3)**.** *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 $j \overset{PW}{\nrightarrow} i$*.*

Proof
By Theorem 2.1 it suffices to show that

$$\forall \psi \in \mathscr{H}_{n-1}^{(j)} \; \langle x_i(n) - \hat{\mathbb{E}}[x_i(n) \mid \mathscr{H}_{n-1}^{(i)}], \psi - \hat{\mathbb{E}}[\psi \mid \mathscr{H}_{n-1}^{(i)}] \rangle = 0.$$

which by the orthogonality principle and by representing $\psi \in \mathscr{H}_{n-1}^{(j)}$ via the action of some strictly causal filter $\Phi(z)$ on $x_j(n)$ is equivalent to

$$\langle x_i(n), \Phi(z)x_j(n) - \hat{\mathbb{E}}[\Phi(z)x_j(n) \mid \mathscr{H}_{t-1}^{(i)}] \rangle = 0. \tag{D 8}$$

If we expand $x_j(n)$ using Equation (8), the left hand side of (D 8) becomes

$$\langle x_i(n), \sum_{k \in \mathscr{A}(j) \cup \{j\}} \Big(\Phi(z)\mathsf{A}_{jk}(z)v_k(n) - \hat{\mathbb{E}}[\Phi(z)\mathsf{A}_{jk}(z)v_k(n) \mid \mathscr{H}_{n-1}^{(i)}] \Big) \rangle.$$

We see that this is 0 by Lemma D.2 since $j \notin \mathscr{A}(i)$, and

$$\mathscr{A}(i) \cap \mathscr{A}(j) = \emptyset \implies \forall k \in \mathscr{A}(j) : k \notin \mathscr{A}(i).$$

□

Proposition D.3 (Pairwise Causation and Confounders; Proposition 2.7)**.** *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 $j \overset{PW}{\to} i$, and if $T_{ij}(z)$ is not anti-causal we have $i \overset{PW}{\to} j$. Moreover, if Assumption 2.1 is satisfied, then $j \overset{PW}{\to} i \iff i \overset{PW}{\to} j$.*

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 $H_i(z)x_i(n) = \hat{\mathbb{E}}[x_i(n) \mid H_{t-1}^{(i)}]$, and (c) follows by performing cancellations 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 $j \in \mathscr{A}(i)$.

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

$$(j \overset{\text{PW}}{\to} i) : \exists F(z) \text{ s.t. } \sum_{k \in \mathscr{A}(i) \cap \mathscr{A}(j)} \langle F(z)\mathsf{A}_{jk}(z)v_k(n), (1 - H_i(z))\mathsf{A}_{ik}(z)v_k(n) \rangle \neq 0, \tag{D 9}$$

$$(i \overset{\text{PW}}{\to} j) : \exists G(z) \text{ s.t. } \sum_{k \in \mathscr{A}(i) \cap \mathscr{A}(j)} \langle G(z)\mathsf{A}_{ik}(z)v_k(n), (1 - H_j(z))\mathsf{A}_{jk}(z)v_k(n) \rangle \neq 0. \tag{D 10}$$

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 cancellation 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

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

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:

$$\sum_{k\in\mathscr{A}(i)\cap\mathscr{A}(j)} \langle \mathsf{A}_{ik}(z^{-1})(1-H_i(z^{-1}))F(z)(1-H_j(z))\mathsf{A}_{jk}(z)v_k(n), v_k(n)\rangle, \tag{D 11}$$

$$\sum_{k\in\mathscr{A}(i)\cap\mathscr{A}(j)} \langle \mathsf{A}_{ik}(z^{-1})(1-H_i(z^{-1}))G(z^{-1})(1-H_j(z))\mathsf{A}_{jk}(z)v_k(n), v_k(n)\rangle. \tag{D 12}$$

Recall the filter from Assumption 2.1

$$T_{ij}(z) = \sum_{k\in\mathscr{A}(i)\cap\mathscr{A}(j)} \sigma_k^2 \mathsf{A}_{ik}(z^{-1})(1-H_i(z^{-1}))(1-H_j(z))\mathsf{A}_{jk}(z). \tag{D 13}$$

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 (D 11) is non-zero and therefore $j \overset{\text{PW}}{\to} i$.

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 D 12 is non-zero and therefore $i \overset{\text{PW}}{\to} j$. □

D.3 The Main Theorem

Theorem D.1 (Pairwise Recovery; Theorem 2.2)**.** *If the Granger causality graph $\mathscr{G}$ for the process $x(n)$ is a strongly causal DAG and Assumption 2.1 holds, then $\mathscr{G}$ can be inferred from pairwise causality tests. The procedure can be carried out, assuming we have an oracle for pairwise causality, via Algorithm (1).*

The proof of this main result will proceed in 5 steps which we state formally as lemmas. The approach is to prove the correctness of Algorithm 1. Firstly, we characterize the sets W and P_k appearing in Algorithm. 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 D.5 (W Represents Ancestor Relations)**.** *In Algorithm 1 we have $(i,j) \in W$ if and only if $i \in \mathscr{A}(j)$. In particular, $W \subseteq \mathscr{E}$.*

Proof
Let $j \in [n]$ and suppose that $i \in \mathscr{A}(j)$. Then $i \overset{\text{PW}}{\to} j$ by Proposition 2.6. Proposition 2.5 ensures that (i,j) are not confounded and Corollary 2.2 that $j \notin \mathscr{A}(i)$ so $j \overset{\text{PW}}{\not\to} i$ by Proposition and therefore 2.4 $(i,j) \in W$.

Conversely, suppose $(i,j) \in W$. Then since $j \overset{\text{PW}}{\not\to} i$, Proposition 2.7 ensures that (j,i) are not confounded and so by Proposition 2.4 we must have $i \in \mathscr{A}(j)$. □

Definition D.2 (Depth)**.** For our present purposes we will define the *depth* $d(j)$ of a node j in $\mathscr{G}$ 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 4 we have $d(3) = 1$ and $d(4) = 2$.

Lemma D.6 (Depth Characterization of P_k and S_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 $\mathscr{G}$ is acyclic and therefore $\mathscr{G}$ 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 D.5) 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 $i \in S_k = S_{k-1} \setminus P_{k-1}$. 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 $i \overset{\text{PW}}{\not\to} s$ so that by Proposition 2.7 there cannot be a confounder of (s,i) (otherwise $i \overset{\text{PW}}{\to} s$) so then by Proposition 2.4 we have $s \in \mathscr{A}(i)$. We have shown that $s \in S_k \implies d(s) \geq k$ and so we must have $d(i) > k$, a contradiction, therefore there is no such $s \in S_k$ so $i \in P_k$.

($d(j) \geq k \implies j \in S_k$): 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 D.7 (Inner Loop)**.** *Fix an integer $k \geq 1$ and suppose that $(i,j) \in E_{k-1}$ if and only if $(i,j) \in \mathscr{E}$ and $d(j) \leq k-1$. Then, we have $(i,j) \in D_{kr}$ if and only if $(i,j) \in \mathscr{E}$, $d(j) = k$, and $d(i) = k-r$.*

Proof
We prove by induction on r, keeping in mind the results of Lemmas D.5 and D.6. For the base case, let $r = 1$ and suppose that $(i, j) \in \mathscr{E}$ with $d(j) = k$ and $d(i) = k - 1$. Then, by Corollary 2.4 $(i, j) \in W$ 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 $(i, j) \in W \implies i \in \mathscr{A}(j)$ implies that $i \in pa(j)$ and $(i, j) \in \mathscr{E}$.

Now, fix $r > 1$ and suppose that the result holds up to $r - 1$. Let $(i, j) \in \mathscr{E}$ 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 $E_{k-1} \cup \left(\bigcup_{\ell=0}^{r-1} D_{kr}\right)$, therefore $(i, j) \in D_{kr}$. Conversely, suppose $(i, j) \in D_{kr}$. Then we have $d(i) = k - r$, $d(j) = k$, and $i \in \mathscr{A}(j)$. Suppose by way of contradiction that $i \notin pa(j)$, then there must be some $u \in pa(j)$ such that $i \in \mathscr{A}(u)$. 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 $E_{k-1} \cup \left(\bigcup_{\ell=0}^{r-1} D_{kr}\right)$, which is a contradiction since we assumed $(i, j) \in D_{kr}$. Therefore $i \in pa(j)$ and $(i, j) \in \mathscr{E}$. □

Lemma D.8 (Outer Loop)**.** *We have $(i, j) \in E_k$ if and only if $(i, j) \in \mathscr{E}$ and $d(j) \leq k$. That is, at iteration k, E_k and $\mathscr{E}$ 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 $(i, j) \in E_k = E_{k-1} \cup \left(\bigcup_{r=1}^{k} D_{rk}\right)$. Then clearly there is some $1 \leq r \leq k$ such that $(i, j) \in D_{kr}$ so that by Lemma D.7 we have $(i, j) \in \mathscr{E}$ and $d(j) = k$.

Conversely, suppose that $(i, j) \in \mathscr{E}$ 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 D.7 $(i, j) \in \bigcup_{r=1}^{k} D_{kr}$ which implies that $(i, j) \in E_k$. □

Lemma D.9 (Finite Termination)**.** *Algorithm 1 terminates and returns the set $E_{k^\star - 1} = \mathscr{E}$ 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 D.8 ensures that E_k coincides with $\{(i, j) \in \mathscr{E} \mid d(j) \leq k\}$ and since $d(j) \leq n - 1$ for any $j \in [n]$ there is some $k^\star \leq n$ such that $E_{k^\star - 1} = \mathscr{E}$. 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 $E_{k^\star - 1} \neq \mathscr{E}$) and therefore the algorithm terminates. □

D.4 Persistent Systems Example

Example D.1 (Ubiquity of Persistent Systems; Example 2.3)**.** Consider a process $x(n)$ generated by the $\mathsf{VAR}(1)$ model[8] having $\mathsf{B}(z) = Bz^{-1}$. If B is diagonalizable, and has at

[8] 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.

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 $\mathsf{VAR}(p)$ systems which are not persistent are pathological, in the sense that their system matrices have zero measure.

Proof
Pick any $i \in [N], j \in \mathscr{A}(i) \setminus \{i\}$. Then the stability of B allows us to write

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

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

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

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 $m_\infty(\mathsf{A}_{ij}) < \infty$, 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.

Using the supposition that B is diagonalizable (*i.e.,J* is a diagonal matrix) with at least 2 distinct eigenvalues (in this case B is *not* nilpotent), we can then rewrite the above as

$$f(r) := \mathsf{tr}[(J^{rk})(vu^{\mathsf{T}})] = \sum_{v=1}^{N} \lambda_v^{rk} v_v u_v := \sum_{v=1}^{N} \lambda_v^{rk} \beta_v$$

where λ_v denotes the eigenvalues of B and $\beta_v = u_v v_v$. Note that $f(0) = 0$ since $i \neq j$ and 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 point wise 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. □