Causal Inference for Time Series: A Linear Systems Approach

Parikshit Shah and Venkatesh Saligrama

October 2, 2017

Abstract

We propose a new algorithm for causal inference of time-series data. The proposed test builds on ideas and techniques grounded in linear systems theory, specifically the Schur-Takagi extension problem. The test requires the solution of a pair of convex optimization problems, one corresponding to the causal direction and the other corresponding to the anti-causal direction. Crucially, in each direction the test seeks for the existence of a stable dynamic system with smallest system norm that explains the input-output data. The optimization problem, relies only on the input and output time series data, and requires a single parameter to account for the noise-level. We provide rigorous guarantees to characterize the performance of the test. We validate our approach with numerical experiments on both synthetic and real data, and compare the performance of our method with competing methods.

1 Introduction

Given two time-series $u = \{u_t\}$, $y = \{y_t\}$, the *causal-inference* problem concerns determining whether u causes y, or the vice-versa. In this paper, we focus on this question from the viewpoint of *linear dynamics*, i.e. whether a relationship between u and y could have arisen as a consequence of u being an input to some unknown linear-time invariant system G or vice-versa.

While causal inference is a classical problem [7], it has recently gained interest in the machine learning community [16, 8, 20, 13, 17] due to applications in neuroscience, climate models and economics. For instance the connectome problem, i.e. the problem of constructing the inter-connections between neurons in brain tissue based on recordings (such as local field potentials via electrodes or fluoroscopy [1]) involves causal inference as a key primitive.

In this paper, we develop a method for causal inference for time-series grounded in linear systems theory. In this setup, the notion of stability is associated to a system rather than a pair of time series, (a system is said to be causal if future inputs cannot influence past outputs). Furthermore, the question of causality itself is intricately linked to the question of dynamic stability. Stability is a point of emphasis in our approach: while a majority of processes observed in the real world are dynamically stable, most

treatments of causal inference do not explicitly constrain the inference test to seek a stable explanation.

In addition to proposing a stability-aware test, we also seek rigorous guarantees for the proposed approach. Our method provably recovers the correct causality direction when the ground truth system consists of linear dynamics. Moreover, our method is provably robust to the presence of noise. The test itself requires tuning a single parameter (the noise level) and the results are robust to the choice of this parameter within a reasonable range. We also briefly discuss other non-ideal behaviour such as the presence of data corruption and distortion using the idea of integral quadratic constraints. Finally our method requires the solution of convex optimization problems (a semidefinite program) [19] for which numerically sound approaches are well-known.

While the linearity assumption may seem restrictive, and many real-world systems are in fact nonlinear; over small time horizons nonlinear dynamics are often well approximated by linear ones [14]. Our approach is especially well-suited for such circumstances - our test works directly in the time domain (as opposed to the "z-domain") and one can use segments of the time domain data to perform causal inference on these short segments.

The key insight of this paper is a new connection between causal inference and the Schur-Takagi extension problem and its solution due to [3] (see [21, 18] for a more applied perspective). This connection transparently yields a causal inference test that reduces to the solution of two semidefinite programming problems (one that represents the causal direction and the other the anti-causal direction). As a consequence of the theorem, in the causal direction the solution of the optimization problem remains bounded by the system norm. The main contribution of this paper is in showing that in the anti-causal direction the problem is either infeasible or the optimization problem grows unbounded. We also show that the test is robust to the presence of noise, and present sample complexity bounds for causal inference. We validate the theory with numerical examples on synthetic as well as real-world data.

The paper is organized as follows. In Section 2 we set up the causal inference problem in the context of linear systems. We also introduce the extension theorem due to [3], and describe how it leads to a test for causal inference. In Section 3 we describe our main results concerning the validity of the proposed causal inference test along with an analysis of the sample complexity in the absence and presence of noise. We discuss properties of our test, such as data rescaling, connection to signal to noise ratio, and the computational implications. We also show how are method can be adapted to handle other forms of distortion using intergral quadratic constraints. In Section 4 we present the results of some numerical experiments on both synthetic and real world data. We also compare the performance of our method to other approaches.

1.1 Related Work

While an extensive body of work is devoted to the problem of causal inference, the situation remains somewhat unsatisfactory, even in the restricted setup of linear dynamics governing the time-series. For instance, the seminal Granger-causality approach and methods that build upon it such as time-reversal based approaches [8, 20] and information transfer based approaches [15] are principled approaches to the problem but

have two drawbacks: (a) it is sensitive to the time-lag provided as input to the test as a parameter, and (b) do not directly incorporate dynamic stability into their framework. In contrast, our method directly incorporates stability. Also, we make no assumption on the noise distribution itself, but rather on the noise being bounded in magnitude.

In other related work, we mention the recently proposed Spectral Independence Criterion (SIC) [16], which while working in a linear systems setup requires a special separable structure, thus potentially limiting the scope of applicability. Other proposals include using spectral approaches [10, 6] for inferring the *undirected* network structure (in this undirected setup, causal inference is not the objective). These approaches rely on estimating the frequency response based on periodograms (doing so stably and accurately is challenging and data intensive), do not provide sample complexity bounds, and do not directly work in the time domain.

1.2 Causality and Stability

A fundamental class of models that is widely used in engineering and science to describe physical systems is that of finite-dimensional linear time invariant systems [4, 5]. The linearity assumption is widely made in applications and standard models in statistics such as ARMA models are special cases of linear time invariant systems. We present briefly the basics of linear time invariant causal operators (see [4] for a comprehensive background). In linear systems theory, a system (which maps inputs to outputs) is often viewed as an operator (we denote the system by G). The inputs and outputs belong to the family of extended sequences $\ell_{2,e}(\mathbb{Z})$, namely, sequences that are square summable over any finite time interval. We let $P_T(u_t) = u_t \mathbf{1}_{\{t \leq T\}}$ be the truncation operator that projects the sequence u onto the components $t \leq T$.

Definition 1. The operator G is causal if $\forall\, T \in \mathbb{Z}$ we have $P_T y = P_T G P_T u,\ \forall\ u,\ y \in \ell_{2,e}(\mathbb{Z})$.

Intuitively, this notion enforces the fact that the current output is not a function of future input. We let $G : \ell_{2,e}(\mathbb{Z}) \to \ell_{2,e}(\mathbb{Z})$ be a linear operator given by: $y_t = \sum_{\tau \in \mathbb{Z}} g_{t\tau} u_\tau$. The map $G \triangleq [g_{t\tau}]$ is linear time-invariant (LTI) iff it is Toeplitz, i.e., G is constant along the diagonals and so $y_t \triangleq \sum_{\tau \in \mathbb{Z}} g_{t-\tau} u_\tau$. Furthermore, G is an LTI causal operator iff it is a-lower triangular Toeplitz operator [5].

Problem Statement: We assume that we are given two signals $u,\ y$ over a finite interval of time $[0,\ T]$. One of two hypotheses is true: either y is generated (possibly noisily) as an output of an LTI causal system when input u is applied or that u is generated as an output of an LTI causal system when input y is applied. Our task is to identify the correct hypothesis.

The task suggests a solution based on leveraging the structure of LTI causal maps by estimating causal maps $u \to y$ and $y \to u$ and pick the direction which has smallest estimation error. Unfortunately, this is not sufficient because without enforcing stability the question of identifying causality is ill-posed. We motivate this point by means of the following process:

$$y_t = y_{t-1}/2 + u_t - 2u_{t-1},\ t \in \mathbb{Z}$$

This relationship can be equivalently re-written as:

$$u_{t+1} = 2u_t + (y_{t+1} - y_t/2),\ t \in \mathbb{Z} \tag{1}$$

Expressed in operator form, both would yield "lower-triangular" representations as described above. However there is a key distinction: the first yields a dynamically stable description wheres the second is unstable. Which of these directions is then true? It is precisely for this reason that stability is inextricably linked with causality in the linear systems viewpoint. Since most physical phenomena of interest are stable, it is natural to ascribe causality to the *stable* explanation, i.e. $u \to y$, since this is the direction in the above example for which the input-output behavior is stable.

Definition 2. Given signals u_t, y_t, we say that u_t causes y_t if there is a stable causal system G such that $y_t = G(u_t)$.

Identifiability: In some situations the system G is stably-causally invertible, namely, both G and G^{-1} are causal and stable. Our problem statement in this context is ambiguous since there are stable explanations $u \to y$ and $y \to u$. In this context we would need prior knowledge (see for e.g. [16]) to disambiguate between different causal explanations. This could include information such as predominant spectral content or other such information. We will later see how one could incorporate such knowledge in the form of quadratic constraints.

Notation: In this paper we will refer to time-series $u = \{u_1, \ldots, u_N\}$ and $y = \{y_1, \ldots, y_N\}$. We reserve G and H to denote linear systems. Linear systems can be presented in several equivalent ways. For instance, G can be thought of as an operator, or an semi-infinite Toeplitz matrix. We will use the notation $G(\cdot)$ to refer to the linear system as an operator operating on input sequences u. We will use $G(z)$ to denote its z-transform representation. We will use G_N to denote the truncation of the semi-infinite Toeplitz matrix to its $N \times N$ principal sub-matrix. A linear system can also be presented as a recursion of the form (2) below (typically called a state-space representation). We refer the reader to [4] for these basic concepts. Given a square matrix M, we will refer to $\rho(M)$ to be its spectral radius, i.e. the maximum (absolute) eigenvalue of M. For a symmetric matrices, we will use $\preceq$ to refer to the standard semidefinite ordering. For a matrix M, we use $\|M\|$ to refer to its spectral norm (i.e. its largest singular value), and $\|M\|_F$ to its Frobenius norm (i.e. the Euclidean norm over matrices).

2 Problem Setup and Preliminaries

Consider a finite dimensional, unknown, single-input single-output linear dynamical system in "state-space" form:

$$\begin{aligned} x_{t+1} &= Ax_t + Bu_t \\ y_t &= Cx_t + Du_t + w_t \end{aligned} \tag{2}$$

In the above, the signal $u_t \in \mathbb{R}$ is the input, $y_t \in \mathbb{R}$ is the output, $w_t \in \mathbb{R}$ is noise, and $x_t \in \mathbb{R}^s$ is the state of the linear system at time t. It is well-known that systems of the above form generalize models such as ARMA which are well-studied in the causality

literature. The above system also has a transfer function representation of the form via the z-transform [4]: $G(z) = C(zI - A)^{-1}B + D$. In this paper we will assume that $D \neq 0$. (Indeed the case $D = 0$ is much easier, as we will argue later). An alternative representation of the above is to express $G(z)$ as: $G(z) = \sum_{i=0}^{\infty} g_i z^i$, where the g_i are referred to as the impulse response coefficients. The g_i correspond to the outputs observed when an input of $\delta_0 = \{1, 0, \ldots, \}$ is applied to G.

Every linear dynamical system can be expressed in the above form. For instance, ARMA models, which underpin Granger causality, can be represented as follows. Consider the ARMA model: $y_t = \sum_{i=1}^{p} a_i y_{t-i} + \sum_{i=0}^{p} b_i u_{t-i}$. The above model has the z-domain representation: $G(z) = \frac{\sum_{i=0}^{p} b_i z^i}{1 - \sum_{i=1}^{p} a_i z^i}$, which in turn can be represented in state-space form [5].

$$A = \begin{bmatrix} -a_1 & 1 & 0 & \ldots & 0 \\ -a_2 & 0 & 1 & \ldots & 0 \\ \vdots & & \ddots & & \\ -a_p & 0 & 0 & \ldots & 0 \end{bmatrix}, \quad B = \begin{bmatrix} b_1 - a_1 b_0 \\ b_2 - a_2 b_0 \\ \vdots \\ b_n - a_n b_0 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \end{bmatrix}, \quad D = b_0.$$

While the above system is in a "canonical" form, one can change coordinates to obtain different realization of A, B, C, D for the same system. Indeed, by suitable change of coordinates, it is always possible to transform the system so that A is in Jordan form, and generically, in diagonal form [5, p. 141].

When $D \neq 0$, the inverse system G^{-1} is also well-defined, and has the realization:

$$\begin{aligned} z_{t+1} &= A_{inv} z_t + B_{inv} y_t \\ u_t &= C_{inv} z_t + D_{inv} y_t, \end{aligned} \tag{3}$$

where $A_{inv} = A - BD^{-1}C, B_{inv} = -BD^{-1}, C_{inv} = D^{-1}C, D_{inv} = D^{-1}$.

Since the above system remains unchanged under non-singular coordinate transformation, one can always transform the system so as to put A_{inv} in a convenient canonical form. Throughout this paper, we will assume that A_{inv} is in fact diagonal. Note that this is a slightly restrictive assumption (in general A_{inv} can be put in Jordan form), though diagonalizability generically holds true. This assumption is made only for convenience to make the subsequent analysis more transparent, and in fact the results hold true even in the more general case.

The system described by (2) is stable (i.e. bounded energy inputs lead to bounded energy outputs) if and only if $\rho(A) < 1$. A related notion, called the $\mathcal{H}_\infty$ norm [5] captures the worst case ratio of output energy to input energy (and is indeed the spectral norm analogue for dynamical systems). We denote the norm of a linear system G with $\|G\|_{\mathcal{H}_\infty}$. Note that $\|G\|_{\mathcal{H}_\infty}$ is finite when G is stable and unbounded when G is unstable.

Given time series $u := \{u_0, \ldots, u_N\}$ and $y := \{y_0, \ldots, y_N\}$, we define the corre-

sponding lower triangular Toeplitz matrices as:

$$U_N = \begin{bmatrix} u_0 & 0 & \dots & 0 \\ u_1 & u_0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ u_N & u_{N-1} & \dots & u_0 \end{bmatrix},\ Y_N = \begin{bmatrix} y_0 & 0 & \dots & 0 \\ y_1 & y_0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ y_N & y_{N-1} & \dots & y_0 \end{bmatrix},\ G_N = \begin{bmatrix} g_0 & 0 & \dots & 0 \\ g_1 & g_0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ g_N & g_{N-1} & \dots & g_0 \end{bmatrix}. \tag{4}$$

In the above, $g = \{g_0, \dots, g_N\}$ correspond to the first N coefficients of the impulse response of the system G which satisfy $y = G(u)$, and consequently $Y_N = G_N U_N$. Note that the set of lower-triangular Toeplitz matrices form an algebra (i.e. are closed under additional and matrix multiplication). Lastly, when $u_0 \neq 0$ (and hence U is invertible), then $G_N = Y_N U_N^{-1}$.

A basic result in operator theory relating to Schur-Takagi type extension problems [3] that has been extensively used in the systems and control literature for the problem of model validation [18] is the following:

Theorem 1 (Schur-Takagi-AAK Theorem [21])**.** *Given a sequence $l = \{l_0, l_1, \dots, l_N\}$, there is a unique system $H(z) = \sum_{i=0}^{\infty} h_i z^i$ with coefficients $h_0 = l_0, h_1 = l_1, \dots, h_N = l_N$ with minimal $\mathcal{H}_\infty$ norm. Furthermore the minimal value is given by $\alpha = \sigma_{\max}(L)$, where L is the lower-triangular Toeplitz matrix with $L_{ij} = l_{i-j}$.*

Theorem 1 resolves an *extension* problem in operator theory, i.e. given finite information about an operator G (i.e. the first N impulse response coefficients), to find an extension compatible with the information and indeed, to find the one with minimal norm.

The theorem immediately suggests the following approach to causal inference. Given time series u, y, we form the Toeplitz matrices U_N, Y_N. The system G that explains these observations must have truncated impulse response coefficients determined by $G_N := Y_N U_N^{-1}$. By the Schur-Takagi-AAK theorem, if it is indeed the case that $y = G(u)$ for a stable, causal system G with norm $\|G\|_{\mathcal{H}_\infty}$, then $\sigma_{\max}(G_N) \leq \|G\|_{\mathcal{H}_\infty}$. The main technical contribution in this paper is to show that, when in fact there is no stable system G, $\sigma_{\max}(G_N)$ diverges. Thus, if the ground is $u \to y$, $\sigma_{\max}(G_N)$ remains bounded, whereas $\sigma_{\max}(G_N^{-1})$ diverges (by assumption G does not have a stable inverse). As we will show later, the behaviour is robust to the presence of noise and other distortions.

2.1 Causal Inference Test

Consider the Toeplitz matrices (4) constructed from the sequences u and y respectively. By reformulating $\sigma_{\max}(Y_N U_N^{-1})$ as an optimization problem, Theorem 1 thus suggests the following pair of optimization problems in support of a causality test:

$$\begin{array}{lll} & \underset{\gamma}{\text{minimize}} \quad \gamma & \underset{\gamma}{\text{minimize}} \quad \gamma \\ \text{subject to} & Y_N^T Y_N \preceq \gamma U_N^{'} U_N & \text{subject to} \quad U_N^T U_N \preceq \gamma Y_N^{'} Y_N \end{array} \tag{5}$$

We denote the optimal value of the above to be $\gamma_{UY}(N)$ and $\gamma_{YU}(N)$ respectively. We then test whether $\gamma_{UY}(N) \gtrless \gamma_{YU}(N)$. This test is quite robust (to scaling) in the ideal case. This is because if there exists a stable causal linear system which explains input-output pair u, y, the value $\gamma_{UY}(N)$ does not exceed $\|G\|_{\mathcal{H}_\infty}$ by Theorem 1. When a causal stable inverse does not exist, while the value $\gamma_{UY}(N)$ remains finite, $\gamma_{YU}(N)$ on the other hand will diverge to infinity with N as we will formally show later.

Modeling Uncertainty: Unfortunately, most cases are far from ideal; for instance $u(t)$ and $y(t)$ measurements maybe noisy or nonlinear data distortions may be present. These distortions can include modulation effects, saturation/non-linear effects, incorrect time-stamps as well as other dynamic effects. We can readily handle noise by posing the problem as a semi-definite program. This follows by accounting for noise by recasting singular values in terms of positive definite inequality and invoking the well-known Schur complement trick (see Appendix):

$$\begin{aligned} &\underset{\gamma,\Delta}{\text{argmin}} \quad \gamma \\ &\text{subject to} \begin{bmatrix} I & \Delta \\ \Delta' & \gamma U_N' U_N - Y_N' Y_N - Y_N' \Delta - \Delta' Y_N \end{bmatrix} \succeq 0 \;\; \Delta \text{ Toeplitz, Lower-triangular } \; |\Delta_{ij}| \leq \eta. \end{aligned} \tag{6}$$

(The quantity above quantity is $\gamma_{UY}(N)$ and $\gamma_{YU}(N)$ is analogously defined.) In general we define a confidence measure with regard to the proposed test since we can be more confident in our conclusion if the two values $\gamma_{UY}^N, \gamma_{YU}^N$ are sufficiently far apart.

Definition 3. The proposed causal inference test is α-confident in predicting the causal direction $u \to y$ if $\log(\frac{\gamma_{YU}}{\gamma_{UY}}) \geq \alpha > 0$.

As we collect more data, we want the confidence level α to increase. We will show that precisely this behavior is manifested. The test succeeds in predicting causality in the correct direction with growing confidence when the problem is identifiable. Indeed, one can define a notion of *sample-complexity*; which is the level of confidence achieved from N samples of data.

We note that the uncertainty matrix Δ (constrained to be lower-triangular) can incorporate different sources of uncertainty:

1. Effect of measurement noise: When the measurements are of the form $y = G(x) + w$, where w is unknown (but bounded) measurement noise, its effect can be captured by the uncertainty variable Δ. Indeed, one can verify that $Y_N = \bar{Y}_N + W_N$, where $\bar{Y}_N$ is the nominal output in the absence of noise and W_N is the lower-triangular Toeplitz matrix composed of the noise components. When the noise is bounded, its effect can be captured by the constraint $|\Delta_{ij}| \leq \eta$.

2. Effect of input noise: When input noise is present, it is manifested as $y = G(x+w)$ where w is the unknown noise signal. Again $Y_N = \bar{Y}_N + G_N W_N$. Once again, $G_N W_N$ is a lower triangular Toeplitz matrix, and when the effect of input noise is bounded, the constraint $|\Delta_{ij}| \leq \eta$ captures the uncertainty.

3. Effect of initial conditions: When unknown initial conditions are present, their effect can be viewed as additional dynamics evolving under the effect of an impulse to G at time 0. Again Δ captures the required effect in manner analogous to the previous point.

While the above formulation accounts for noise, it does not account for distortions and causal ambiguities. Handling these issues can be important. For instance, consider Eq. 1 and suppose that the time-stamp for $y(t)$ were incorrect and in reality we had measured a time-series $z(t)$ that is a delayed version of $y(t)$, namely, $z(t) = y(t-k)$. In this situation the causal explanation may fail in both directions. In general there could be other types of distortions such as saturation effects, signal modulation (frequency shift) etc. The question arises as to how to handle these types of effects. In general if we had prior information we could account for them within the SDP framework by employing integral-quadratic-constraints to model these effects.

Data distortion: When the observed data u_t, y_t has undergone a distortion with known properties we may view it as an underlying signal pair $(v(t), z(t))$ such that $u(t), y(t)$ are noisy/distorted versions of $(v(t), z(t))$. While, the pair $(v(t), z(t))$ is ideal, (i.e. we can infer the correct direction if we had access to these measurements), only certain properties of this distortion are available. We can model the relationship between $v(t) \mapsto (t)$ and $z(t) \mapsto y(t)$ by means of integral quadratic constraints (IQC) [11]. Indeed [11] provides a library of various types of scenarios including uncertain or time-varying multiplicative gains, uncertain delays, non-linear distortions etc. We briefly describe them here for our purposes. These constraints are quadratic forms over the signal space and take the form: $\sigma(u, v) = \sum_{t=0}^{N} u(t) P_t u'(t) - \sum_{t=0}^{N} v(t) Q_t v'(t) \geq 0$ where P_t, Q_t are known positive scalars (or PSD matrices in case we have multi-variate time-series) that are shaped to incorporate prior knowledge of the distorting effects. For instance, to incorporate bounded time delay we first let $P_t = Q_t$. We then choose the multiplier P_t such that it has a real-rational bounded Fourier transform [11]. In this way we can introduce a collection of such constraints to incorporate both prior information such as frequency content as well as other distortions. We could also incorporate additive noise, w, in measuring u, with in this setup using two IQCs ($\sigma(u, v - w) > 0$ and $\|w\|_2^2 \leq \eta$. These lead to the following problem:

$$\gamma_{UY}^{N} = \underset{\gamma}{\text{argmin}} \quad \gamma$$
$$\text{subject to} \quad Z_N^{'} Z_N \preceq \gamma V_N^{'} V_N,\ \sigma_1(u, v - w) \geq 0,\ \sigma_2(y, z - n),\ \|w\|_2^2 \leq \eta,\ \|n\|_2^2 \leq \eta$$

Using Schur complementation we can again reduce it to a semi-definite program.

3 Theoretical Guarantees

We now present our main theoretical guarantees. The proofs are available in the supplementary material. We first begin with the noiseless case:

Theorem 2. *Let $u, y \in \mathbb{R}^N$ be generated so that $y = G(u)$ for a stable linear system G of the form* (2)*. Let $\gamma_{UY}(N)$ and $\gamma_{YU}(N)$ be obtained via the solution of* (5)*,* (**??**)

respectively. Then the following hold:

a) If $D = 0$, we have that $\gamma_{UY}(N) \leq \|G\|^2_{\mathcal{H}_\infty}$, $\gamma_{YU}(N) = +\infty$ (i.e. problem is infeasible), and thus unbounded confidence.

b) Suppose $D \neq 0$, and suppose the inverse system G^{-1} defined in (3) *is unstable with $\rho := \rho(A_{inv}) > 1$. Then there is a constant $c > 0$ such that $\gamma_{UY}(N) \leq \|G\|^2_{\mathcal{H}_\infty}$ for all N, $\gamma_{YU}(N) \geq c\frac{\rho^{2N}}{N}$. Consequently the proposed test achieves a sample complexity for α-confidence with $\alpha \sim O(N)$.*

Remarks

In contrast with most causal inference approaches e.g. [7, 16] our result provides quantitative, non-asymptotic bounds required to ascertain the confidence level for our causality test. When working with very limited data, our method is extremely appealing, since it implies a low sample complexity. To achieve a confidence level α, only $O(\alpha)$ samples are required (ignoring logarithmic factors). In terms of computational complexity, the noiseless case requires the solution of a generalized eigenvalue problem (GEVP), a canonical problem in numerical linear algebra for which efficient and scalable approaches are well-studied. We next study the case where measurement noise is present:

Theorem 3. *Let $u, y \in \mathbb{R}^N$ be generated so that $y = G(u) + w$ for a stable linear system G of the form* (2) *where w is some measurement noise. Let η be chosen such that $|w_i| \leq \eta$ for all $i = 0, \ldots, N$. Let $\gamma_{UY}(N)$ and $\gamma_{YU}(N)$ be obtained via the solution of* (9)*,* (10) *respectively. Let $D \neq 0$, and suppose the inverse system G^{-1} defined in* (3) *is unstable with $\rho := \rho(A_{inv}) > 1$. Assume that $\frac{\sigma_{\min}(G_N U_N)}{\sigma_{\max}(W_N)} > 2 + \delta$ for some $\delta > 0$ and $\sigma_{\min}(G_N U_N) \geq C$. Then there are constants $c_1 > 0$, $c_2 > 0$ such that $\gamma_{UY}(N) \leq \|G\|^2_{\mathcal{H}_\infty}$ for all N, $\gamma_{YU}(N) \geq \left(c_0 \frac{\rho^N}{\sqrt{N}} - c_1 \eta N\right)^2$, and hence the proposed test succeeds with $O(N)$ confidence (ignoring logarithmic factors).*

3.1 Discussion and Implementation Details

1. Sample Complexity: The above theorem establishes non-asymptotic bounds for the confidence of the causal inference test. Moreover, our result does not make distributional assumptions, such as Gaussianity, stationarity, etc.

2. Computation: The test requires the solution of convex optimization problems (i.e. semidefinite programs). These are somewhat expensive to solve using off-the-shelf solvers. However there is special structure in these problems which make them amenable to much more efficient computation. Due to space constraints we do not discuss this aspect here.

3. Signal to Noise Ratio: The assumption $\frac{\sigma_{\min}(G_N U_N)}{\sigma_{\max}(W_N)} > 2 + \delta$ may be viewed as a requirement on the *signal-to-noise* ratio, i.e. we require that the signal encoded in the Toeplitz matrix GU be sufficiently larger than the noise level. We also have a requirement on the size of $G_N U_N$ itself, this is to prevent the situation where for large N the spectral norm of $G_N U_N$ to vanish. A parameter that needs to be picked for the test is η, and it needs to be picked so that it acts as an upper bound on the noise. In numerical

experiments our test is seen to be extremely robust to this parameter.

4. Data rescaling: Suppose under the ground-truth that $u \to y$, one were to scale the output y by a factor $\beta > 1$ (so that inherently the underlying system G is scaled by β. Then the $\mathcal{H}_\infty$ norm also gets scaled by β. The value of γ_{UY} then gets scaled up also by α. However the value of γ_{YU} is nevertheless divergent. However, to capture this phenomenon from finite data (i.e. to achieve a fixed level of α-confidence), additional data is now required. In Theorem 3, this is reflected in the fact that $\gamma_{UY}(N)$ increases as the squared of the system norm, thereby reducing α by a factor of β^2. This point is intimately related to the notion of "stability margin" in linear systems theory. Indeed, stable systems with higher norm are considered closer to instability, and more data is required to infer causality consequently.

Nevertheless, we could incorporate scaling for finite data for wide-sense stationary signals $u(t), y(t)$ with good mixing properties. One way of doing so involves forming periodograms, $\Phi_u(\exp(-j\omega)), \Phi_y(\exp(-j\omega))$ of $u(t)$ and $y(t)$ respectively and estimating the gain at zero frequency, i.e., $\hat{\beta}^2 = \frac{\Phi_u(\exp(-j0))}{\Phi_u(\exp(-j0))}$. For the direction $u \to y$ we discount the measured γ_{UY}^N by this value and set $\hat{\gamma}_{UY}^N = \gamma_{UY}^N / \hat{\beta}^2$ and in the reverse direction we amplify it to $\hat{\gamma}_{YU}^N = \hat{\beta}^2 \gamma_{UY}^N$. A scale invariant test would then compare these modified values. For wide-sense stationary processes with sufficiently high mixing rate this ratio can be estimated somewhat accurately (even though the underlying system relating them could be unstable).

4 Comparison to Related Approaches and Numerical Experiments

We first begin with some synthetic numerical examples.

Example 1. We begin with a simple system consisting of a delay, i.e., $y_t = u_{t-1}$, where $u_t \in R^{50}$ is chosen to be a random input distributed i.i.d. normally with zero mean and unit variance. Solving the optimization problems (5) we get the $\gamma_{UY}(50) = 1, \gamma_{YU}(50) = +\infty$.

Continuing with this example, suppose that $y_t = u_{t-1} + w_t$, where w_t is a noise process distributed i.i.d. normally with zero mean and variance 0.25. We solve the optimization problems (9), (10) with $\eta = 0.5$ to obtain $\gamma_{UY}(50) = 0.7322$ (i.e. infeasible) whereas $\gamma_{YU}(50) = +\infty$, thereby obtaining the correct causality direction. We note that the test is stable with respect to the choice of η as well as the number of samples as can be verified by varying η and the number of samples. On the same synthetic data, Granger causality also has a p-value of 0 in the causal direction and $.0525$ in the anti-causal direction when a time-lag of 1 unit is specified. On the same random instance, SIC yields a value of 0.9432 in the causal direction and a value of $.9612$ in the anti-causal direction, thus mis-predicting the causality direction. (Note that the ideal values are both 1 when the periodograms are computed exactly, but due to effect of finite data these values are away from 1).

Example 2. Our next synthetic example involves the following system: $G(z) = \frac{1-az}{z-a}$, where $a \in (0, 1)$. The input is a i.i.d. random normally distributed signal with zero-

mean and unit variance. In Fig. 1, we plot how the causality confidence parameter α increases with the number of samples for $a = 0.9$. Note that for $N = 30$ samples, $\alpha = 556.57$. The Granger-causality test with time lag set to 1 has a p-value of 0.039 in the causal direction and 0.55 in the anti-causal direction (neither is considered statistically significant). The SIC values in the causal and anti-causal direction respectively are 0.7646 and 1.0233 (thereby mis-predicting the causal direction). Fig. 1 also shows how the confidence increases in the presence of signal corrupted by a Gaussian noise with mean zero and variance .03 (and the parameter η set to .1).

4.1 Experiments with Real Data

1. Robotic Arm: We use data from a system identification dataset from the database known as "Daisy" [2]. The specific dataset chosen corresponds to a robotic arm with known input in the form of a voltage signal to a motor and measured output using a sensor [2]. We use this data to validate the proposed causal inference test. Based on $N = 100$ samples, we obtain $\gamma_{UY}(100) = 4.88$, $\gamma_{YU}(50) = 12.17$, thereby yielding a confidence of $\alpha = 2.49$. Granger-causality with a unit time-lag obtains a p-value of 0.0023 in the causal direction and 0.135 in the anti-causal direction. The SIC in the causal direction is 0.691 whereas in the anti-causal direction it is 0.037 thereby mis-predicting the direction.

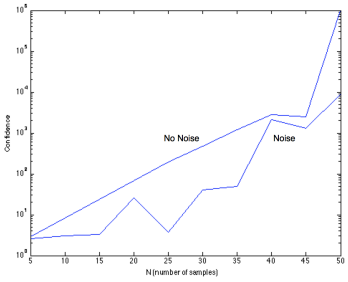

Figure 1: The figure shows the growth of the confidence parameter α with the number of samples in (a) the absence of noise and (b) the presence of noise.

2. Milk and Cheese Prices: We consider 25 years of monthly milk and block cheddar cheese prices available at `http://future.aae.wisc.edu/tab/prices.html`. The ground truth is considered to be that price of milk drives the price of cheese (and indeed other dairy products such as butter). Based on $N = 50$ samples, we obtain $\gamma_{UY}(50) = .003$, $\gamma_{YU}(50) = 77.3$, resulting in a causal prediction of MilkPrice $\rightarrow$ CheesePrice (consistent with the ground-truth) with confidence $\alpha = 25935.1$. For Granger causality, the p-value in the causal direction was found to be .0352 and in the anti-causal direction to be 3.5×10^{-7}, hence Granger causality mis-predicts the direction (here time-lag was set to 1 unit; similar trends were observed for longer time-lags). The SIC has a value of 0.389 in the causal direction and .250 in the anti-causal direction.

3. Connectome Data: We consider data obtained by performing measurements of neural activity in a population of 100 neurons. The data is available at `https://www.kaggle.com/c/connectomics/data`. (We consider data from the "small" dataset. The activity of each neuron is measured by fluorescence [12] and this is represented as a time series. The neural connectivity of this population is known a priori. We take two neurons (neurons 1 and 11) between which there is known to be a

directional connectivity, and consider the time series generate by the fluorescence activity. Based on $N = 60$ samples of these time-series, we obtain $\gamma_{UY}(60) = 35.59$, $\gamma_{YU}(60) = +\infty$ (infeasible). Granger-causality with a unit time-lag obtains a p-value of 0.0057 in the causal direction and 0.36 in the anti-causal direction. The SIC in the causal direction is 0.068 whereas in the anti-causal direction it is 0.479 thereby mis-predicting the direction.

References

[1] ChaLearn Connectomics Challenges. `http://connectomics.chalearn.org/`. Accessed: 2016-02-04.

[2] DaISy: STADIUS' Identification Database. `http://homes.esat.kuleuven.be/~smc/daisy/daisydata.html`. Accessed: 2016-02-04.

[3] Vadim M Adamjan, Damir Z Arov, and MG Krein. Analytic properties of schmidt pairs for a hankel operator and the generalized schur-takagi problem. *Sbornik: Mathematics*, 15(1):31–73, 1971.

[4] F.M. Callier and C.A. Desoer. *Linear System Theory*. Springer Texts in Electrical Engineering. Springer New York, 1994.

[5] Geir E Dullerud and Fernando Paganini. *A course in robust control theory: a convex approach*, volume 36. Springer Science & Business Media, 2013.

[6] J. Etesami and N. Kiyavash. Directed information graphs: A generalization of linear dynamical graphs. In *2014 American Control Conference*, pages 2563–2568, June 2014.

[7] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

[8] Stefan Haufe, Vadim V. Nikulin, and Guido Nolte. Alleviating the influence of weak data asymmetries on granger-causal analyses. In *Proceedings of the 10th International Conference on Latent Variable Analysis and Signal Separation*, LVA/ICA'12, pages 25–33, Berlin, Heidelberg, 2012. Springer-Verlag.

[9] Roger A Horn and Charles R Johnson. *Topics in Matrix Analysis*. Cambridge University Press, New York, NY, USA, 1986.

[10] D. Materassi and M. V. Salapaka. On the problem of reconstructing an unknown topology via locality properties of the wiener filter. *IEEE Transactions on Automatic Control*, 57(7):1765–1777, July 2012.

[11] A. Megretski and A. Rantzer. System analysis via integral quadratic constraints. *Automatic Control, IEEE Transactions on*, 42(6):819–830, Jun 1997.

[12] Yuriy Mishchencko, Joshua T Vogelstein, and Liam Paninski. A bayesian approach for inferring neuronal connectivity from calcium fluorescent imaging data. *The Annals of Applied Statistics*, pages 1229–1261, 2011.

[13] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, pages 154–162, 2013.

[14] Shankar Sastry. *Nonlinear systems: analysis, stability, and control*, volume 10. Springer Science & Business Media, 2013.

[15] Thomas Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000.

[16] Naji Shajarisales, Dominik Janzing, Bernhard Schoelkopf, and Michel Besserve. Telling cause from effect in deterministic linear dynamical systems. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 285–294, 2015.

[17] Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030, December 2006.

[18] R.S. Smith and J.C. Doyle. Model validation: a connection between robust control and identification. *Automatic Control, IEEE Transactions on*, 37(7):942–952, Jul 1992.

[19] Lieven Vandenberghe and Stephen Boyd. Semidefinite programming. *SIAM Rev.*, 38(1):49–95, March 1996.

[20] I. Winkler, D. Panknin, D. Bartz, K.-R. Müller, and S. Haufe. Validity of time reversal for testing Granger causality. *ArXiv e-prints*, September 2015.

[21] Tong Zhou and Hidenori Kimura. Time domain identification for robust control. *Systems and Control Letters*, 20(3):67–178, 1993.

5 Proofs

Lemma 1. *Let M, N be square matrices and let N be invertible. Then the following statements are equivalent:*

1. $\sigma_{max}(MN^{-1}) \leq \gamma$
2. $M^{'}M \preceq \gamma^2 N^{'}N$.

Proof. The proof directly follows from the fact that $\sigma_{max}(MN^{-1}) \leq \gamma$ is equivalent to ${N^{-1}}^{'} M^{'} M N^{-1} \preceq \gamma^2 I$. □

Lemma 2. *Let H be a system of the form* (2) *with $\rho := \rho(A) > 1$. Then there is a constant c, and N_0 such that $\|H_N\| \geq \frac{c}{\sqrt{N}}\rho^N$ for all $N \geq N_0$.*

Proof. We assume that (2) is transformed so that A is in diagonal form. We then have

$$H(z) = C(zI - A)^{-1}B + D = \sum_{i=1}^{r} \frac{c_i}{z - a_{ii}} + D.$$

Without loss of generality assume that a_{11} has maximum modulus, i.e. $|a_{11}| \geq |a_{ii}|$ for all $i = 1, \ldots, r$. Since $\rho(A) > 1$ note that $|a_{11}| > 1$. The impulse response coefficients of this system are then precisely:

$$h_0 = D \qquad h_k = \sum_{i=1}^{r} c_i a_{ii}^k.$$

Since $\sigma_{\max}(H_N)^2 \|x\|^2 \geq x^{'} H_N x$ for all x, we trivially have $\sigma_{\max}(H_N)^2 \geq \frac{1}{N}\|H_N\|_F^2$ by picking $x = \frac{1}{\sqrt{N}}\mathbf{1}$, the vector of all ones.

Note that

$$\begin{aligned}
|h_k|^2 &= \left| \sum_{i=1}^{r} c_i a_{ii}^k \right|^2 \\
&= |a_{11}|^{2k} \left| c_1 + \sum_{i=2}^{r} c_i \left(\frac{a_{ii}}{a_{11}} \right)^k \right|^2 \\
&\geq |a_{11}|^{2k} \left| c_1 - \sum_{i=2}^{r} \left| c_i \left(\frac{a_{ii}}{a_{11}} \right) \right|^k \right|^2
\end{aligned}$$

Hence, there exists an N_0 such that for all $N \geq N_0$ we have:

$$|h_N|^2 \geq \frac{c_1}{2}|a_{11}|^{2N} = \frac{c_1}{2}\rho^{2N}.$$

Noting that $\|H_N\|_F^2 = \sum_{ij} |H_{Nij}|^2 \geq |h_N|^2$, we have the required result. □

Proof of Theorem 2. a) Note that when $D = 0$ we have $u_0 \neq 0$, but $y_0 = 0$. Hence Y_N, which is a lower-triangular Toeplitz matrix has zeroes along the diagonal, and is therefore singular. Since $u_0 \neq 0$, U_N is invertible, on the other hand and by Lemma 1, $Y_N^{'} Y_N \preceq \gamma U_N^{'} U_N$ is equivalent to $\sigma_{\max}(Y_N U_N^{-1}) \leq \sqrt{\gamma}$. However, in the absence of noise, $Y_N U_N^{-1} = G_N$, hence we are seeking a γ such that $\sigma_{\max}(G_N) \leq \sqrt{\gamma}$. By

Theorem 1, an upper bound on $\sqrt{\gamma}$ is simply $\|G\|_{\mathcal{H}_\infty}$. Hence $\gamma_{UY}(N) \leq \|G\|^2_{\mathcal{H}_\infty}$ for all N.

On the other hand, Y_N is singular, hence $Y_N^{'}Y_N$ drops rank whereas $U_N^{'}U_N$ is full rank. Hence the inequality $U_N^{'}U_N \preceq \gamma Y_N^{'}Y_N$ is infeasible, and $\gamma_{YU}(N) = +\infty$.

b) By the same argument as part (a) we see that $\gamma_{UY}(N) \leq \|G\|^2_{\mathcal{H}_\infty}$. When $D \neq 0$, $y_0 = Du_0 \neq 0$. Since Y_N is thus a lower-triangular matrix with non-zero diagonal, Y_N is non-singular. By Lemma 1, we have $U_N^{'}U_N \preceq \gamma Y_N^{'}Y_N$ is equivalent to $\sigma_{\max}(U_N Y_N^{-1}) \leq \sqrt{\gamma}$. However, $U_N Y_N^{-1} = G_N^{-1}$. Observing that G_N^{-1} is a Toeplitz matrix whose entries are the impulse response coefficients of G^{-1} (which exists because $D \neq 0$ by (3)), together with Lemma 2 we have that $\gamma_{YU}(N) \geq c\frac{\rho^{2N}}{N}$.

□

We now proceed to the analysis of the noisy case, i.e. the case where a noise process $w = \{w_0, \ldots, w_N\}$ corrupts the observations u. Consider the case where the noise is non-zero but bounded, i.e. $|w_j| \leq \eta$. Over a time horizon of k it is convenient to write the dynamical system as:

$$Y_N = G_N U_N + W_N,$$

where Y_N, U_N, W_N, G_N are all lower-triangular Toeplitz matrices, and $u_0 \neq 0$. Note that $U, W, G, \bar{Y}$ that the diagonal elements $u_0 \neq 0$, $g_0 \neq 0$ (since $D \neq 0$), and $w_0 \neq 0$. As a consequence of the above matrices being lower triangular with nonzero diagonals, these matrices are easily seen to be invertible.

Lemma 3. *Let $U_N, W_N, G_N, \Delta, Y_N$ be lower triangular invertible Toeplitz matrices as described above. Then we have:*

$$\|(U_N - \Delta)Y_N^{-1}\| \geq \|G_N^{-1}\|\sigma_{\min}\left(I - W_N Y_N^{-1}\right) - \|\Delta Y_N^{-1}\|.$$

Proof. Note that since $Y_N = G_N U_N + W_N$, we have by a simple rearrangement of terms:

$$G_N^{-1}(I - W_N Y_N^{-1}) = U_N Y_N^{-1},$$

and hence,

$$(U_N - \Delta)Y_N^{-1} = G_N^{-1}(I - W_N Y_N^{-1}) - \Delta Y_N^{-1}.$$

We this have:

$$\begin{aligned}\|(U_N - \Delta)Y_N^{-1}\| &\geq \|G_N^{-1}(I - W_N Y_N^{-1})\| - \|\Delta Y_N^{-1}\| \\ &\geq \|G_N^{-1}\|\sigma_{\min}(I - W_N Y_N^{-1}) - \|\Delta Y_N^{-1}\|.\end{aligned}$$

(In the above, we have used the triangle inequality in the first step, and the fact that $\sigma_{\max}(AB) \geq \sigma_{\max}(A)\sigma_{min}(B)$). □

Proof of Theorem 3. Let w be the spcific noise realization, and W_N be the corresponding Toeplitz matrix. By assumption $\eta \geq \max_i w_i$. Hence $\Delta = W_N$ is feasible with respect to (9). For this value of Δ we then have $(Y_N - \Delta)U_N^{-1} = G_N$, and hence $\sigma_{\max}((Y_N - \Delta)U_N^{-1}) \leq \|G\|_{\mathcal{H}_\infty}$. Hence, the pair $\Delta = W_N, \gamma = \|G\|^2_{\mathcal{H}_\infty}$ is feasible with respect (9). Hence $\gamma_{UY}(N) \leq \|G\|^2_{\mathcal{H}_\infty}$.

We now establish that $\gamma_{YU}(N)$ grows exponentially. By Lemma 3 we have

$$\|(U_N - \Delta)Y_N^{-1}\| \geq \|G_N^{-1}\|\sigma_{\min}\left(I - W_N Y_N^{-1}\right) - \|\Delta Y_N^{-1}\|.$$

Note that by assumption G^{-1} is unstable, and hence $\rho(A_{inv}) > 1$. By Lemma 2 we therefore have

$$\|G_N^{-1}\| \geq c\frac{\rho^N}{\sqrt{N}}. \tag{7}$$

We further have

$$\begin{aligned}
\sigma_{\min}\left(I - W_N Y_N^{-1}\right) &\geq 1 - \sigma_{\max}\left(W_N(G_N U_N + W_N)^{-1}\right) \\
&\geq 1 - \frac{\sigma_{\max}(W_N)}{\sigma_{\min}(G_N U_N + W_N)} \\
&\geq 1 - \frac{\sigma_{\max}(W_N)}{\sigma_{\min}(G_N U_N) - \sigma_{\max}(W_N)} \\
&= 1 - \frac{1}{\frac{\sigma_{\min}(G_N U_N)}{\sigma_{\max}(W_N)} - 1}
\end{aligned}$$

where the second and third inequality follow from Weyl's inequalities concerning singular values [9, p. 171]. By assumption,

$$\frac{\sigma_{\min}(G_N U_N)}{\sigma_{\max}(W_N)} > 2 + \delta,$$

so that

$$1 - \frac{1}{\frac{\sigma_{\min}(G_N U_N)}{\sigma_{\min}(W_N)} - 1} > 0.$$

As a consequence of (7), we have

$$\|G_N^{-1}\|\sigma_{\min}\left(I - W_N Y_N^{-1}\right) > c_0\frac{\rho^N}{\sqrt{N}}, \tag{8}$$

for some constant c_0.

Now, to bound the last term we have:

$$\begin{aligned}
\|\Delta Y_N^{-1}\| &\leq \frac{\|\Delta\|}{\sigma_{\min}(GU + W)} \\
&\leq \frac{\|\Delta\|}{\sigma_{\min}(GU) - \sigma_{\max}(W)} \\
&\leq \frac{2\|\Delta\|}{\sigma_{\min}(GU)} \qquad \left(\text{since } \frac{\sigma_{\min}(GU)}{\sigma_{\max}(W)} \geq 2 + \delta\right) \\
&\leq \frac{\sqrt{2}N\eta}{C}.
\end{aligned}$$

Putting together this inequality with (8) and Lemma 3 we have the required result. □

6 Appendix

6.1 Semidefinite Program for the Noisy Case

When uncertainty is present, we account for its presence by altering the optimization problems (5), to the following:

$$\begin{array}{ll} \underset{\gamma,\Delta}{\text{minimize}} & \gamma \\ \text{subject to} & (Y_N - \Delta)^{'}(Y_N - \Delta) \preceq \gamma U_N^{'} U_N \\ & \Delta \text{ Toeplitz, Lower-triangular} \\ & |\Delta_{ij}| \leq \eta. \end{array} \tag{9}$$

$$\begin{array}{ll} \underset{\gamma,\Delta}{\text{minimize}} & \gamma \\ \text{subject to} & (U_N - \Delta)^{'}(U_N - \Delta) \preceq \gamma Y_N^{'} Y_N \\ & \Delta \text{ Toeplitz, Lower-triangular} \\ & |\Delta_{ij}| \leq \eta. \end{array} \tag{10}$$

where η is a parameter that controls the noise-level. As above, we will denote the optimal solutions to be $\gamma_{UY}(N)$ and $\gamma_{YU}(N)$.

Note that by taking Schur complements, problem (9) can be transformed to a semidefinite programming problem as follows (similarly for (10)):

$$\begin{array}{ll} \underset{\gamma,\Delta}{\text{minimize}} & \gamma \\ \text{subject to} & \\ & \begin{bmatrix} I & \Delta \\ \Delta^{'} & \gamma U_N^{'} U_N - Y_N^{'} Y_N - Y_N^{'}\Delta - \Delta^{'} Y_N \end{bmatrix} \succeq 0 \\ & \Delta \text{ Toeplitz, Lower-triangular} \\ & |\Delta_{ij}| \leq \eta. \end{array} \tag{11}$$