Telling Cause from Effect in Deterministic Linear Dynamical Systems

Naji Shajarisales[1] NAJI@TUEBINGEN.MPG.DE
Dominik Janzing[1] DOMINIK.JANZING@TUEBINGEN.MPG.DE
Bernhard Schölkopf[1] BS@TUEBINGEN.MPG.DE
Michel Besserve[1,2] MICHEL.BESSERVE@TUEBINGEN.MPG.DE
[1] MPI for Intelligent Systems, Tuebingen, Germany
[2] MPI for Biological Cybernetics , Tuebingen, Germany

Abstract

Telling a cause from its effect using observed time series data is a major challenge in natural and social sciences. Assuming the effect is generated by the cause through a linear system, we propose a new approach based on the hypothesis that nature chooses the "cause" and the "mechanism generating the effect from the cause" independently of each other. Specifically we postulate that the power spectrum of the "cause" time series is uncorrelated with the square of the frequency response of the linear filter (system) generating the effect. While most causal discovery methods for time series mainly rely on the noise, our method relies on asymmetries of the power spectral density properties that exist even in deterministic systems. We describe mathematical assumptions in a deterministic model under which the causal direction is identifiable. In particular, we show a scenario where the method works but Granger causality fails. Experiments show encouraging results on synthetic as well as real-world data. Overall, this suggests that the postulate of Independence of Cause and Mechanism is a promising principle for causal inference on observed time series.

1. Introduction

A major challenge in the study of complex natural systems is to infer the causal relationships between elementary characteristics of these systems. This provides key information to understand the underlying mechanisms at play and possibly allows to intervene on them to influence the overall behavior of the system. While causal knowledge is traditionally built by performing experiments, boiling down to modifying a carefully selected parameter of the system and analyzing the resulting changes, many natural systems do not allow such interventions without tremendous cost or complexity. For example, it is very difficult to influence the activity of a specific brain region without influencing other properties of the neural system (Logothetis et al., 2010). Causal inference methods have been developed to avoid such interventions and infer the causal relationships from observational data only (Spirtes et al., 1993; Pearl, 2000). To be able to build such knowledge without interventions, these approaches have to rely on key assumptions pertaining to the mechanisms generating the observed data.

The framework of causality described in Spirtes et al. (1993) and Pearl (2000) has originally addressed this question by modelling observations as i.i.d. random variables. However, observed data from complex natural systems are often not i.i.d. and time dependent information reflects key aspects of those systems. Many causal inference methods for time series, including the widely used Granger causality (Granger, 1969), assume the data is generated from a stochastic model through a structural equation linking past values to future ones through an i.i.d. additive noise term, the "innovation of the process" (Granger, 1969; Peters et al., 2013). While these methods can successfully estimate the causal relationships when empirical data is generated according to the model assumptions, the results can be misleading when the model is misspecified. In particular, this is the case when the data generating model has unknown measurement delays.

In this paper, we introduce a new approach for inferring causal directions in time series, the Spectral Independence Criterion (SIC). The idea behind SIC, as well as several new approaches to causal inference (Daniusis et al., 2010; Janzing et al., 2010; Mooij et al., 2010; Zscheischler et al., 2011; Janzing et al., 2012; Sgouritsa et al., 2015), is to rely on the "philosophical" principle that the cause and

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

the mechanism that generates the cause from the effect are chosen independently by Nature. Thus, these two objects should not contain any information about each other (Janzing & Schölkopf, 2010; Lemeire & Janzing, 2012; Schölkopf et al., 2012). Here, we refer to this abstract principle as the postulate of Independence of Cause and Mechanism (ICM). The above mentioned methods based on ICM refer to different domains and rely on quite different formalizations of the concept of "independence". SIC formalizes the ICM postulate in the context where both cause and effect are stationary time series and the cause generates the effect through a linear time invariant filter. The SIC postulate assumes that the frequency spectrum of the cause does not correlate with the frequency response of the filter. This assumption will be justified by its connection to the Trace Method (Janzing et al., 2010) and by a generative model of the system. Under this postulate, we prove that SIC can tell the causal direction of the system from its anti-causal counterpart. Moreover, we elaborate on the connection between this novel framework and linear Granger causality, showing they are exploiting fundamentally different information from the observed data. In addition, superiority to Granger causality is shown theoretically in the context of time series measurements perturbed by an unknown time lag. We perform extensive experimental comparisons, both on simulated and real datasets. In particular, we show that our approach outperforms Granger causality to estimate the direction of causation between two structures of rat hippocampus using Local Field Potential (LFP) recordings.

Overall, the proposed method constitutes a novel approach to causal inference for time series data with identifiability results, and shows unprecedented robustness to measurement delays. The encouraging results on real data suggest ICM is a promising principle to design new causal inference methods for empirical time series.

2. Spectral Independence Criterion (SIC)

2.1. Notations and Model Description

We refer to a sequence of real or complex numbers $\mathbf{a} = \{a_t, t \in \mathbb{Z}\}$ as a *deterministic time series*. Its discrete Fourier transform (represented by $\widehat{\mathbf{a}}$) is defined by

$$\widehat{\mathbf{a}}(\nu) = \sum_{t\in\mathbb{Z}} a_t \exp(-\mathbf{i}2\pi\nu t),\ \nu \in [-1/2,\ 1/2] =: \mathcal{I}$$

The *energy* of the deterministic time series is the squared l^2 norm: $\|\mathbf{a}\|_2^2 = \sum_t |a_t|^2$. For ease of notation we will also use the Z-transform of $\mathbf{a}$

$$\widetilde{\mathbf{a}}(z) = \sum_{t\in\mathbb{Z}} a_t z^{-t},\ z \in \mathbb{C}$$

such that $\widehat{\mathbf{a}}(\nu) = \widetilde{\mathbf{a}}(\exp(\mathbf{i}2\pi\nu))$.

We assume that the causal mechanism is given by a (deterministic) Linear Time Invariant (LTI) filter. That is, the causal mechanism is formalized by the convolution

$$\mathbf{y} = \{y_t\} = \{\sum_{\tau\in\mathbb{Z}} x_{t-\tau} h_\tau\} = \mathbf{x} * \mathbf{h}, \tag{1}$$

where $\mathbf{h}$ denotes the *impulse response*, $\mathbf{x}$ the input time series and $\mathbf{y}$ the output. We will assume that the filter satisfies the Bounded Input Bounded Output (BIBO) stability property (Proakis, 2001), which boils down to the condition $\|\mathbf{h}\|_1 = \sum_t |h_t| < +\infty$. Under this assumption, the Fourier transform $\widehat{\mathbf{h}}$ is well defined and we call it the *frequency response* of the system. We assume that the input time series $\mathbf{x}$ is a sample drawn from a *stochastic process*, $\{X_t, t \in \mathbb{Z}\}$. We use $\{X_t\}$ or simply $\mathbf{X}$ to represent the complete stochastic process. We use $\mathbf{X}_{t:s}$ to indicate the random vector corresponding to the restriction of the time series to the integer interval $[t\,..\,s]$. Assuming $\mathbf{X}$ is a zero mean stationary process (in this paper, stationary will always stand for weakly or wide-sense stationary as defined in Brockwell & Davis (2009)), we will denote by $C_{xx}(\tau) = \mathbb{E}[X_t X_{t+\tau}]$ the autocovariance function of the process and assume it is absolutely summable. Then, we can define its *Power Spectral Density* (PSD) $S_{xx} = \widehat{C_{xx}}$. Under these assumptions, the power of the process $P(\mathbf{X}) = \mathbb{E}(|X_t|^2)$ is finite and $P(\mathbf{X}) = \int_{-1/2}^{1/2} S_{xx}(\nu)d\nu$, such that S_{xx} belongs to L^1. Moreover, we recall the following basic property for our model:

Proposition 1. *Assume the weakly stationary input* $\mathbf{X}$ *is filtered by the BIBO linear system of impulse response* $\mathbf{h}_{\mathbf{X}\to\mathbf{Y}}$ *to provide the output* $\mathbf{Y}$*. Then* $\|\mathbf{h}_{\mathbf{X}\to\mathbf{Y}}\|_2^2 < +\infty$*,* $\widehat{\mathbf{h}} \in L^\infty$ *and* $\mathbf{Y}$ *is weakly stationary with summable autocovariance such that*

$$S_{yy}(\nu) = |\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}(\nu)|^2 S_{xx}(\nu), \nu \in \mathcal{I}. \tag{2}$$

Proof. Results from elementary properties of the Fourier transform and Proposition 3.1.2. in Brockwell & Davis (2009). □

If such a linear filtering relationship exists for $\mathbf{X}$ as input and $\mathbf{Y}$ as output, but not in the opposite way, we can use this information to infer that $\mathbf{X}$ is causing $\mathbf{Y}$ and not the other way around. If there exist such impulse responses for both directions, say $\mathbf{h}_{\mathbf{X}\to\mathbf{Y}}$ and $\mathbf{h}_{\mathbf{Y}\to\mathbf{X}}$, their Fourier transforms are related by

$$\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}} = \frac{1}{\widehat{\mathbf{h}}_{\mathbf{Y}\to\mathbf{X}}}, \tag{3}$$

and we have to resort to a more refined criterion for the causal inference. We will assume this situation in the remaining of the paper.

2.2. Definition of SIC

Assume we are given the two processes $\mathbf{X} := \{X_t, t \in \mathbb{Z}\}$ and $\mathbf{Y} := \{Y_t, t \in \mathbb{Z}\}$. Moreover, we assume that exactly one of the following two alternatives is true: (1) $\mathbf{X}$ causes $\mathbf{Y}$ or (2) $\mathbf{Y}$ causes $\mathbf{X}$. Suppose that there are no unobserved common causes of $\mathbf{X}$ and $\mathbf{Y}$. Our causal inference problem thus reduces to a binary decision. In the spirit of ICM, we propose a Spectral Independence Criterion (SIC) which assumes that in case (1), $\mathbf{X}$ and $h_{\mathbf{X}\to\mathbf{Y}}$ should not contain information about each other. To formalize this idea, we assume:

Postulate 1 (Spectral Independence Criterion (SIC)). *If $\mathbf{Y}$ is the effect generated by $\mathbf{X}$ through a LTI system with impulse response $\mathbf{h}_{\mathbf{X}\to\mathbf{Y}}$, then we have:*

$$\langle |\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}|^2 S_{xx}\rangle = \langle |\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}|^2\rangle\langle S_{xx}\rangle\,, \tag{4}$$

where $\langle f\rangle = \int_{\mathcal{I}} f(\nu)d\nu$ denotes the average over the unit frequency interval $\mathcal{I}$.

Note that the left hand side of (4) is the average intensity of the output signal $\{Y_t, t \in Z\}$ over all frequencies. Hence, SIC states that the average output intensity is the same as amplifying all frequencies by the average amplifying factor. To motivate why we call (4) an *independence* criterion we note that the difference between the l.h.s and r.h.s can be written as a covariance:

$$\langle S_{xx}\cdot|\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}|^2\rangle - \langle S_{xx}\rangle\langle|\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}|^2\rangle = \mathbb{C}\mathrm{ov}\left[S_{xx}, |\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}|^2\right],$$

where we consider S_{xx} and $|\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}|^2$ as functions of the random variable ν uniformly distributed on $\mathcal{I}$. As a consequence statistical independence between those random variables implies that (4) is satisfied.

Note that the criterion (4) can be rephrased in terms of the power spectra of $\mathbf{X}$ and $\mathbf{Y}$ alone using (2), which are closer to observable quantities than $\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}$:

Proposition 2 (SIC in terms of power spectra). *If $\mathbf{Y}$ is the effect generated by $\mathbf{X}$ through a LTI system, the SIC postulate is equivalent to:*

$$\langle S_{yy}\rangle = \langle S_{xx}\rangle\langle S_{yy}/S_{xx}\rangle\,. \tag{5}$$

2.3. Quantifying Violation of SIC

This motivates us to define a measure of dependence between the input PSD on the one hand and frequency response of the mechanism on the other hand. To assess to what degree such a relation holds we introduce a scale invariant expression $\rho_{\mathbf{X}\to\mathbf{Y}}$, that we call the Spectral Dependency Ratio (SDR) from $\mathbf{X}$ to $\mathbf{Y}$:

$$\rho_{\mathbf{X}\to\mathbf{Y}} := \frac{\langle S_{yy}\rangle}{\langle S_{xx}\rangle\langle S_{yy}/S_{xx}\rangle} \tag{6}$$

Here, the value 1 means independence in the sense of SIC, which becomes more obvious by rewriting (6) as

$$\rho_{\mathbf{X}\to\mathbf{Y}} = \frac{\mathbb{C}\mathrm{ov}[S_{xx}, |\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}|^2]}{\langle S_{xx}\rangle\langle|\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}|^2\rangle} + 1\,.$$

We define $\rho_{\mathbf{Y}\to\mathbf{X}}$ as well by exchanging the roles of $\mathbf{X}$ and $\mathbf{Y}$:

$$\rho_{\mathbf{Y}\to\mathbf{X}} := \frac{\langle S_{xx}\rangle}{\langle S_{yy}\rangle\langle S_{xx}/S_{yy}\rangle}. \tag{7}$$

2.4. Identifiability Results

In order to identify the true causal direction from SIC, it is necessary to show that $\rho_{\mathbf{X}\to\mathbf{Y}}$ and $\rho_{\mathbf{Y}\to\mathbf{X}}$ take characteristic values that are informative about this inference problem. The following crucial result shows explicitly how dependence measures in both directions are related:

Proposition 3. (Forward-backward inequality) *For a given linear filter with input PSD S_{xx}, output PSD S_{yy} and a frequency response $\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}$ of non-constant modulus, we have*

$$\rho_{\mathbf{X}\to\mathbf{Y}}\cdot\rho_{\mathbf{Y}\to\mathbf{X}} = \frac{1}{\langle|\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}|^2\rangle\langle 1/|\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}|^2\rangle} < 1\,. \tag{8}$$

Moreover, if it exists $\alpha > 0$ such that, for all $\nu \in \mathcal{I}$, $|\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}(\nu)|^2 \leq (2-\alpha)\|\boldsymbol{h}_{\mathbf{X}\to\mathbf{Y}}\|_2^2$, then

$$\rho_{\mathbf{X}\to\mathbf{Y}}\cdot\rho_{\mathbf{Y}\to\mathbf{X}} \leq \left[1 + \alpha CV\left(|\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}|^2\right)^2\right]^{-1} < 1\,. \tag{9}$$

where $CV(|\widehat{\mathbf{h}}|^2)$ denotes the coefficient of variation of $|\widehat{\mathbf{h}}|^2$ along the frequency axis, i.e. the ratio of the standard deviation to the mean.

The proof of this proposition is given in the *supplementary material*. Assuming the SIC postulate is satisfied in the forward direction such that $\rho_{\mathbf{X}\to\mathbf{Y}} = 1$, it follows naturally from 8 that $\rho_{\mathbf{Y}\to\mathbf{X}} < 1$. However $\rho_{\mathbf{X}\to\mathbf{Y}}$ cannot be guaranteed to be exactly one in practice, due to statistical fluctuations. According to equation (9), the more $|\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}|^2$ fluctuates around its mean (resulting in a large coefficient of variation), the more the product of the independence measures can be bounded away from 1. Equation (9) thus guaranties that the two causal directions can still be distinguished whenever $\rho_{\mathbf{X}\to\mathbf{Y}}$ fluctuates around one. In particular, if $\rho_{\mathbf{X}\to\mathbf{Y}}$ is slightly below 1, the bound (9) still guarantees that $\rho_{\mathbf{Y}\to\mathbf{X}}$ is even smaller provided that the coefficient of variation of $|\widehat{\mathbf{h}}_{\mathbf{X}\to\mathbf{Y}}|^2$ is sufficiently large. As a consequence, our causal inference rule will select the causal direction to be the one with the largest ρ value.

How well real-world systems satisfy SIC in causal direction cannot be answered by theory alone. This is because

bounding the probability for large deviations of $\rho_{\mathbf{X}\to\mathbf{Y}}$ from 1 relies on simplistic models where linear systems are generated by random processes. We now describe such a model that independently generates $\mathbf{X}$ and $\mathbf{h}_{\mathbf{X}\to\mathbf{Y}}$ according to some simple prior without claiming that this is an appropriate description of how nature generates causal relations. We start with a Finite Impulse Response (FIR) $\mathbf{h}$, that is, $h_\tau = 0$ for all $\tau < k$ and $\tau \geq k+m$, for some m and k. Then $\mathbf{h}$ is given by m real numbers $b_1, \ldots, b_m$ such that

$$h_{i+k} = b_i \quad i = 0, \ldots, m-1\,.$$

We then apply an orthogonal transformation $\mathbf{U}$, randomly drawn from the orthogonal group $O(m)$ according to the 'uniform distribution' on $O(m)$, that is, the Haar measure. In this way, we generate a new impulse response function

$$h'_{i+k} := (\mathbf{U}\mathbf{b})_i \quad i = 0, \ldots, m-1\,, \tag{10}$$

where $h'_\tau = 0$ for all $\tau < k$ and $\tau \geq k+m$. Since orthogonal transformations preserve the Euclidean norm by definition, they preserve the energy of the filter. Our procedure thus chooses a random filter among the set of filters having the same support of length m and the same energy. We now show that for large m the resulting filter will approximately satisfy SIC with high probability:

Theorem 1. (Concentration of Measure for FIR filters) *For some fixed S_{xx}, let $\rho^U_{\mathbf{X}\to\mathbf{Y}}$ be the SDR obtained from $\mathbf{h}'$ in (10). If $\mathbf{U}$ is chosen from the Haar measure on $O(m)$, then for any given $\varepsilon > 0$*

$$|\rho^U_{\mathbf{X}\to\mathbf{Y}} - 1| \leq \frac{2\varepsilon}{P(\mathbf{X})} \max_\nu S_{xx}(\nu)\,.$$

with probability $\delta := 1 - \exp(-\kappa(m-1)\varepsilon^2)$ where κ is a positive global constant independent of m, ε, $\mathbf{X}$ and $\mathbf{Y}$.

The proof of this theorem is provided in the *supplementary material*. This result provides a justification for using SIC provided that the number of nonzero elements of $\mathbf{h}$, m, is large enough. The relevance of m will be investigated in practice in the experimental section.

Note also the following Bayesian generalization of Theorem 1: Whenever one assigns independent priors to $\mathbf{h}_{\mathbf{X}\to\mathbf{Y}}$ and $\mathbf{X}$ and the former one is $O(m)$-invariant, the above bound holds regardless of how the latter has been chosen. This way, we avoid the question of defining appropriate priors for the input signal.

2.5. Relation to the Trace Condition

We now describe the relation between SIC and a causal inference method called Trace Method (Janzing et al., 2010). Let X and Y be n- and m-dimensional variables respectively, causally related by the linear structural equation

$$Y = AX + E\,,$$

where A is an $m \times n$ structure matrix and E is a m-dimensional noise variable independent of X. Janzing et al. (2010) postulate the following independence condition between the covariance matrix of input distribution Σ_X and A:

Postulate 2 (Trace Condition)**.**

$$\tau_m(A\Sigma_X A^T) = \tau_n(\Sigma_X)\tau_m(AA^T)\,, \tag{11}$$

approximately, where $\tau_n(B)$ denotes the renormalized trace $\mathrm{tr}(B)/n$.

The postulate can be justified by random matrix theory with large m when A and Σ_X are independently chosen according to a distribution satisfying appropriate symmetry assumptions (Janzing et al., 2010). To link SIC and trace method we only need square matrices, i.e, $m = n$.

To quantify the violation of (11) we introduce the following quantity:

Definition 1 (Tracial Dependency Ratio (TDR))**.** *The tracial dependency ratio is given by*

$$r_{X\to Y} := \frac{\tau_n(A\Sigma_X A^T)}{\tau_n(\Sigma_X)\tau_m(AA^T)}\,. \tag{12}$$

We thus can see that the TDR plays a role analogue to our SDR ρ in the finite dimensional case. We can actually show that SIC can be viewed as a limit case of the Trace Condition by defining the following truncated system.

Definition 2. *For any given infinite dimensional linear system $\mathbf{X} \mapsto \mathbf{Y} = \boldsymbol{h} * \mathbf{X}$, the truncated system of order N is defined by zeroing the input and the output values for integers k such that $-N \leq k < N$:*

$$\mathbf{X}'_N = \mathbf{X}_{-N:N-1} \mapsto \mathbf{Y}'_N = (\mathbf{h} * \mathbf{X}'_N)_{-N:N-1}.$$

Note that in this definition for each N, the vectors $\mathbf{Y}'_{-N:N-1}$ are inherently different. To reduce the notational complications, for any stochastic time series $\mathbf{Z}$ and a given N we represent $\mathbf{Z}_{-N:N-1}$ as $\mathbf{Z}_N$. Every truncated system defines a $2N$ dimensional deterministic linear structural equation. We then have the following result showing that SIC can be obtained from the Trace Condition as an appropriate limit:

Theorem 2. *Let $r_{\mathbf{X}'_N\to\mathbf{Y}'_N}$ represent the TDR for the truncated systems of order N for a given linear system with SDR $\rho_{\mathbf{X}\to\mathbf{Y}}$. Then*

$$\lim_{N\to\infty} r_{\mathbf{X}'_N\to\mathbf{Y}'_N} = \rho_{\mathbf{X}\to\mathbf{Y}}$$

The proof, together with two necessary lemmas is available in *supplementary material*.

3. SIC for Vector Autoregressive Models

SIC and Granger causality rely on completely different assumptions but both apply to linear systems. In this section, we study the classical Vector Autoregressive (VAR) model used in Granger causality from the SIC perspective to better understand how they are related.

3.1. VAR Model

We assume the observed time series are generated by a VAR model such that $\mathbf{X}$ Granger causes $\mathbf{Y}$, but $\mathbf{Y}$ *does not* Granger causes $\mathbf{X}$:

$$X_t = \sum_k a_k X_{t-k} + \epsilon_t \tag{13}$$

$$Y_t = \sum_k b_k Y_{t-k} + \sum_k c_k X_{t-k} + \xi_t \tag{14}$$

Both noise terms $\boldsymbol{\epsilon}$ and $\boldsymbol{\xi}$ in this expression are i.i.d normal noises.

3.2. Applying SIC to VAR Models

We want to rewrite this expression such that $\mathbf{Y}$ is generated from $\mathbf{X}$ by a deterministic linear time invariant filter. We observe that the VAR model can be cast as a linear time invariant filter if we neglect the additive noise ξ. Indeed, then the mechanism is the following noiseless ARX (AutoRegressive with eXogenous input) model (Keesman, 2011):

$$Y_t = \sum_k b_k Y_{t-k} + \sum_k c_k X_{t-k}. \tag{15}$$

As a consequence, testing SIC on the VAR model in the forward direction amounts (when neglecting the filtered noise ξ) to test independence between

$$S_{xx}(\nu) = \frac{1}{|1 - \sum_k a_k \exp(-2\pi \mathbf{i} k\nu)|^2} \tag{16}$$

and

$$|\widehat{\mathbf{h}}(\nu)|^2 = \frac{|\sum_k c_k \exp(-2\pi \mathbf{i} k\nu)|^2}{|1 - \sum_k b_k \exp(-2\pi \mathbf{i} k\nu)|^2}, \tag{17}$$

parametrized by the coefficients $\{a_k\}$ and $\{b_k, c_k\}$ respectively (these expression are derived in the *supplementary material*). We conjecture that a concentration of measure result similar to Theorem 1 holds, stating that independent choice of the coefficients from an appropriate symmetric distribution typically yields small correlations between the functions defined in (16) and (17). This will be tested empirically in Section 4. The robustness of our approach to noise in the VAR model remains an interesting question to be addressed in future work.

3.3. Comparison of SIC and Granger Causality

The bivariate VAR model above is the typical model where Granger causality works. To recall the idea of the latter, note that it infers that there is an influence from $\mathbf{X}$ to $\mathbf{Y}$ whenever predicting $\mathbf{Y}$ from its past is improved by accounting for the past of $\mathbf{X}$. Within the context of the above linear model, knowing $X_{t-1}, X_{t-2}, \ldots$ reduces the variance of Y_t, given $Y_{t-1}, Y_{t-2}, \ldots$ because then the noise ϵ_t is the only remaining source of uncertainty. Without knowing $X_{t-1}, X_{t-2}, \ldots$, we have additional uncertainty due to the contribution of $\epsilon_{t-1}, \epsilon_{t-2}, \ldots$.

SIC, on the other hand, does not rely on detecting whether $\mathbf{X}$ helps in improving the prediction of $\mathbf{Y}$. As demonstrated above, SIC applied to a bivariate VAR model boils down to quantifying independence between two linear filters parametrized by sets of coefficients, the filter generating the input with frequency response $\widehat{n}$ and the filter of the mechanism with frequency response $\widehat{m}$. This is a completely different concept. One can easily imagine that the coefficients $\{b_k, c_k\}$ and $\{a_k\}$ can be hand-designed such that the functions (17) and (16) are correlated. This would spoil SIC, but leave Granger unaffected. On the other hand, the following subsection describes a scenario where Granger fails but SIC still works.

3.4. Sensitivity to Time Lag

Consider two time series $\{X_t\}$ and $\{Y_t\}$ where $\{X_t\}$ is white noise and

$$\forall t \in \mathbb{Z}, \quad Y_t = cY_{t-1} + X_{t-1},$$

for a given c. It can be easily seen that this type of input and output can be simulated using an IIR filter with $(a_1, a_2) = (1, c)$ and $b_1 = 1$ in (18) where the rest of the coefficients are zero (please refer to the definition of coefficients in section 4.1). The infinite DAG for this causal structure can be seen in Fig. 1.

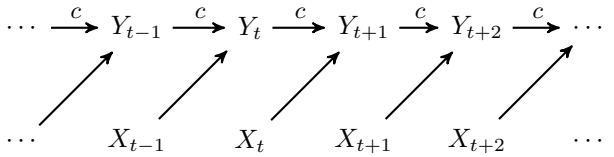

Figure 1. The original causal structure of section 3.4

Now if there would be a measurement delay of length k for $\mathbf{Y}$, the observed values will be a new time series, say $\tilde{\mathbf{Y}}$, where $\tilde{Y}_t = Y_{t-k}$. Although the ground truth is $\mathbf{X} \to \tilde{\mathbf{Y}}$ independent of k, Granger causality only infers the correct causal structure if $k \geq 0$ (where there is a lag in measurement of $\mathbf{X}$, but not $\mathbf{Y}$). However SIC always infers the correct direction (except when $c = 0$ and the time structure

is spoiled). This is because the PSD of the white noise $\mathbf{X}$ is constant and depends only on the total power, i.e,

$$S_{xx}(\nu) = \mathrm{Var}(X_t) = P(\mathbf{X}),$$

for all $\nu \in [-1/2, 1/2]$, and obviously this constant remains the same for the lagged time series. In addition, time lags are irrelevant for SIC since they do not affect the power spectrum.

4. Experiments

In this section we study our causal inference algorithm using synthetic experiments and also apply it to several real world data sets.

4.1. Synthetic Data: ARMA Filters and Processes

We designed synthetic experiments to assess the validity of the SIC approach. The data generating process is as follows. The LTI filter $\mathcal{S}$ modeling the mechanism is chosen among the family of *ARMA*$(FO(\mathcal{S}), BO(\mathcal{S}))$ filters with parameters $(\mathbf{a}, \mathbf{b})$ defined by input-output difference equation:

$$y_n = \frac{1}{a_0}\left(\sum_{i=0}^{FO(\mathcal{S})} b_i x_{n-i} + \sum_{j=1}^{BO(\mathcal{S})} a_j y_{n-j}\right). \qquad (18)$$

For these filters, $FO(\mathcal{S})$ and $BO(\mathcal{S})$ (FO and BO for short) are the feedforward and feedback order respectively. a_i's are feedforward and b_i's are feedback coefficients. Note that when $FO = 0$, the filter is called an autoregressive filter. Alternatively, $BO = 0$ corresponds to the family of causal Finite Impulse Response (FIR) filters. Whenever $BO \neq 0$, the filter has Infinite Impulse Response (IIR). We chose two filters $\mathcal{S}$ and $\mathcal{S}'$, with parameters $(\mathbf{a}, \mathbf{b})$ and $(\mathbf{a}', \mathbf{b}')$ respectively. To simulate a cause effect pair $\mathbf{X}$,$\mathbf{Y}$, we generated the cause $\mathbf{X}$ by applying $\mathcal{S}$ to a normally distributed i.i.d noise. Then, we generated $\mathbf{Y}$ by applying $\mathcal{S}'$ to $\mathbf{X}$.

In each trial all the elements of vectors $\mathbf{a}$, $\mathbf{a}'$, $\mathbf{b}$ and $\mathbf{b}'$ except the first ones (i.e. a_0, b_0, a'_0, b'_0 which were fixed to one) were sampled from an isotropic multidimensional Gaussian distribution with variance 0.01. Coefficients are sampled using rejection sampling such that only BIBO-stable filters are kept.

We simulated sequences of length 1000. The PSD of $\mathbf{X}$ and $\mathbf{Y}$ were estimated using Welch's method (Welch, 1967). We repeated this experiment 1000 times. Figure 2 shows an example of the distribution of $\rho_{\mathbf{X}\to\mathbf{Y}}$ and $\rho_{\mathbf{Y}\to\mathbf{X}}$, as well as their difference using $FO(\mathcal{S}) = BO(\mathcal{S}) = FO(\mathcal{S}') = BO(\mathcal{S}') = 10$.

The SDR for the correct direction is concentrated around one, while in the wrong direction most of the probability mass is below 1 (in this example 99.1%). Consequently the SDR in correct direction is larger than the one in wrong direction in 88.3% of the cases. Accordingly, our inference algorithm based on the sign of this difference will select the correct direction in most of the cases. Using this inference algorithm, we tested the effect of the filter orders on the performance of the method. We evaluated the performance of each setting of $FO(\mathcal{S})$, $FO(\mathcal{S}')$, $BO(\mathcal{S})$ and $BO(\mathcal{S}')$ over 1000 trials. We varied the orders between 2 and 21 and compared the performance of the cases $FO(\mathcal{S}') = BO(\mathcal{S}')$, $FO(\mathcal{S}') = 0$ and $BO(\mathcal{S}') = 0$. Considering that the experiments are independent and based on the assumption that our method is successful with probability p where p has a binomial distribution, we calculated confidence intervals using Wilson's score interval (Wilson, 1927) where $\alpha = 0.05$ (and therefore $z_{\alpha/2} = 1.96$). The performance increases rapidly with filter order, as can be seen in the plots of Fig. 3. Moreover, the feedforward filter coefficients seem the most beneficial to the approach, since their absence leads to the worst performance (Fig. 3, red line).

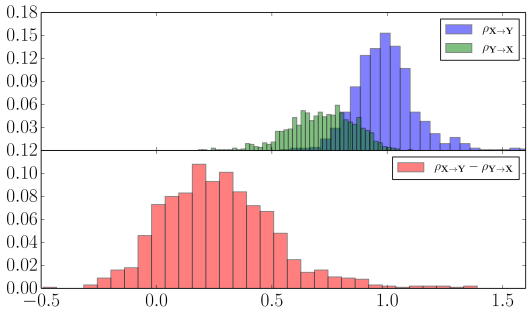

Figure 2. Top plot: Histogram for the estimators of $\rho_{\mathbf{X}\to\mathbf{Y}}$ and $\rho_{\mathbf{Y}\to\mathbf{X}}$. Bottom plot: Histogram of the estimated difference $\rho_{\mathbf{X}\to\mathbf{Y}} - \rho_{\mathbf{Y}\to\mathbf{X}}$

Algorithm 1 SIC_Inference

function SIC_Inference($\mathbf{X}$,$\mathbf{Y}$)
  Calculate $\rho_{\mathbf{X}\to\mathbf{Y}}$ and $\rho_{\mathbf{Y}\to\mathbf{X}}$ using (6)
  if $\rho_{\mathbf{X}\to\mathbf{Y}} > \rho_{\mathbf{Y}\to\mathbf{X}}$ return $\mathbf{X} \to \mathbf{Y}$
  else return $\mathbf{Y} \to \mathbf{X}$
end function

4.2. Real World Examples

We tried our method over several examples of real data where the ground truth about the causal structure of the data is known a priori and the data is labelled in a way that the ground truth is $\mathbf{X} \to \mathbf{Y}$.

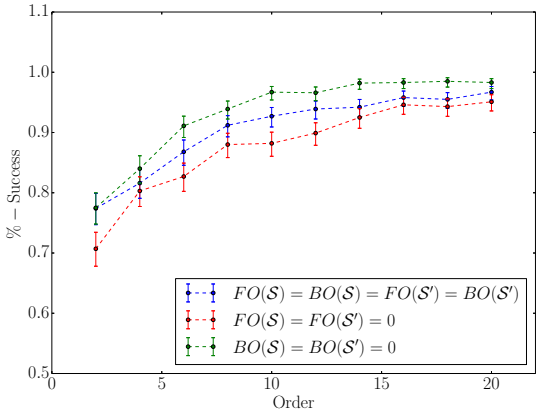

Figure 3. SIC performance against filter order for synthetic experiments for different types of filters (see text).

4.2.1. Gas Furnace (Box et al., 2013)

This dataset consists of 296 time points, with $\mathbf{X}$ the gas rate consumed by a gas furnace and $\mathbf{Y}$ the produced rate of CO_2. Fig. 4 shows $\rho_{\mathbf{X}\to\mathbf{Y}} - \rho_{\mathbf{Y}\to\mathbf{X}}$ against the window length, which was ranging from 50 to 150 points. As illustrated, the difference is always positive and our method is able to correctly infer the right causal direction independent from window length. TiMiNO and Granger causality correctly identified the ground truth in this case as well (Peters et al., 2013).

4.2.2. Old Faithful Geyser (Azzalini & Bowman, 1990)

$N = 298$: $\mathbf{X}$ contains the duration of an eruption and $\mathbf{Y}$ is the time interval to the next eruption of the Old Faithful geyser. Figure 4 represents the difference in SDRs as a function of window length with the same configuration as the gas furnace experiment. Again the correct causal direction is inferred by our method independently of the window length as illustrated in Fig. 4. In this case TiMiNO correctly identifies the cause from effect but neither linear nor nonlinear Granger causality infer the correct causal direction (Peters et al., 2013).

4.2.3. LFP recordings of the Rat Hippocampus

It is known that contrary to neocortex where connectivity between areas is bidirectional, monosynaptic connections between several regions of the hippocampus are mostly unidirectional (Andersen et al., 2006). An important example of such connectivity is between the CA3 and CA1 subfields (Andersen et al., 2006). Despite this anatomical fact, a study of causality based on Local Field Potentiareportl (LFP) recordings of CA1 and CA3 of the hippocampus of the rat during sleep reports that Granger causality

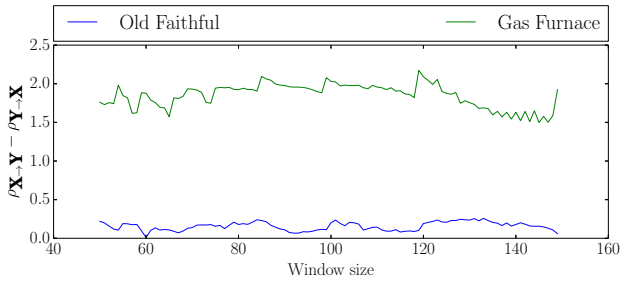

Figure 4. Difference between the estimators of SDRs in both directions against window length of the Welch periodogram.

infers strong bidirectional relations between the two areas (Baccala et al., 1998). Baccala et al. (1998) explains the possible reasons of such result as feedback loops involving cortex and medial septum, and diffuse connections going from CA1 to CA3.

To do a comparison with Granger causality, we applied our framework to recordings from those regions using a publicly available dataset[1] (Mizuseki et al., 2009; 2006). LFP's were recorded using a 8 shank probe having 64 channels downsampled to 1252Hz. Shanks were attributed by experimentalists to the CA1 and CA3 areas (leaving 32 channels for each area). For more information on the details of gathered data please refer to Mizuseki et al. (2006). We used the data for rat "vvp01" during a period of sleep and a period of active behaviour in a linear environment. We applied linear Granger causality using an implementation from the *statsmodel* Python library[2]. We considered a forced decision scheme for Granger causality (to make it comparable to our method), where we selected the correct Granger causal direction as the one having the lowest p-value for the null hypothesis of absence of causal influence. Following the usual methodology of causality analysis (Baccala et al., 1998; Cadotte et al., 2010), we divided the duration of ten minutes into 300 intervals of two seconds ($N = 2504$) to reduce the effect of nonstationarity in data analysis, and performed SIC causal inference on each interval for each electrode pair. We took two different approaches to report the performance of methods: one, based on a majority vote over all 300 intervals for each channel pair, and two, by assessing the average performance based on individual time intervals. The results are plotted as histograms in Fig. 5 and they show that SIC clearly outperforms Granger causality on this dataset. The confidence intervals are once again based on Wilson score but obviously this time the independence assumption between the trials is not well justified.

[1] http://crcns.org/data-sets/hc

[2] Statsmodels: Statistical library for Python. More details on the null hypothesis for Granger causality can be found on the website.

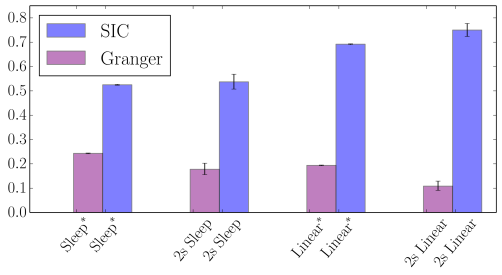

Figure 5. Average performance of the linear Granger causality and SIC methods for deciding CA3→CA1 ground truth direction against the opposite. For both the linear and sleep sessions the performance is significantly above the chance level for SIC. $*$ indicates the use of a majority voting scheme.

4.2.4. Characterizing the Echo

The echo effect of a room on a sound generated in the room can be well estimated by a convolution of the generated signal with a function known as room Impulse Response Function (IRF). In this experiment we used an open source database of room IRFs available at the Open AIR library[3]. We chose the IRFs for Elevden Hall, Elevden, Suffolk, England and Hamilton Mausoleum, Hamilton, Scotland. We convolved these signals with 30 ± 5 seconds segments of two classical music pieces: The first movement of Vivaldi's Winter Concerto consisting of 9190656 data points, and the Lacrimosa of Mozart's Requiem, consisting of 8842752 points, both '.wav' files with the rate of 44100Hz. Regardless of the segment the SDR in forward direction is considerably larger than the SDR in the backward direction as can be seen in Fig. 6. In another experiment we used a computer to play the musical pieces above in an academic Lecture Hall (labelled as "Hall" in plots) and in an office room (labelled as "Room" in plots) and recorded the echoed version in the environment. In a series of different tests, we splitted the data into $9, 17, 33, 65, 129$ pieces, and we ignored the last piece so that all the pieces would have an equal length. In each test we averaged the performance of our causal inference method over all the segments and plotted this performance against the size of the window length in Welch method. The window size was varied between 500 and half of the length of the music segment length (which is dependent on the number of segments). The results can be found in Fig. 7 and show a very good performance of the approach for large window lengths.

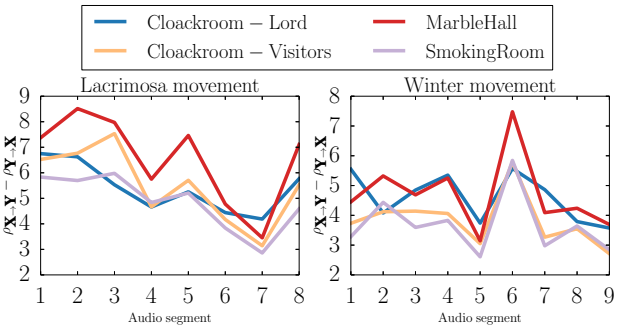

Figure 6. The plots represent the value $\rho_{\mathbf{X}\to\mathbf{Y}} - \rho_{\mathbf{Y}\to\mathbf{X}}$ for 4 different environments as a function of different music segments. The method correctly infers the causal direction in all the cases.

[3]Open AIR: Open source library for acoustic IRFs.

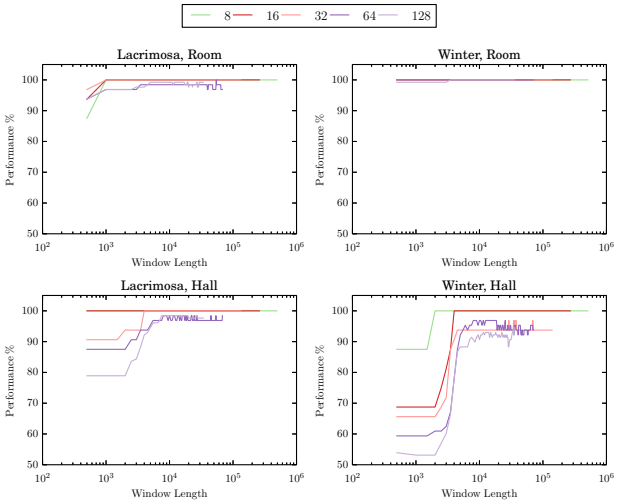

Figure 7. The performance of the method over real echoed audio signals recorded simultaneously by playing the piece in two different closed environments that have their own acoustic structure.

5. Conclusion

We have introduced a causal discovery method for time series based on the SIC postulate, assuming a LTI relationship for a given pair of time series $\mathbf{X}$ and $\mathbf{Y}$, such that either $\mathbf{X} \to \mathbf{Y}$ or $\mathbf{Y} \to \mathbf{X}$. Theoretical justifications are provided for this postulate to lead to identifiability. The method provides an extension of the recently proposed Trace Method for time series. Encouraging experimental results have been presented on real world and synthetic data. In particular, this method proved to be more effective than linear Granger causality on LFP recordings from the CA1 and CA3 areas of rat hippocampus, assuming a ground truth causal direction from CA3 to CA1 based on anatomy. We suggest that this method provides a new perspective for causal inference in time series based on assumptions fundamentally different from Granger causality. Including confounders, establishing a statistical significance test (for example using a procedure inspired by (Zscheischler et al., 2011)), and extending this method to multivariate time series is left to future work.

References

Andersen, P., Morris, R., Amaral, D., Bliss, T., and O'Keefe, J. *The Hippocampus Book*. Oxford University Press, 2006.

Azzalini, A. and Bowman, A. W. A look at some data on the Old Faithful geyser. *Applied Statistics*, pp. 357–365, 1990.

Baccala, L. A., Sameshima, K., Ballester, G., Do Valle, A. C., and Timo-Iaria, C. Studying the interaction between brain structures via directed coherence and Granger causality. *Applied Signal Processing*, 5(1):40, 1998.

Box, G. E. P., Jenkins, G. M., and Reinsel, G. *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, 2013.

Brockwell, P. J. and Davis, R. A. *Time Series: Theory and Methods*. Springer Science & Business Media, 2009.

Cadotte, A. J., DeMarse, T. B., Mareci, T. H., Parekh, M. B., Talathi, S. S., Hwang, D., Ditto, W. L., Ding, M., and Carney, P. R. Granger causality relationships between local field potentials in an animal model of temporal lobe epilepsy. *Journal of Neuroscience Methods*, 189(1):121–129, 2010.

Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K., and Schölkopf, B. Inferring deterministic causal relations. In Grünwald, P. and Spirtes, P. (eds.), *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010).*, pp. 143–150, 2010.

Granger, C. W. J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

Gray, R. M. *Toeplitz and Circulant Matrices: A Review*. Now Publishers Inc., 2006.

Janzing, D. and Schölkopf, B. Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, 56(10):5168–5194, 2010.

Janzing, D., Hoyer, P. O., and Schölkopf, B. Telling cause from effect based on high-dimensional observations. In *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pp. 479–486, 2010.

Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniušis, P., Steudel, B., and Schölkopf, B. Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182:1–31, 2012.

Keesman, K. J. *System Identification: an Introduction*. Springer Science & Business Media, 2011.

Lemeire, J. and Janzing, D. Replacing causal faithfulness with algorithmic independence of conditionals. *Minds and Machines*, pp. 1–23, 7 2012.

Logothetis, N. K., Augath, M., Murayama, Y., Rauch, A., Sultan, F., Goense, J., Oeltermann, A., and Merkle, H. The effects of electrical microstimulation on cortical signal propagation. *Nature Neuroscience*, 13:1283–1291, 2010.

Mizuseki, K., Sirota, A., Pastalkova, E., Diba, K., and Buzsáki, G. Multiple single unit recordings from different rat hippocampal and entorhinal regions while the animals were performing multiple behavioral tasks. crcns.org, 2006.

Mizuseki, K., Sirota, A., Pastalkova, E., and Buzsáki, G. Theta oscillations provide temporal windows for local circuit computation in the entorhinal-hippocampal loop. *Neuron*, 64(2):267–280, 2009.

Mooij, J. M., Stegle, O., Janzing, D., Zhang, K., and Schölkopf, B. Probabilistic latent variable models for distinguishing between cause and effect. In *Advances in Neural Information Processing Systems*, pp. 1687–1695, 2010.

Pearl, J. *Causality: Models, Reasoning and Inference*. Cambridge Univ Press, 2000.

Peters, J., Janzing, D., and Schölkopf, B. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, pp. 154–162, 2013.

Proakis, J. G. *Digital Signal Processing: Principles Algorithms and Applications*. Pearson Education India, 2001.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pp. 1255–1262, 2012.

Serre, D. *Matrices: Theory and Applications*. Graduate Texts in Mathematics. Springer, 2010. ISBN 9781441976833.

Sgouritsa, E., Janzing, D., Hennig, P., and Schölkopf, B. Inference of cause and effect with unsupervised inverse regression. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*. JMLR.org, 2015.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search (Lecture Notes in Statistics)*. Springer-Verlag, New York, NY, 1993.

Welch, P. D. The use of fast fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967.

Wilson, E. B. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Yeh, J. *Real Analysis: Theory of Measure and Integration*, volume 2. World Scientific Hackensack, 2006.

Zscheischler, J., Janzing, D., and Zhang, K. Testing whether linear equations are causal: A free probability theory approach. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pp. 839–846, 2011.