

Validity of time reversal for testing Granger causality

Irene Winkler, Danny Panknin, Daniel Bartz, Klaus-Robert Müller and Stefan Haufe

Abstract—Inferring causal interactions from observed data is a challenging problem, especially in the presence of measurement noise. To alleviate the problem of spurious causality, Haufe et al. (2013) proposed to contrast measures of information flow obtained on the original data against the same measures obtained on time-reversed data. They show that this procedure, time-reversed Granger causality (TRGC), robustly rejects causal interpretations on mixtures of independent signals. While promising results have been achieved in simulations, it was so far unknown whether time reversal leads to valid measures of information flow in the presence of true interaction. Here we prove that, for linear finite-order autoregressive processes with unidirectional information flow between two variables, the application of time reversal for testing Granger causality indeed leads to correct estimates of information flow and its directionality. Using simulations, we further show that TRGC is able to infer correct directionality with similar statistical power as the net Granger causality between two variables, while being much more robust to the presence of measurement noise.

Index Terms—Granger causality, time reversal, noise, TRGC

I. INTRODUCTION

THE estimation of causal relations between time series is a signal processing topic promising to enhance our understanding of dynamical systems in numerous application domains. For data with time structure, the concept of Granger causality (GC) has gained popularity as a simple testable definition of causality based on temporal precedence. Signal processing techniques based on Granger-causality have been studied in a variety of fields such as econometrics [1], neuroscience [2], [3], [4], [5], and climate science [6], [7].

In its original formulation, a time series x_t is said to Granger-cause a time series y_t , if the past of x_t helps to predict y_t above what can be predicted by using ‘all other information in the universe’ besides the past of x_t [8]. In the bivariate framework, it is common to consider only the information contained in the past of x_t and y_t (cf. [9]).

A serious problem for the estimation of information flow using Granger causality is that spurious Granger causality can occur due to measurement noise. On one hand, if two sensors

measuring the same signal are superimposed with noise, they mutually help predicting each other’s future [10], [11]. This is a problem especially in the study of brain connectivity using non-invasive electrophysiology, where the activity at a given sensor is typically a mixture of contributions from several neuronal sources due to the volume conduction of electric currents in the head [12], [13], [14], [15], [16]. On the other hand, noise that is correlated across sensors has a similar adverse effect on estimates of directed interaction even if the actual signals-of-interest are not mixed into different sensors [17], [18]. Such spurious causality can occur in any measure based on the concept of Granger causality, including multivariate [19], [20] and non-linear [21], [22], [23] variants.

Recently, a number of ways to make causality estimates more robust to the presence of mixed signals and noise have been proposed. These include novel measures of directed information flow [12], [10], [11] as well as novel ways of assessing their statistical significance [24], [23], [17], [16], [18]. Recently, Haufe et al. [17], [16] suggested to contrast causality scores obtained on the original time series to those obtained on time-reversed signals. The intuitive idea behind this approach is that, if temporal order is crucial to tell a driver from a recipient, directed information flow should be reduced (if not reversed) if the temporal order is reversed. In fact, Haufe et al. showed that for correlated, but non-interacting signals, the use of time reversal for testing Granger causality scores (here referred to as time-reversed Granger causality, TRGC) and other metrics based on cross-spectral estimates or linear autoregressive modeling correctly leads to rejection of causal interpretations. This was confirmed for Granger causality in an independent simulation study [18] showing that TRGC leads to a much smaller fraction of false positive detections compared to the original Granger causality index, and also compares favorably against the Phase Slope Index (PSI) [11].

While time-reversed Granger causality thus displays an intriguing noise robustness property, and yields very encouraging results in simulations, its behavior *in the presence of causal interactions* is still poorly understood. In particular, it is currently unclear how Granger causality scores computed on time-reversed signals link to the causal interactions on the original time-series, and therefore whether TRGC correctly indicates the direction of causality. Theoretical guarantees have only been derived for special cases in which either the signal’s auto- and cross-covariances are very small in magnitude, or in which both signals have very similar autocorrelations [18].

The aim of this paper is two-fold. In the theory section, we provide new theoretical insights on time-reversal for testing

This work was supported by a Marie Curie International Outgoing Fellowship (grant No. 625991) within the 7th European Community Framework Program, the BMBF project ALICE II, Autonomous Learning in Complex Environments (01IB15001B), and the Brain Korea 21 Plus Program as well as the SGER Grant 2014055911 through the National Research Foundation of Korea funded by the Ministry of Education.

I. Winkler, D. Panknin, D. Bartz, K.-R. Müller and S. Haufe are with the Machine Learning Group, Technische Universität Berlin, Germany. K.-R. Müller is also with the Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea. S. Haufe is also with the Laboratory for Intelligent Imaging and Neural Computing, Columbia University, New York, USA. Correspondence to: {i.winkler,stefan.haufe}@tu-berlin.de.

Granger causality between two variables. After introducing the concepts of linear autoregressive modeling, Granger causality, and time-reversed Granger causality (Section II-A and II-B), we elaborate on the existing result of Haufe et al. [16] showing that, for mixtures of independent signals, causality measures based on cross-covariances are invariant to the reversal of the temporal order (Section II-C). This is the theoretical basis for the noise-robustness property of time-reversal testing of causality scores. We then investigate the time-reversal of a process fulfilling the assumptions typically made by Granger causality estimators: a finite-order vector autoregressive (VAR) process that is unaffected by measurement noise. We review what is known about the time-reversal of a VAR process (Section II-D), based on what we provide an analytic description of Granger causality scores of time-reversed signals in terms of their autoregressive coefficients (Section II-E) and a minimal example (Section II-F). Using these insights, we prove our main result stating that, in the case of unambiguous unidirectional information flow from x_t to y_t , time reversal leads to a decrease of the Granger-causal net information flow relative to the original time series. The difference of net Granger causality scores obtained on original and time-reversed data thus indicates the correct direction of interaction (Section II-G).

In the second part of the paper (Section III), we revisit scenarios known to cause problems for conventional Granger causality. Using simulations, we illustrate when and how the theoretical guarantees of TRGC lead to measurable performance increases in practice. We point out the implications of our theoretical and empirical results in Section IV, along with a discussion of ambiguities in causal interpretation caused by the presence of correlated residuals in VAR models.

II. THEORY

Vectors are considered to be column vectors (unless otherwise stated), and are generally typed in bold. The symbol \cdot^\top denotes the transpose operator, I the identity matrix, and $[\cdot, \cdot]$ concatenation. The symbol \otimes refers to the Kronecker product, and $\text{vec}(\cdot)$ to the vectorization operator, which converts a matrix into a column vector. The symbol $\langle \cdot \rangle$ denotes expectation. The cross-covariance matrices of a stationary process \mathbf{z}_t are denoted by

$$C_{\mathbf{z}}(h) := \left\langle (\mathbf{z}_t - \langle \mathbf{z} \rangle)(\mathbf{z}_{t-h} - \langle \mathbf{z} \rangle)^\top \right\rangle \quad \forall h \in \mathbb{Z}.$$

We use the notation \mathbf{z}_t both for an observed time series and its underlying data generating process. We denote all quantities related to the time-reversed process $\tilde{\mathbf{z}}_t := \mathbf{z}_{-t}$ with a tilde.

A process ϵ_t is said to be *white noise* if it is stationary with mean zero, finite covariance and zero autocorrelation; that is, if $C_\epsilon(h) = 0 \forall h \in \mathbb{Z} \setminus \{0\}$. Note that the covariance matrix $C_\epsilon(0)$ is not necessarily diagonal, and that neither independence nor joint Gaussianity is required.

A. Granger causality and the linear VAR model

Consider a stable bivariate vector autoregressive process of lag order p (VAR(p) process), $\mathbf{z}_t = \begin{bmatrix} x_t \\ y_t \end{bmatrix} \in \mathbb{R}^2$,

$$\mathbf{z}_t = A_1 \mathbf{z}_{t-1} + A_2 \mathbf{z}_{t-2} + \dots + A_p \mathbf{z}_{t-p} + \epsilon_t, \quad (1)$$

where $\epsilon_t \in \mathbb{R}^2$ is a 2-dimensional white noise process (that is, $\langle \epsilon_t \rangle = 0$, $\langle \epsilon_t \epsilon_{t-h}^\top \rangle = 0$ for $h \in \mathbb{Z} \setminus \{0\}$, and $\langle \epsilon_t \mathbf{z}_{t-h}^\top \rangle = 0$ for $h \in \mathbb{N} \setminus \{0\}$) with residual covariance matrix

$$\Sigma = \langle \epsilon_t \epsilon_t^\top \rangle = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy} & \Sigma_{yy} \end{bmatrix}. \quad (2)$$

The noise variables ϵ_t are also called *innovations* or *residuals*. Stability requires that $\det(I - A_1 \lambda - \dots - A_p \lambda^p) \neq 0$ for all $\lambda \in \mathbb{C}$ with $|\lambda| \leq 1$.

Following [25], x_t and y_t possess themselves autoregressive (AR) representations, which we denote by

$$x_t = \sum_{k=1}^{\infty} a_k x_{t-k} + \xi_t^x, \quad \text{Var}(\xi_t^x) =: \Sigma_x \quad \text{and} \quad (3)$$

$$y_t = \sum_{k=1}^{\infty} b_k y_{t-k} + \xi_t^y, \quad \text{Var}(\xi_t^y) =: \Sigma_y. \quad (4)$$

The residuals ξ_t^x and ξ_t^y of these two univariate processes are each serially uncorrelated, but may be correlated with each other at various lags. Importantly, even though the bivariate autoregressive process (1) is of finite order, the univariate processes (3) and (4) are in general of infinite order. We refer to (1) as the *unrestricted* or *full* model, while (3) and (4) contain the *restricted* models.

Directed Granger-causal information flow is defined based on the so-called *Granger-scores* [25]

$$F_{y \rightarrow x} := \log \left(\frac{\Sigma_x}{\Sigma_{xx}} \right) \quad \text{and} \quad F_{x \rightarrow y} := \log \left(\frac{\Sigma_y}{\Sigma_{yy}} \right). \quad (5)$$

Granger causality from x_t to y_t implies that information from the past of x_t improve the prediction of the present of y_t compared to what can be predicted by the past of y_t alone. That is, the residual variance Σ_{yy} of the unrestricted model is required to be smaller than the residual variance Σ_y of the restricted model. Under the assumption of Gaussian-distributed residuals, $F_{y \rightarrow x}$ and $F_{x \rightarrow y}$ are asymptotically χ^2 distributed, giving rise to an analytical test of their significance [25]. An asymptotically equivalent test is given by an F-test of the goodness-of-fit of the two models (cf. [9], [4]). We refer to this approach as *standard Granger causality* (standard GC).

As variables in physical systems often mutually influence each other, it is also of interest to determine the *net* driver of the interaction by assessing whether more information is flowing from x_t to y_t then from y_t to x_t or vice versa. Following [11], [26], *net Granger causality* (Net-GC) is defined as the difference of the Granger causality scores, that is

$$F_{x \rightarrow y}^{(net)} := F_{x \rightarrow y} - F_{y \rightarrow x} \quad \text{and} \quad F_{y \rightarrow x}^{(net)} := -F_{x \rightarrow y}^{(net)}. \quad (6)$$

As the analytical distributions of these differences are unknown, statistical significance of Net-GC scores needs to be assessed using resampling methods as outlined in Section III-A.

B. Time-reversed Granger causality (TRGC)

To avoid false detections of causal interactions, Haufe et al. proposed to contrast causality measures applied to the original time series with the same measures obtained from

time-reversed signals $\tilde{z}_t := \mathbf{z}_{-t}$ [17], [16]. Here, we formalize this idea in the context of Granger causality.

Given a bivariate VAR(p) process, its time-reversed process $\tilde{\mathbf{z}}_t$ also possesses a VAR(p) representation, which we derive in Section II-D. We denote the residual covariance matrix of the time-reversed process by

$$\tilde{\Sigma} = \begin{bmatrix} \tilde{\Sigma}_{xx} & \tilde{\Sigma}_{xy} \\ \tilde{\Sigma}_{xy} & \tilde{\Sigma}_{yy} \end{bmatrix}. \quad (7)$$

The restricted AR models of the time-reversed data have a simple structure, as they are concerned with univariate time series. The autocovariance function of a univariate time series is symmetric, i.e., we have $C_x(h) = C_x(-h)$ and $C_y(h) = C_y(-h)$ for all $h \in \mathbb{Z}$. As a result of this and (19) (Section II-C), the time-reversed signals will have the same autocovariances as the original series. Because the AR representation is uniquely determined by the autocovariance function (cf. Section II-D1), they also share the same AR representation. The restricted models of the time-reversed univariate processes are thus given by

$$x_t = \sum_{k=1}^{\infty} a_k x_{t+k} + \tilde{\xi}_t^x, \quad \text{Var}(\tilde{\xi}_t^x) =: \tilde{\Sigma}_x \quad \text{and} \quad (8)$$

$$y_t = \sum_{k=1}^{\infty} b_k y_{t+k} + \tilde{\xi}_t^y, \quad \text{Var}(\tilde{\xi}_t^y) =: \tilde{\Sigma}_y \quad (9)$$

with

$$\tilde{\Sigma}_x = \Sigma_x \quad \text{and} \quad \tilde{\Sigma}_y = \Sigma_y. \quad (10)$$

In analogy to the original time series, we define the time-reversed Granger scores as

$$\tilde{F}_{\tilde{y} \rightarrow \tilde{x}} := \log \left(\frac{\tilde{\Sigma}_x}{\tilde{\Sigma}_{xx}} \right) \quad \text{and} \quad \tilde{F}_{\tilde{x} \rightarrow \tilde{y}} := \log \left(\frac{\tilde{\Sigma}_y}{\tilde{\Sigma}_{yy}} \right), \quad (11)$$

and the net Granger causality scores as

$$\tilde{F}_{\tilde{x} \rightarrow \tilde{y}}^{(net)} := \tilde{F}_{\tilde{x} \rightarrow \tilde{y}} - \tilde{F}_{\tilde{y} \rightarrow \tilde{x}} \quad \text{and} \quad \tilde{F}_{\tilde{y} \rightarrow \tilde{x}}^{(net)} := -\tilde{F}_{\tilde{x} \rightarrow \tilde{y}}^{(net)}. \quad (12)$$

Finally, the differences of the Granger scores obtained on original and time-reversed signals are given by

$$\tilde{D}_{y \rightarrow x} := F_{y \rightarrow x} - \tilde{F}_{\tilde{y} \rightarrow \tilde{x}}, \quad (13)$$

$$\tilde{D}_{x \rightarrow y} := F_{x \rightarrow y} - \tilde{F}_{\tilde{x} \rightarrow \tilde{y}}, \quad \text{and} \quad (14)$$

$$\tilde{D}_{x \rightarrow y}^{(net)} := F_{x \rightarrow y}^{(net)} - \tilde{F}_{\tilde{x} \rightarrow \tilde{y}}^{(net)}. \quad (15)$$

Time-reversed Granger causality can be applied in the following variants.

a) Conjunction-based time-reversed Granger causality (Conj-TRGC): Here, net information flow from x_t to y_t is inferred if

$$F_{x \rightarrow y}^{(net)} > 0 \quad \text{and} \quad \tilde{F}_{\tilde{x} \rightarrow \tilde{y}}^{(net)} < 0, \quad (16)$$

that is, if the directionality of net Granger causality reverses for time-reversed signals. This variant has been investigated in [18].

b) Difference-based time-reversed Granger causality (Diff-TRGC): Here, net information flow from x_t to y_t is inferred if

$$\tilde{D}_{x \rightarrow y}^{(net)} > 0, \quad (17)$$

that is, we require that net Granger causality from x_t to y_t is reduced on the time-reversed signals. Note that this is a weaker requirement than conjunction-based TRGC, as all signals for which (16) holds also fulfill (17).

c) Conjunction of Net-GC and Diff-TRGC: Finally, we can require both the time-reversed net difference and the net Granger score to be significantly larger than zero in order to infer net information flow from x_t to y_t , that is

$$\tilde{D}_{x \rightarrow y}^{(net)} > 0 \quad \text{and} \quad F_{x \rightarrow y}^{(net)} < 0. \quad (18)$$

Just as for Net-GC, statistical significance of Conj-TRGC and Diff-TRGC, as well as the combination of Net-GC and Diff-TRGC can be assessed using resampling techniques (see Section III-A).

C. Robustness of time-reversed Granger causality (TRGC)

In [16] it is pointed out that time-reversed Granger causality robustly rejects causal interpretations for mixtures of non-interacting signals such as correlated noise sources. The mathematical basis for this noise robustness property is the fact that the cross-covariance matrices $\tilde{C}_{\tilde{\mathbf{z}}}(\cdot)$ of the time-reversed signals are equal to the transposed cross-covariance matrices of the original signals, that is

$$\tilde{C}_{\tilde{\mathbf{z}}}(h) = \langle \tilde{\mathbf{z}}_t \tilde{\mathbf{z}}_{t-h}^T \rangle = \langle \mathbf{z}_t \mathbf{z}_{t+h}^T \rangle = C_{\mathbf{z}}(-h) = (C_{\mathbf{z}}(h))^T \quad (19)$$

for all $h \in \mathbb{Z}$. If a series η_t only contains a mixture of independent signals, all its cross-covariance matrices are symmetric [27]: consider $\eta_t = M s_t$ where s_t contains a number of independent sources. Then, for all $h \in \mathbb{Z}$, $C_s(h) = \text{diag}$ and thus $C_{\eta}(h) = M C_s(h) M^T$ is symmetric. For mixtures of independent noise sources, any causality measure that is solely based on a series' cross-covariance matrices therefore yields the same result on the original and the time-reversed signals. This includes Granger causality, but also other popular variants such as directed transfer function (DTF) [19] and partial directed coherence (PDC) [20]. Given sufficient amounts of data, the conditions for Conj-TRGC and Diff-TRGC cannot be fulfilled for mixtures of independent sources using these measures, preventing the detection of spurious interaction.

D. The VAR representation of a time-reversed process

There is so far no theoretical argument guaranteeing that time-reversed Granger causality correctly indicates the presence of information flow as well as its direction *in the presence of actual interaction*. In order to provide such a guarantee, we here study the time-reversal of (linear) finite-order VAR processes. Note that studying this case is sufficient since, as a results of Wold's decomposition theorem, every stationary, purely nondeterministic, process can be approximated well by a finite order VAR process [28], [1].

We start by briefly revisiting the link between cross-covariance matrices and VAR representation, which we use

throughout the paper, in Section II-D1. In Sections II-D2 and II-D3, we then review the theoretical result of Andel [29] stating that the time-reversed signal of any VAR(p) process has again a VAR(p) representation that can be expressed analytically in terms of the original process. As the description for $p > 1$ is mathematically involved, we only treat the case $p = 1$ in the main paper, while the proof for arbitrary p is presented in Appendix A.

We use these results to provide an analytic description of difference-based TRGC scores in terms of their autoregressive coefficients (Section II-E), give a minimal example (Section II-F), and prove our main result stating that, in the case of unambiguous unidirectional information flow, difference-based time-reversed Granger causality indeed yields the correct result (Section II-G).

1) *The cross-covariance function of a VAR process:* Most of the insights in this paper are based on the direct link between autoregressive coefficient matrices A_1, \dots, A_p and residual covariance matrices Σ on one hand, and cross-covariance matrix $C_z(\cdot)$ on the other hand. This link is established by the Yule-Walker equations as follows (see e.g. [1]). For a VAR(1) process

$$\mathbf{z}_t = A_1 \mathbf{z}_{t-1} + \epsilon_t, \quad (20)$$

the Yule-Walker equations read

$$C_z(0) = A_1 \cdot C_z(0) \cdot A_1^\top + \Sigma \quad \text{and} \quad (21)$$

$$C_z(h) = A_1 \cdot C_z(h-1) \quad (\forall h \in \mathbb{N} \setminus \{0\}). \quad (22)$$

Given A_1 and Σ , the cross-covariances are uniquely determined from (21) through

$$\text{vec}(C_z(0)) = (I - A_1 \otimes A_1)^{-1} \text{vec} \Sigma, \quad (23)$$

while higher-order cross-covariances $C_z(h)$ can be recursively computed using (22). Conversely, A_1 and Σ are uniquely determined by the cross-covariances through

$$A_1 = C_z(1)C_z(0)^{-1} \quad \text{and} \quad (24)$$

$$\Sigma = C_z(0) - A_1 C_z(0) A_1^\top. \quad (25)$$

Results on VAR(1) processes can typically be extended to higher-order VAR(p) processes by reducing VAR(p) processes to their VAR(1) form. The VAR(1) representation of a VAR(p) process as well as the Yule-Walker equations for general VAR(p) processes are provided in Appendix A-A.

2) *The VAR representation of a time-reversed VAR(1) process:* The time-reversed autoregressive representation of a VAR(1) process \mathbf{z}_t has been derived by Bartlett in 1955 [30]. Suppose we generate an infinite sequence of \mathbf{z}_t according to the VAR(1) process (20). The VAR representation of the *time-reversed* or *backward* process is given by

$$\mathbf{z}_t = \tilde{A}_1 \mathbf{z}_{t+1} + \tilde{\epsilon}_t, \quad (26)$$

where

$$\tilde{A}_1 = C_z(0) \cdot A_1^\top \cdot C_z(0)^{-1}, \quad (27)$$

and where the reversed residuals $\tilde{\epsilon}_t$ are calculated from \mathbf{z}_t as

$$\tilde{\epsilon}_t := \mathbf{z}_t - \tilde{A}_1 \mathbf{z}_{t+1} \quad (28)$$

with residual covariance matrix

$$\tilde{\Sigma} = \langle \tilde{\epsilon}_t \tilde{\epsilon}_t^\top \rangle = C_z(0) - C_z(0) \cdot A_1^\top \cdot C_z(0)^{-1} \cdot A_1 \cdot C_z(0). \quad (29)$$

It is easy to show that the sequence $\tilde{\epsilon}_t$ is indeed white noise, that is for all $h \in \mathbb{Z} \setminus \{0\}$: $\langle \tilde{\epsilon}_t \cdot \tilde{\epsilon}_{t+h}^\top \rangle = 0$ and for all $h \in \mathbb{N} \setminus \{0\}$: $\langle \tilde{\epsilon}_t \cdot \mathbf{z}_{t+h}^\top \rangle = 0$.

From (27), we see that the time-reversed coefficient matrix \tilde{A}_1 is similar to A_1 , and thus shares some of its properties, notably its eigenvalues, determinant, trace and rank. However, in the context of Granger causality, it is important to note that many properties of A_1 do not transfer to \tilde{A}_1 . In particular, if A_1 is triangular, diagonal, or symmetric, this is not generally the case for \tilde{A}_1 .

3) *The VAR representation of a time-reversed VAR(p) process:* The result of Bartlett on the time-reversed VAR(1) process has been generalized to VAR(p) processes by Andel in 1972 [29], in a paper that received, so far, little attention. Andel showed that any stable VAR(p) process (1) has a time-reversed representation

$$\mathbf{z}_t = \tilde{A}_1 \mathbf{z}_{t+1} + \tilde{A}_2 \mathbf{z}_{t+2} + \dots + \tilde{A}_p \mathbf{z}_{t+p} + \tilde{\epsilon}_t \quad (30)$$

that is again of order p with uniquely defined autoregressive coefficients $\tilde{A}_1, \dots, \tilde{A}_p$ and residual covariance matrix $\tilde{\Sigma}$. We reproduce this result in Appendix A-B. Note that, while we only treat bivariate VAR processes in this paper, the analytic description of the time-reversed VAR process holds for processes of arbitrary dimensionality.

E. Analytic description of Diff-TRGC

Contrasting Granger scores obtained on original with those obtained on time-reversed signals is simplified by the fact that the AR representation of a univariate time series does not depend on the direction of time. It follows immediately from (10), that the differences of the Granger scores related to original and time-reversed data do not depend on the restricted models:

$$\begin{aligned} \tilde{D}_{y \rightarrow x} &= F_{y \rightarrow x} - \tilde{F}_{\tilde{y} \rightarrow \tilde{x}} = \log \tilde{\Sigma}_{xx} - \log \Sigma_{xx} \\ \tilde{D}_{x \rightarrow y} &= F_{x \rightarrow y} - \tilde{F}_{\tilde{x} \rightarrow \tilde{y}} = \log \tilde{\Sigma}_{yy} - \log \Sigma_{yy} \\ \tilde{D}_{x \rightarrow y}^{(net)} &= F_{x \rightarrow y}^{(net)} - \tilde{F}_{\tilde{x} \rightarrow \tilde{y}}^{(net)} \\ &= (F_{x \rightarrow y} - F_{y \rightarrow x}) - (\tilde{F}_{\tilde{x} \rightarrow \tilde{y}} - \tilde{F}_{\tilde{y} \rightarrow \tilde{x}}) \\ &= \log \tilde{\Sigma}_{yy} - \log \tilde{\Sigma}_{xx} - \log \Sigma_{yy} + \log \Sigma_{xx}. \end{aligned} \quad (31)$$

The Granger score differences $\tilde{D}_{y \rightarrow x}$, $\tilde{D}_{x \rightarrow y}$, and $\tilde{D}_{x \rightarrow y}^{(net)}$ thus only depend on the residual covariance matrices of the full models of the original and time-reversed data. For the VAR(1) process, these are given in (25) and (29). For VAR(p) processes, the residual covariance matrices can be obtained through (56) and (59) as described in Appendix A-A and A-B.

Please note that while (31) implies that the unrestricted models can be neglected when computing Granger scores differences, we might gain from including them in finite sample settings. We investigate this issue through simulations in Section III.

F. A minimal example

It is not intuitive to see how the residual variance of the time-reversed process, and thus Granger causality, depends on the autoregressive coefficients of the model. Interpretation is made difficult by the occurrence of $C_{\mathbf{z}}(0)^{-1}$ in (29).

Let us therefore consider the following minimal case: a VAR(1) process \mathbf{z}_t with $C_{\mathbf{z}}(0) = I$. In that case, $C_{\mathbf{z}}(h) = A_1^h$ and $C_{\mathbf{z}}^\top(h) = (A_1^\top)^h$ for all $h \in \mathbb{Z} \setminus \{0\}$ (from (22)). All asymmetries in the cross-covariance matrices $C_{\mathbf{z}}(h)$ are thus due to asymmetries in A_1 .

Furthermore, time-reversing the signal leads to transposition of the autoregressive coefficient matrix $\tilde{A}_1 = A_1^\top$ as a result of (27). The residual covariance matrices (25) and (29) are now given by

$$\Sigma = I - A_1 \cdot A_1^\top \quad \text{and} \quad \tilde{\Sigma} = I - A_1^\top \cdot A_1.$$

Denote with $A_1 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ the autoregressive coefficients. We then have

$$\Sigma_{xx} = 1 - a_{11}^2 - a_{12}^2, \quad \tilde{\Sigma}_{xx} = 1 - a_{11}^2 - a_{21}^2,$$

and

$$\begin{aligned} \tilde{D}_{y \rightarrow x} = \log \tilde{\Sigma}_{xx} - \log \Sigma_{xx} > 0 &\Leftrightarrow \Sigma_{xx} < \tilde{\Sigma}_{xx} \\ &\Leftrightarrow a_{12}^2 > a_{21}^2. \end{aligned}$$

The difference of the Granger scores computed on the original and time-reversed time series thus indicates the correct *net* direction of information flow. We will in general not be able to infer whether x_t has a Granger-causal influence on y_t . However, we will be able to tell whether x_t Granger-causes y_t more than y_t Granger-causes x_t , or vice versa.

While this simple case will almost never occur in practice, we give theoretical guarantees for more general cases in the next section.

G. Validity of TRGC for unidirectional information flow

We now prove our main result, the validity of difference-based time-reversed Granger causality in the presence of unidirectional information flow. Consider a bivariate VAR(p) process with unambiguous unidirectional information flow. This is the case when all coefficient matrices are triangular and the residual covariance matrix Σ is diagonal. Then the following theorem holds.

Theorem 1. Let $\mathbf{z}_t = \begin{bmatrix} x_t \\ y_t \end{bmatrix} \in \mathbb{R}^2$ be a stable bivariate VAR(p) process (1) with the time-reversed representation (30). Under the assumptions

- (A1) A_1, \dots, A_p are lower triangular matrices (i.e., x_t may Granger-cause y_t , but y_t does not Granger-cause x_t), and
- (A2) Σ is a diagonal matrix, i.e. $\Sigma_{xy} = 0$ (the residuals are uncorrelated), and
- (A3) $C_{\mathbf{z}}(0)$ is invertible,

it holds that

$$\tilde{\Sigma}_{xx} \leq \Sigma_{xx}, \quad (32)$$

and that

$$\tilde{\Sigma}_{yy} \geq \Sigma_{yy}. \quad (33)$$

Corollary 1. Under assumptions (A1)–(A3), Theorem 1 and (31) immediately imply the following inequalities for the differences of Granger scores:

$$\tilde{D}_{y \rightarrow x} = F_{y \rightarrow x} - \tilde{F}_{\tilde{y} \rightarrow \tilde{x}} \leq 0 \quad (34)$$

$$\tilde{D}_{x \rightarrow y} = F_{x \rightarrow y} - \tilde{F}_{\tilde{x} \rightarrow \tilde{y}} \geq 0 \quad (35)$$

$$\tilde{D}_{x \rightarrow y}^{(net)} = (F_{x \rightarrow y} - F_{y \rightarrow x}) - (\tilde{F}_{\tilde{x} \rightarrow \tilde{y}} - \tilde{F}_{\tilde{y} \rightarrow \tilde{x}}) \geq 0. \quad (36)$$

As a result of Corollary 1, net Granger-causal information flow from x_t to y_t is reduced or remains the same when the signal is time-reversed. Thus, in the case of unambiguous unidirectional information flow, difference-based time-reversed Granger causality yields the correct result. Note that it is not true in general that the net flow between the time-reversed signals \tilde{x}_t and \tilde{y}_t , $\tilde{F}_{\tilde{x} \rightarrow \tilde{y}}^{(net)}$, is negative (reverses compared to the original series). That is, conjunction-based TRGC might in some cases incorrectly reject the presence of true causal interaction.

Corollary 1 states that each of the three difference scores, $\tilde{D}_{y \rightarrow x}$, $\tilde{D}_{x \rightarrow y}$, and $\tilde{D}_{x \rightarrow y}^{(net)}$ alone is sufficient to infer the correct directionality under assumptions (A1)–(A3). As (A1) requires information flow to be unidirectional, the individual scores $\tilde{D}_{y \rightarrow x}$ and $\tilde{D}_{x \rightarrow y}$ only indicate *net* information flow, which is what is also observed in Section II-F.

The three scores will behave differently if the assumption of uncorrelated residuals (A2) is violated. Then, $\tilde{\Sigma}_{xx} \leq \Sigma_{xx}$ and $\tilde{D}_{y \rightarrow x} \leq 0$ still hold, but the inequalities $\tilde{\Sigma}_{yy} \geq \Sigma_{yy}$, $\tilde{D}_{x \rightarrow y} \geq 0$ and $\tilde{D}_{x \rightarrow y}^{(net)} \geq 0$ do not. On average, the net difference $\tilde{D}_{x \rightarrow y}^{(net)}$ (which equals $\tilde{D}_{x \rightarrow y} - \tilde{D}_{y \rightarrow x}$) is less affected by the presence of correlations in the residuals than any of the individual scores, which is why we defined difference-based TRGC based on $\tilde{D}_{x \rightarrow y}^{(net)}$ in (17). Nevertheless, all three scores are valid measures for net information flow, as residuals should be uncorrelated if the VAR model accurately describes a physical process. The significance of uncorrelated as opposed to correlated residuals is discussed in Section IV-A.

Sketch of the proof. The first inequality (32) is relatively easy to prove. The intuition is the following: Since y_t does not Granger-cause x_t , the prediction of x_t is only based on past x_t . In contrast, the coefficient matrices $\tilde{A}_1, \dots, \tilde{A}_p$ of the time-reversed representation are in general not triangular. This means that prediction of the time-reversed signals \tilde{x}_t is not only based on past \tilde{x}_t , but can also use information from past \tilde{y}_t . We would thus expect that \tilde{x}_t can be better predicted than x_t , and that the corresponding residuals are smaller.

The proof of the second inequality (33) is more involved. The intuition is the following: we would expect that the ‘amount’ of unexplainable variance is the same for both the original and the time-reversed process. Thus, since the residual variance of x_t decreases, the residual variance of y_t should increase. Mathematically, we prove that

$$\det(\Sigma) = \det(\tilde{\Sigma}). \quad (37)$$

The proof of (37) is the only part that requires the analytic description of $\tilde{\Sigma}$, and is the main difficulty of the overall

proof. It is not straightforward, because $\tilde{\Sigma}$ depends on the inverse of the covariance matrix $C_{\mathbf{z}}(0)$, while we only have an analytic description of $\text{vec } C_{\mathbf{z}}(0)$. From (37), it is easy to infer $\tilde{\Sigma}_{yy} \leq \Sigma_{yy}$, which completes the proof. It is only in this final step that we need assumption (A2) that Σ is diagonal.

Proof (Part 1: Proof that $\tilde{\Sigma}_{xx} \leq \Sigma_{xx}$):

As A_1, \dots, A_p are lower triangular matrices (assumption (A1)), x_t is an autoregressive process of order p ,

$$x_t = a_1 x_{t-1} + \dots + a_p x_{t-p} + \xi_t^x \quad \text{with} \quad \text{Var}(\xi_t^x) = \Sigma_{xx} \quad (38)$$

Its time-reversed representation (cf. Section II-B) is

$$x_t = a_1 x_{t+1} + \dots + a_p x_{t+p} + \tilde{\xi}_t^x, \quad \text{with} \quad \text{Var}(\tilde{\xi}_t^x) = \Sigma_{xx} \quad (39)$$

Because the unrestricted (or full) model (30) extends the restricted model by including y_t , (32) follows:

$$\tilde{\Sigma}_{xx} \leq \text{Var}(\tilde{\xi}_t^x) = \Sigma_{xx}. \quad (40)$$

□

Proof (Part 2: Proof that $\tilde{\Sigma}_{yy} \leq \Sigma_{yy}$):

As mentioned in the proof sketch, we need to derive (37), the equality of the determinants $\det \Sigma$ and $\det \tilde{\Sigma}$. To improve readability, we here treat only the case $p = 1$, and derive (37) for general $p \in \mathbb{N} \setminus \{0\}$ in Appendix A-C.

The proof relies on Sylvester's determinant theorem [31], which states that for any matrices $K \in \mathbb{R}^{n \times m}$, $L \in \mathbb{R}^{m \times n}$:

$$\det(I + KL) = \det(I + LK). \quad (41)$$

We then have:

$$\begin{aligned} \det \Sigma &\stackrel{(21)}{=} \det(C_{\mathbf{z}}(0) - A_1 \cdot C_{\mathbf{z}}(0) \cdot A_1^\top) \\ &= \det(C_{\mathbf{z}}(0)) \cdot \det(I - A_1 \cdot C_{\mathbf{z}}(0) \cdot A_1^\top \cdot C_{\mathbf{z}}(0)^{-1}) \\ &\stackrel{(41)}{=} \det(C_{\mathbf{z}}(0)) \cdot \det(I - C_{\mathbf{z}}(0) \cdot A_1^\top \cdot C_{\mathbf{z}}(0)^{-1} A_1) \\ &= \det(C_{\mathbf{z}}(0) - C_{\mathbf{z}}(0) \cdot A_1^\top \cdot C_{\mathbf{z}}(0)^{-1} A_1 \cdot C_{\mathbf{z}}(0)) \\ &\stackrel{(29)}{=} \det \tilde{\Sigma}. \end{aligned}$$

From the result of *Part 1* (32), the equality of residual covariance determinants (37) (derived for general p in Appendix A-C), and assumption (A2) of uncorrelated residuals in Σ , we then obtain:

$$\begin{aligned} \Sigma_{xx} \Sigma_{yy} &\stackrel{(A2)}{=} \det \Sigma \stackrel{(37)}{=} \det \tilde{\Sigma} = \tilde{\Sigma}_{xx} \tilde{\Sigma}_{yy} - \tilde{\Sigma}_{xy} \tilde{\Sigma}_{yx} \\ &\leq \tilde{\Sigma}_{xx} \tilde{\Sigma}_{yy} \\ &\stackrel{(32)}{\leq} \Sigma_{xx} \tilde{\Sigma}_{yy} \\ &\Leftrightarrow \Sigma_{yy} \leq \tilde{\Sigma}_{yy}. \end{aligned}$$

□

Strict inequality. Let us further note that inequality (36) for difference-based TRGC is strict in the presence of causal interaction. The following theorem holds.

Theorem 2. *Under assumptions (A1)-(A3), it holds that*

$$\tilde{D}_{x \rightarrow y}^{(net)} = 0 \Leftrightarrow A_1, \dots, A_p \text{ are diagonal}. \quad (42)$$

The proof is provided in Appendix A-D. Combined with Corollary 1, Theorem 2 immediately implies that $\tilde{D}_{x \rightarrow y}^{(net)} > 0$ in the presence of unidirectional information flow from x_t

to y_t . That is, net Granger-causal information flow from x_t to y_t is truly reduced and cannot remain the same when the signal is time-reversed.

III. EXPERIMENTS

In this section, we provide an empirical investigation of model violations and other factors influencing the performance of Granger causal measures using numerical simulations. After describing the tested methods and performance measures (Section III-A), we compare several variants of TRGC in either the presence or absence of noise (Section III-B). We then investigate the influence of common drivers, various types of noise (Section III-C and III-D) and downsampling (Section III-E) on standard Granger causality and Diff-TRGC.

A. Experimental setup

We consider bivariate time series in the presence of unidirectional information flow ($x_t \rightarrow y_t$) as well as in the absence of causal interaction. Unless otherwise stated, time series of length $T = 2000$ are generated from stationary VAR(5) processes, whose autoregressive coefficients are drawn from a normal distribution with mean 0 and standard deviation $\sigma_A = 0.2$. The absence of causal interaction is modeled by setting respective AR coefficients to zero. Residuals are generated from a normal distribution with diagonal covariance matrix, whose entries are drawn from the standard uniform distribution.

We compare standard GC as well as Net-GC to Diff-TRGC (see (17)). In Section III-B, we also include Conj-TRGC (see (16)), the conjunction of Net-GC and Diff-TRGC (see (18)), and a variation of Diff-TRGC, in which $\tilde{D}_{x \rightarrow y}^{(net)}$ is computed using only the full bivariate models according to (31). This variant is denoted by *Diff-TRGC (full)*.

All statistical tests are performed at significance level $\alpha = 0.05$. For standard GC, we perform two separate F-tests, one to assess whether x_t Granger-causes y_t , and one to assess whether y_t Granger-causes x_t . It is possible that both variables are estimated to Granger-cause each other. In contrast, all other metrics indicate net directionality. We assess their statistical significance by bootstrapping residuals from the regression model: We regress \mathbf{z}_t on its past and future values $\mathbf{z}_{t-p}, \dots, \mathbf{z}_{t-1}, \mathbf{z}_{t+1}, \dots, \mathbf{z}_{t+p}$, and retain the fitted values $\hat{\mathbf{z}}_t$ and residuals $\hat{\epsilon}_t := \mathbf{z}_t - \hat{\mathbf{z}}_t$. In each bootstrap repetition, causality metrics are computed on synthetic variables $\mathbf{z}_t^* := \hat{\mathbf{z}}_t + \hat{\epsilon}_s$, where s is selected randomly for each t . Percentile confidence intervals are then constructed from the bootstrap sampling distribution. Significance is determined by evaluating if the confidence interval does not contain 0. We use 500 bootstrap samples and select the number of lags p as the optimizer of Schwarz's Bayesian Information Criterion (BIC) [32].

All experiments are repeated 300 times. In each run, a true positive (TP) is defined as a significant detection of the true direction of interaction. The *true positive rate* (TPR) is the fraction of true positives among all runs. It is here also referred to as the *sensitivity* or *power*. A false positive (FP) is defined as a significant detection of the wrong direction of interaction, or a significant detection of causal interaction in the absence

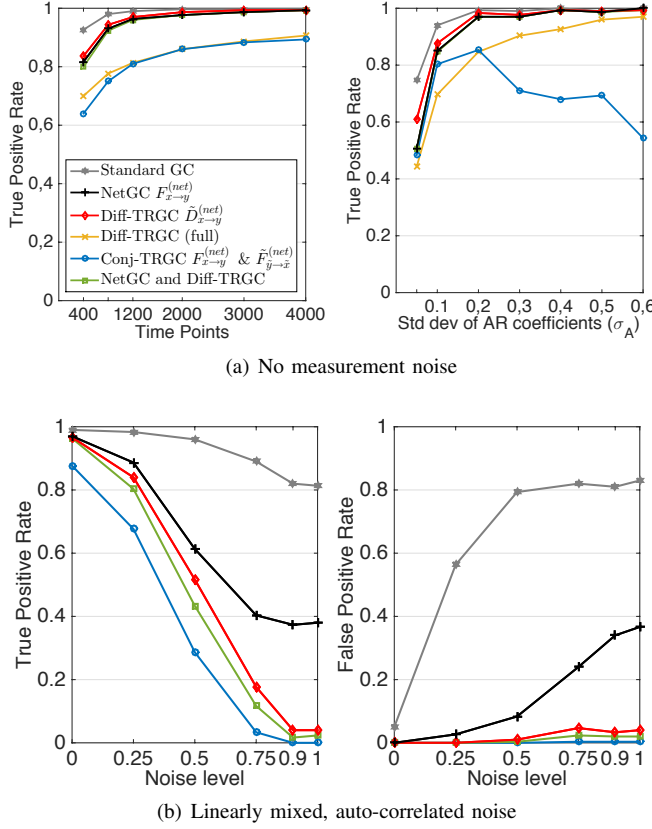


Fig. 1. Performance of Granger causality and different variants of time-reversed Granger causality (TRGC). (a) True positive rate in the noiseless case as a function of the number of samples T for fixed standard deviation $\sigma_A = 0.2$ of the AR coefficients, and as a function of σ_A for fixed $T = 2000$. (b) True and false positive rates as a function of the SNR for additive mixed autocorrelated noise (according to (43)) for $T = 2000$ and $\sigma_A = 0.2$.

of any causal interaction. The *false positive rate* (FPR) is the fraction of false positives among all tested runs.

B. Comparison of TRGC variants under interaction

We assess Granger causality and time-reversed Granger causality in the presence of unidirectional interaction considering differing sample sizes, standard deviations of the AR parameters, noise types and signal-to-noise ratios (SNR).

In a first experiment, we consider the noiseless case, and vary the sample size from 400 to 4000 for a fixed standard deviation $\sigma_A = 0.2$ of the AR coefficients. In a second experiment, we vary the standard deviation σ_A at a constant sample size of $T = 2000$. This experiment thus tests the impact of the strength of the causal connections relative to the innovation noise. The standard deviations tested are 0.05, 0.1, 0.2, ..., and 0.6. Finally, for a fixed standard deviation $\sigma_A = 0.2$, and a fixed sample size $T = 2000$, we add linearly mixed, autocorrelated measurement noise $\eta_t \in \mathbb{R}^2$ to each system according to

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = (1 - \gamma) \begin{bmatrix} x_t^{(l)} \\ y_t^{(l)} \end{bmatrix} + \gamma \cdot \eta_t, \quad (43)$$

where the subscript $^{(l)}$ denotes the underlying *latent* variables and γ defines the signal-to-noise ratio (SNR). Noise η_t is

generated by multiplying two independent AR(5) time-series with a random matrix B , with $\det(B) = 1$. We consider the signal-to-noise ratios 0, 0.25, 0.5, 0.75, 0.9 and 1.

The TP and FP rates attained in the three experiments are depicted in Figure 1. From Figure 1(a), we see that Diff-TRGC (full), which computes the difference score $\tilde{D}_{x \rightarrow y}^{(net)}$ only using the full model according to (31), seems to be suboptimal for finite samples. While we have demonstrated the equivalence of (31) to the original definition (15) for infinite samples in Section II-E, this equivalence does not hold for the finite samples studied here. Estimating residuals from the restricted models increased the power of the test for all investigated parameter settings.

Conj-TRGC has lower power relative to Diff-TRGC. This is particularly so for high σ_A , which corresponds to a dominance of the dynamical and causal aspects of the model comprised in the AR coefficients relative to the innovation noise. This result is not unexpected, as time-reversing the signals does not necessarily reverse the direction of information flow. Note that, on the other hand, Conj-TRGC is the more conservative measure compared to Diff-TRGC and could be expected to produce fewer spurious results in the presence of noise. However, as we see in Figure 1(b), both variants yield almost no spurious results in the presence of measurement noise. We will therefore use Diff-TRGC in the remaining experiments.

C. Impact of latent variables and measurement noise in the absence of causal interaction

Already Granger pointed out that standard Granger causality can lead to spurious results if not all relevant variables are incorporated in the model [8]. In a bivariate system, GC cannot determine whether the observed variables x_t and y_t are both driven by a third common cause. This argument extends to multivariate systems, if a relevant confounding variable is not part of the measurement. Furthermore, standard GC is susceptible to *measurement noise* [33], [10], [11], [26], [34], [18] and to *instantaneous linear mixing* of activity, which is a major problem for example in the analysis of electroencephalographic (EEG) recordings [13], [14], [16]. We demonstrate these effects here in additional simulations, in all of which no actual interaction occurs. We consider three different scenarios.

(A) *Linear mixing*. The observed time series x_t and y_t are a linear mixture of two independent signals $x_t^{(l)}$, $y_t^{(l)}$, that is

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = M \begin{bmatrix} x_t^{(l)} \\ y_t^{(l)} \end{bmatrix}, \quad (44)$$

where $M \in \mathbb{R}^{2 \times 2}$ denotes the mixing matrix. $x_t^{(l)}$ and $y_t^{(l)}$ were generated as two independent univariate AR(5) processes.

(B) *Common hidden cause*. The observed time series x_t and y_t are driven by a common unobserved cause g_t . Time series x_t , y_t , and g_t are generated from a three-dimensional VAR(5) model with $\sigma_A = 0.3$, in which g_t Granger-causes x_t and y_t , with no causal interaction between x_t and y_t as modeled by the respective AR coefficients being set to zero.

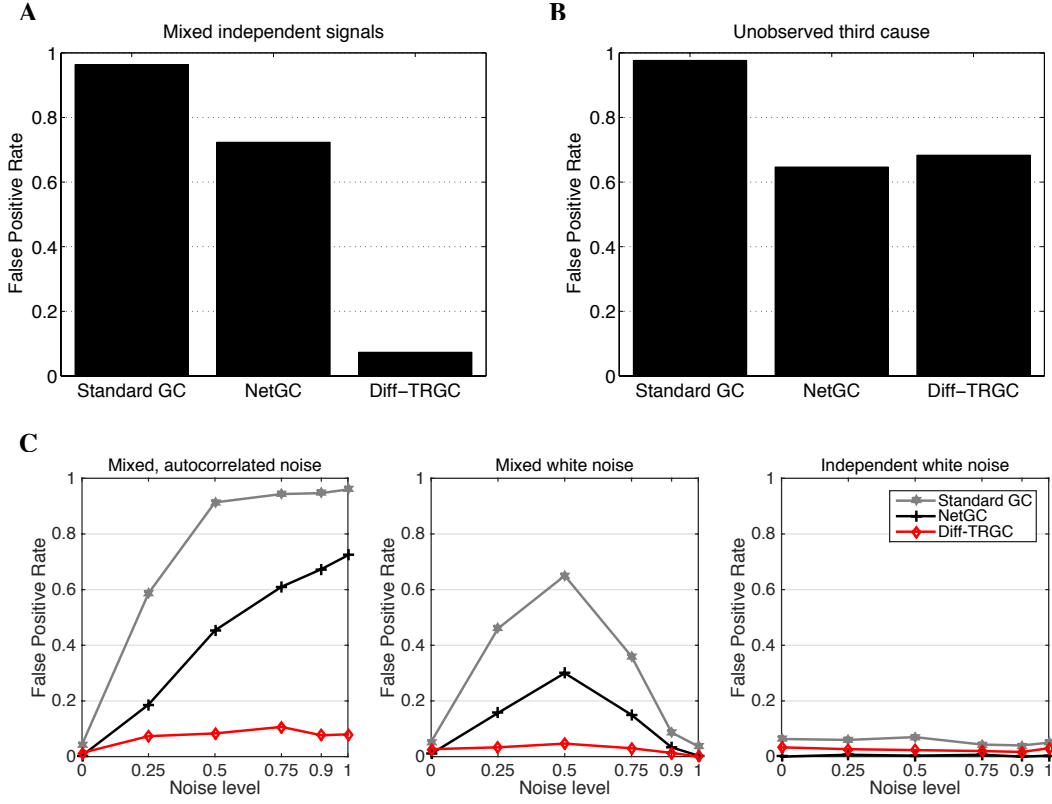


Fig. 2. False positive rates of Granger causality (standard GC and Net-GC) and difference-based time-reversed Granger causality (Diff-TRGC) as a function of the SNR for two signals lacking any causal connection. (A) Instantaneous linear mixture of two independent univariate AR(5) processes. (B) Common unobserved cause. x_t and y_t . (C) Superposition of two independent univariate AR(5) processes with additive Gaussian noise.

(C) *Additive noise.* The observed time series x_t and y_t are a superposition of two independent univariate AR(5) processes $x_t^{(l)}$, $y_t^{(l)}$ and additive noise η_t as in (43), with $\gamma \in \{0, 0.25, 0.5, 0.75, 0.9, 1\}$ adjusting the SNR. We consider three different types of noise. *Independent white noise* is generated from a normal distribution with diagonal covariance matrix, whose entries are drawn from the standard uniform distribution. *Mixed white noise* is created by multiplying independent noise with a random matrix B with $\det(B) = 1$. *Mixed autocorrelated noise* is created by multiplying two independent AR(5) time-series with B .

Figure 2 illustrates the behavior of standard Granger causality, Net-GC and Diff-TRGC in the various simulation settings. Values on the y-axis indicate the FP rate at significance level $\alpha = 0.05$. As all experiments are characterized by the absence of any interaction between x_t and y_t , any significant detection of information flow either from x_t to y_t or y_t to x_t is counted as a false positive.

It is apparent from Figure 2 that standard GC and Net-GC lead to spurious detection of causality in all tested scenarios. Their behavior in the presence of noise (panel C) depends on the properties of that noise. Mixed noise (left and center plots of panel C) is generally very problematic, especially if it is also autocorrelated (left part). As x_t and y_t are already independent, adding independent noise (obviously) does not pose a problem here (right part of panel C).

In contrast to standard GC and Net-GC, time-reversed Granger causality implemented through Diff-TRGC is insen-

sitive to mixtures of independent sources regardless of their spatial and temporal correlation structure (see panels A and C). This behavior thus reflects its known theoretical properties discussed in Section II-C. The presence of a hidden common confounder, however, cannot be ruled out by using time-reversed Granger causality (panel B).

D. Impact of noise in the presence of causal interaction

We further study the behavior of standard GC, Net-GC and Diff-TRGC in the presence of unidirectional causal interactions superimposed with noise. Four different scenarios are considered. In all cases, data are generated according to (43) with $x_t^{(l)}$ Granger-causing $y_t^{(l)}$. In the first three scenarios, (A-C), interacting signals from bivariate VAR(5) models are superimposed with noise. As in Section III-C, we use mixed autocorrelated noise (scenario A), mixed white noise (B), and independent white noise (C). The same signal to noise ratios as in Section III-C are used.

In the fourth scenario, (D), we simulate the following VAR(1) process with long memory:

$$\begin{aligned} \begin{bmatrix} x_t^{(l)} \\ y_t^{(l)} \end{bmatrix} &= \begin{bmatrix} 0.95 & 0 \\ 1 & 0.5 \end{bmatrix} \begin{bmatrix} x_{t-1}^{(l)} \\ y_{t-1}^{(l)} \end{bmatrix} + \epsilon_t \quad \epsilon_t \sim \mathcal{N}(0, I) \\ x_t &= (1 - \gamma) \cdot x_t^{(l)} + \gamma \cdot \eta_t \quad \eta_t \sim \mathcal{N}(0, 1) \\ y_t &= y_t^{(l)}, \end{aligned} \quad (45)$$

adopted from [26], where \mathcal{N} denotes the normal distribution.

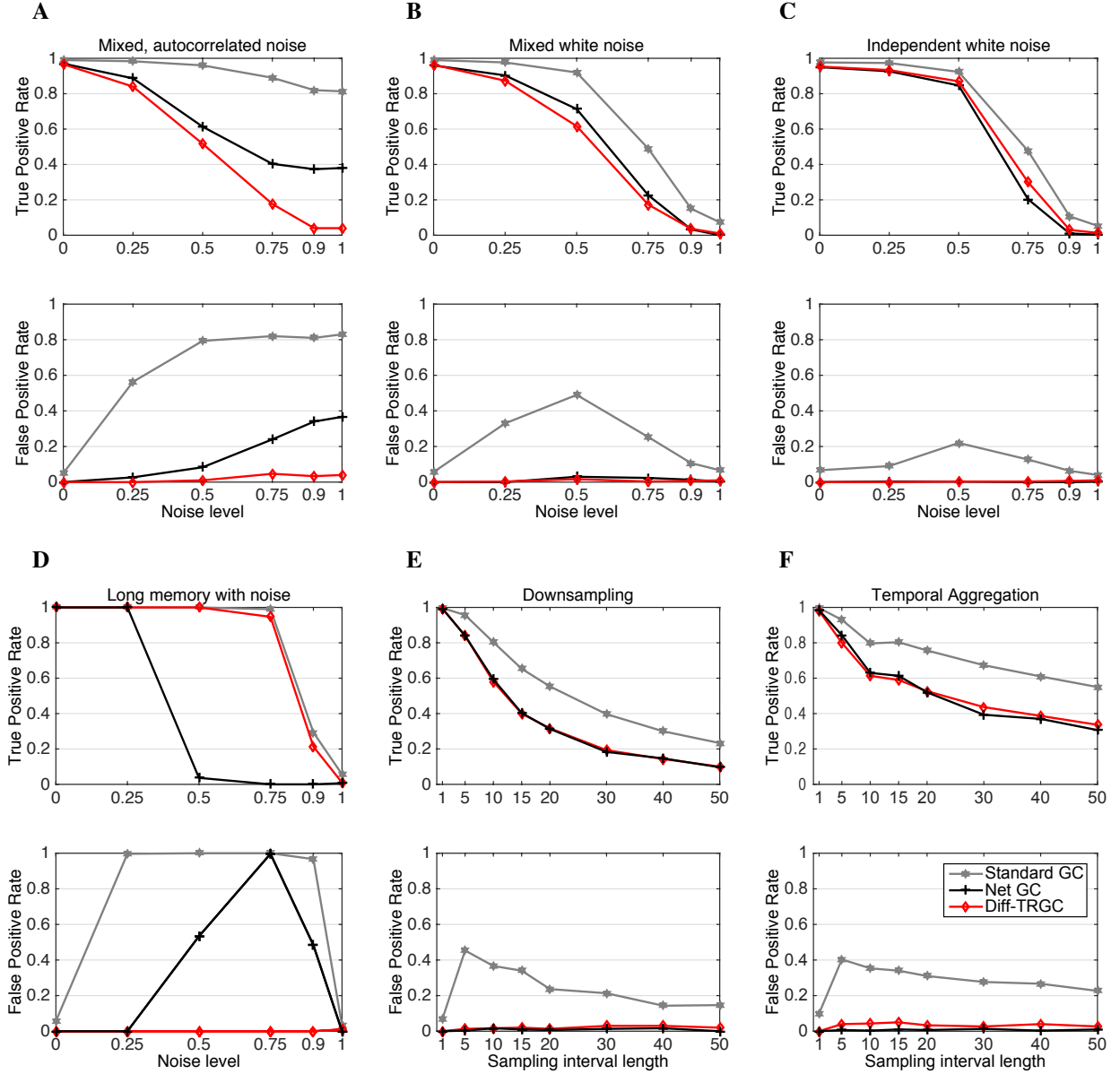


Fig. 3. Performance of Granger causality (standard GC and Net-GC) and difference-based time-reversed Granger causality (Diff-TRGC) for two signals with unidirectional information flow from x_t to y_t . Shown are the fractions of true positives ($x_t \rightarrow y$ detected) and false positives ($y_t \rightarrow x_t$ detected), when x_t and y_t are corrupted by noise (A-D), downsampling (E), and temporal aggregation (F). The underlying latent signals $x^{(l)}$ and $y^{(l)}$ were generated from VAR(5) processes with random AR coefficients, except for D, in which signals follow a VAR(1) process with long memory according to (45).

True positive and false positive rates as estimated from 300 simulation runs are reported in Figure 3 (A-D). Just as in the absence of causality (cf. Section III-C), we observe that linearly mixed, autocorrelated noise leads to the highest numbers of false detections for standard GC, while independent white noise leads to lowest FP rates. Diff-TRGC is characterized by negligible amounts of false positives in all cases at the cost of slightly decreased sensitivity as compared to standard GC in scenarios (A-C). Interestingly, Net-GC behaves very similar to Diff-TRGC in the presence of non-autocorrelated noise both in terms of sensitivity and specificity (B-C). In these settings, spurious causality could already be almost entirely eliminated by testing for net Granger causality. This result, however, does not imply that Net-GC cannot be affected by

non-autocorrelated noise in general. A counterexample is the system with long memory studied in scenario (D). Here, Net-GC (as well as standard GC) fails, because y_t contains delayed but cleaner information about $x_t^{(l)}$ than x_t itself and thus may help to predict future x_t . Diff-TRGC, however, robustly identifies x_t as the driver.

Our examples show time-reversed Granger causality almost completely eliminates spurious causalities arising from any kind of additive noise. At the same time, it exhibits similar statistical power as net Granger causality. We also observe that net Granger causality is typically more robust with respect to additive noise than standard Granger causality.

E. Impact of downsampling and temporal aggregation

Spurious Granger causality has also been reported to arise due to downsampling and temporal aggregation [35], [36], [37], posing serious problems, for example, in functional magnetic resonance imaging (fMRI) [38], [39].

We generate data using a VAR(5) model with random coefficients with $\sigma_A = 0.3$, in which x_t Granger-causes y_t . These data are decimated at different factors τ in two ways. In the downsampling scenario (E), causal measures are applied to time series of length $T = 2000$ constructed from the original time series by skipping $\tau - 1$ time points in between sampled data points. In the temporal aggregation scenario (F), time series of length $T = 2000$ are constructed from the original time series by averaging over τ data points. No noise was added.

Figure 3 (E-F) depicts TP and FP rates attained in the two scenarios as a function of τ . We see that Net-GC and Diff-TRGC are more robust than standard GC. Both Net-GC and Diff-TRGC did not result in spurious causality.

IV. DISCUSSION

We established the theoretical guarantee that difference-based time-reversed Granger causality (Diff-TRGC) indicates the correct direction of causality in bivariate autoregressive processes characterized by unambiguous unidirectional information flow. Our results complement previous work by [16], [17] showing that TRGC in general correctly rejects causal interpretations for mixtures of non-interacting sources (thus, in the absence of any causality). While further compelling intuitive ideas for robust causality measures have been presented [11], [23], [18], our result provides, to the best of our knowledge, the first proof of the correctness of one of such techniques (Diff-TRGC) for a relatively general class of time-series models.

Our theory is accompanied by simulations, in which we confirmed that time-reversed Granger causality robustly detects the presence of true causal interactions in various realistic scenarios including mixed noise and downsampling. We showed that Diff-TRGC is able to infer correct directionality with similar power as net Granger causality, while at the same time producing fewer (in most cases, negligible amounts of) false alarms than Net-GC and standard GC. We therefore suggest to use Diff-TRGC whenever the data under study are likely to be corrupted by noise.

A. Correlated residuals

To define an unambiguous uni-directional information flow, our theory assumes uncorrelated residuals, as is common in the literature. Correlated residuals indicate instantaneous effects that the variables exert on each other. While we would not expect correlated residuals if the VAR model accurately describes the data generating process, such effects are likely to occur in practice (e.g., if the sampling rate of the acquired data falls below the time scale of the causal interactions). They pose severe problems for causal estimation, because they can be explained by several possible data generating models,

the coefficients of which cannot be uniquely identified using second order information only.

Data generating models. Instantaneous interactions can be modeled implicitly through correlated residuals in classical VAR processes, or explicitly, for example using so-called ‘structural’ VAR (SVAR) processes [1], [40], [41]. By augmenting the VAR model with an instantaneous mixing matrix Γ_0 , the SVAR model

$$\mathbf{z}_t = \sum_{h=0}^p \Gamma_h \mathbf{z}_{t-h} + \bar{\epsilon}_t, \quad (46)$$

achieves that the residuals $\bar{\epsilon}_t$ are uncorrelated. Here, the diagonal of Γ_0 is assumed to be zero.

Correlated residuals emerge naturally in electrophysiological neuroimaging data, where the signals observable at the sensors (e.g., EEG electrodes) are a linear mixture of the latent activity of possibly interacting neuronal populations within the brain. A model for such mixtures of potentially interacting sources is given by

$$\mathbf{z}_t = M \mathbf{z}_t^{(l)}, \quad \mathbf{z}_t^{(l)} = \sum_{h=1}^p B_h \mathbf{z}_{t-h}^{(l)} + \epsilon_t, \quad (47)$$

where $\mathbf{z}_t \in \mathbb{R}^d$ denotes the observed data, $\mathbf{z}_t^{(l)} \in \mathbb{R}^d$ denotes the activity of underlying latent variables (e.g., brain sources) following a VAR(p) process with uncorrelated residuals ϵ_t , and $M \in \mathbb{R}^{d \times d}$ is an unknown mixing matrix (representing, e.g., the volume conduction effect of the human head). We call (47) the ‘mixture of interacting sources’ model.

Note that VAR models with correlated residuals, SVAR models, and mixture of interacting sources models can be used interchangeably to represent the same statistical process. For example, an interacting sources model (47) can be equivalently written as a VAR(p) process (1) with coefficients

$$A_h = M B_h M^{-1}, \quad h \in \{1, \dots, p\} \quad (48)$$

and correlated residuals $\epsilon_t = M \bar{\epsilon}_t$. Likewise, an SVAR(p) process (46) can be converted into a VAR(p) process with correlated residuals $\epsilon_t = (I - \Gamma_0)^{-1} \bar{\epsilon}_t$ and coefficients

$$A_h = (I - \Gamma_0)^{-1} \Gamma_h, \quad h \in \{1, \dots, p\}. \quad (49)$$

The reverse transformations from VAR models to SVAR or interacting source models, as well as the transformations between SVAR and interacting source models, are, however, not unique (see *Model identifiability*).

Ambiguous causal interpretations can emerge in cases where one of the three models indicates time-delayed causal interactions through non-zero off-diagonal coefficients in the A_h , B_h or Γ_h , while another one does not. This ambiguity can in general only be resolved if the model generating the data is known a-priori. In case of EEG data, for example, (47) reflects the true data-generating process. Therefore, only the parameters B_h of the source VAR process (47) permit meaningful causal interpretation (wrt. to the source variables $\mathbf{z}_t^{(l)}$), while, for example, the VAR parameters in (48) are distorted by the mixing matrix M .

Model identifiability. A further complication in the presence of instantaneous effects in the data is that for mixture of interacting sources as well as SVAR models, the parameters are not uniquely defined from second order information only. This can be best seen for the latter model (47). Identifying the model parameters requires the estimation of a full factorization of the data into a mixing matrix M and source time series $\mathbf{z}_t^{(l)}$. This means that the estimation problem falls into the blind source separation (BSS) setting, in which Gaussianity of the factors is not sufficient for their identification. The classical approach to BSS, independent component analysis (ICA) assumes statistical independence and non-Gaussianity of the sources $\mathbf{z}_t^{(l)}$ to ensure identifiability. This concept can be adopted in the context of source AR models by enforcing independence/non-Gaussianity of the residuals of the source AR process in (47) [13], [15], [42]. In a similar way, independence of residuals has been used in the identification of SVAR models [40], [43].

Example. Consider the following VAR(1) process with correlated residuals:

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 0.7 & 0 \\ -0.12 & 0.9 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \epsilon_t, \quad \langle \epsilon_t \epsilon_t^\top \rangle = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}.$$

This process can also be represented by the SVAR(1) model

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0.6 & 0 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} + \begin{bmatrix} 0.7 & 0 \\ -0.54 & 0.9 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \bar{\epsilon}_t$$

as well as the mixtures of interacting sources model

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0.6 & 0.8 \end{bmatrix} \begin{bmatrix} x_t^{(l)} \\ y_t^{(l)} \end{bmatrix}, \quad \begin{bmatrix} x_t^{(l)} \\ y_t^{(l)} \end{bmatrix} = \begin{bmatrix} 0.7 & 0 \\ 0 & 0.9 \end{bmatrix} \begin{bmatrix} x_{t-1}^{(l)} \\ y_{t-1}^{(l)} \end{bmatrix} + \epsilon_t^{(l)},$$

with uncorrelated residuals $\langle \bar{\epsilon}_t \bar{\epsilon}_t^\top \rangle = \begin{bmatrix} 1 & 0 \\ 0 & 0.64 \end{bmatrix}$, $\langle \epsilon_t^{(l)} \epsilon_t^{(l)\top} \rangle = I$. Note that both the VAR(1) and the SVAR(1) representation indicate unidirectional causal interaction between the observed variables x_t and y_t , whereas the mixture model suggests that the observed data can also arise from a mixture of two independent latent sources $x_t^{(l)}$ and $y_t^{(l)}$. However, another equivalent mixture model

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} -\sqrt{0.2} & \sqrt{0.8} \\ \sqrt{0.2} & \sqrt{0.8} \end{bmatrix} \begin{bmatrix} x_t^{(l)} \\ y_t^{(l)} \end{bmatrix}, \quad \begin{bmatrix} x_t^{(l)} \\ y_t^{(l)} \end{bmatrix} = \begin{bmatrix} 0.86 & 0.08 \\ 0.08 & 0.74 \end{bmatrix} \begin{bmatrix} x_{t-1}^{(l)} \\ y_{t-1}^{(l)} \end{bmatrix} + \tilde{\epsilon}_t$$

with $\langle \tilde{\epsilon}_t \tilde{\epsilon}_t^\top \rangle = I$ suggests bidirectional informational flow on the source level. Similarly, the following SVAR(1) model indicates bidirectional flow

$$\begin{bmatrix} x_t \\ y_t \end{bmatrix} = \begin{bmatrix} 0 & 0.6 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} + \begin{bmatrix} 0.772 & -0.54 \\ -0.12 & 0.9 \end{bmatrix} \begin{bmatrix} x_{t-1} \\ y_{t-1} \end{bmatrix} + \bar{\epsilon}_t, \quad \langle \bar{\epsilon}_t \bar{\epsilon}_t^\top \rangle = \begin{bmatrix} 0.64 & 0 \\ 0 & 1 \end{bmatrix}.$$

B. Future work

Further effort is required to investigate the behavior of TRGC in the presence of bidirectional information flow. Also, our theoretical analysis only covers the bivariate framework. Both Granger causality and TRGC can result in spurious causality when relevant variables are not included (cf. Fig. 2, panel B). Therefore, an extension of the analysis of time-reversal to general multivariate signals would be very interesting. Furthermore, it would be desirable to obtain theoretical guarantees for the performance of TRGC in the presence of true interaction superimposed by noise in the form of bounds on the false positive rate. A major difficulty here is to obtain the residual covariance of the superposition of a VAR process and additive noise. Analytically computing Granger causality in the presence of noise is mathematically involved even for special cases [44].

Finally, [16] showed that for any causality measure based on cross-covariances, differences of the scores obtained on the original and time-reversal signals correctly indicate the absence of causality on mixtures of independent sources. While we focused here on Granger causality, it remains to be shown whether validity of time-reversal in the presence of causal interaction can also be demonstrated for other causality measures.

APPENDIX A PROOFS FOR VAR(p)

A. The VAR(p) process and its cross-covariance function

Consider a stable bivariate VAR(p) process, $\mathbf{z}_t \in \mathbb{R}^2$, as defined in (1),

$$\mathbf{z}_t = A_1 \mathbf{z}_{t-1} + A_2 \mathbf{z}_{t-2} + \dots + A_p \mathbf{z}_{t-p} + \epsilon_t,$$

where $\epsilon_t \in \mathbb{R}^2$ is a 2-dimensional white noise process (i.e. $\langle \epsilon_t \rangle = 0$, $\langle \epsilon_t \epsilon_{t-h}^\top \rangle = 0$ for $h \in \mathbb{Z} \setminus \{0\}$, and $\langle \epsilon_t \mathbf{z}_{t-h}^\top \rangle = 0$ for $h \in \mathbb{N} \setminus \{0\}$) with residual covariance matrix $\Sigma = \langle \epsilon_t \epsilon_t^\top \rangle$.

Many results on VAR(1) processes can be extended to higher order VAR(p) processes by considering their VAR(1) form. Given the 2-dimensional VAR(p) process \mathbf{z}_t , the corresponding $2p$ -dimensional VAR(1) representation is defined as

$$Z_t = \mathbf{A} Z_{t-1} + E_t, \quad (50)$$

with

$$Z_t = \begin{bmatrix} \mathbf{z}_t \\ \mathbf{z}_{t-1} \\ \vdots \\ \mathbf{z}_{t-p+1} \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} A_1 & A_2 & \dots & A_{p-1} & A_p \\ I & 0 & \dots & 0 & 0 \\ 0 & I & \dots & 0 & 0 \\ & & \ddots & & \\ 0 & 0 & \dots & I & 0 \end{bmatrix}, \quad E_t = \begin{bmatrix} \epsilon_t \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

and residual covariance matrix

$$\Sigma_E = \langle E_t E_t^\top \rangle = \begin{bmatrix} \Sigma & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}. \quad (51)$$

The cross-covariances of Z_t are linked to the cross-covariances of \mathbf{z}_t through

$$C_Z(h) = \begin{bmatrix} C_{\mathbf{z}}(h) & C_{\mathbf{z}}(h+1) & \dots & C_{\mathbf{z}}(h+p-1) \\ C_{\mathbf{z}}(h-1) & C_{\mathbf{z}}(h) & \dots & C_{\mathbf{z}}(h+p-2) \\ \vdots & \vdots & \ddots & \vdots \\ C_{\mathbf{z}}(h-p+1) & C_{\mathbf{z}}(h-p+2) & \dots & C_{\mathbf{z}}(h) \end{bmatrix} \quad (52)$$

for all $h \in \mathbb{Z}$. The Yule-Walker equations can then be expressed as

$$C_Z(0) = \mathbf{A} \cdot C_Z(0) \cdot \mathbf{A}^\top + \Sigma_E \quad \text{and} \quad (53)$$

$$C_Z(h) = \mathbf{A} \cdot C_Z(h-1) \quad \forall h \in \mathbb{N} \setminus \{0\}. \quad (54)$$

Given A_1, \dots, A_p , and Σ , the cross-covariances are uniquely determined: Equation (53) implies that $\text{vec}(C_Z(0)) = (I - \mathbf{A} \otimes \mathbf{A})^{-1} \text{vec} \Sigma_E$, while $C_Z(h)$ for $h > 1$ can be recursively computed using (54). Conversely, A_1, \dots, A_p and Σ are uniquely determined by the cross-covariances through

$$[A_1, A_2, \dots, A_p] = [C_Z(1), C_Z(2), \dots, C_Z(p)] \cdot C_Z(0)^{-1} \quad (55)$$

and

$$\Sigma = C_Z(0) - [A_1, A_2, \dots, A_p] \cdot C_Z(0) \cdot [A_1, A_2, \dots, A_p]^\top. \quad (56)$$

B. The time-reversed VAR(p) process

The results of Bartlett on the analytical description of time-reversed VAR(1) processes have been generalized to VAR(p) processes by Andel in 1972 [29]. Given a 2-dimensional VAR(p) process \mathbf{z}_t as in (1), Andel considers a second VAR(p) process

$$\tilde{\mathbf{z}}_t = \tilde{A}_1 \tilde{\mathbf{z}}_{t-1} + \tilde{A}_2 \tilde{\mathbf{z}}_{t-2} + \dots + \tilde{A}_p \tilde{\mathbf{z}}_{t-p} + \mathbf{e}_t, \quad (57)$$

where \mathbf{e}_t is white noise with covariance matrix $\tilde{\Sigma} = \langle \mathbf{e}_t \mathbf{e}_t^\top \rangle$.

Now, denote with $Q := C_Z(0)^{-1}$ the inverse of the covariance of Z_t , with block matrix notation

$$Q =: (Q_{lk})_{l,k=1}^p = \begin{bmatrix} Q_{1,1} & Q_{1,2} & \dots & Q_{1,p} \\ Q_{2,1} & Q_{2,2} & \dots & Q_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{p,1} & Q_{p,p-1} & \dots & Q_{p,p} \end{bmatrix} \in \mathbb{R}^{2p \times 2p},$$

where Q_{lk} are 2×2 blocks.

Andel proves that $C_{\tilde{\mathbf{z}}}(h) = C_Z(-h)$ for all $h \in \mathbb{Z}$ (that is, $\tilde{\mathbf{z}}_t$ has the same cross-covariance matrices as \mathbf{z}_t reversed in time), if and only if $\tilde{A}_1, \dots, \tilde{A}_p$ and $\tilde{\Sigma}$ are defined as follows: for $1 \leq j \leq p$,

$$\tilde{A}_j = -(Q_{pp} + A_p^\top \Sigma^{-1} A_p)^{-1} (Q_{p,p-j} + A_p^\top \Sigma^{-1} A_{p-j}) \quad (58)$$

and

$$\tilde{\Sigma} = (Q_{pp} + A_p^\top \Sigma^{-1} A_p)^{-1}, \quad (59)$$

where $Q_{p,0} := 0$ and $A_0 := -I$. Andel further proves that $\tilde{A}_p \neq 0$, if and only if $A_p \neq 0$, and that, if \mathbf{z}_t is stable, so is $\tilde{\mathbf{z}}_t$.

Note that, while we only treat bivariate VAR processes in this paper, the analytic description reviewed above holds for arbitrary dimensionality.

C. Proof that $\det(\Sigma) = \det(\tilde{\Sigma})$ for general p – this completes the proof of Theorem 1

Given Andel's result, we can complete the proof for Theorem 1. The only missing part of the proof (cf. Section II-G) is the proof of (37), $\det(\Sigma) = \det(\tilde{\Sigma})$, for arbitrary $p \in \mathbb{N} \setminus \{0\}$.

Preliminaries. We will make use of the following well-known equalities. Let K be a positive definite matrix with $L = K^{-1}$, and let

$$K = \begin{bmatrix} K_{1,1} & K_{1,2} \\ K_{2,1} & K_{2,2} \end{bmatrix}, \quad L = \begin{bmatrix} L_{1,1} & L_{1,2} \\ L_{2,1} & L_{2,2} \end{bmatrix}$$

be the block matrix notations of L and K , where $K_{1,1}$ is a square matrix of the same size as $L_{1,1}$. Then the standard Schur complement formula (e.g. [45], Theorem 2.7) is given as

$$L_{2,2} = [K_{2,2} - K_{2,1} K_{1,1}^{-1} K_{1,2}]^{-1}. \quad (60)$$

Let T and W be invertible matrices, then for all matrices U and V of fitting size

$$\det(T + UWV) = \det(W^{-1} + VT^{-1}U) \det(T) \det(W). \quad (61)$$

This relation is known as the generalized *matrix determinant lemma* and a straightforward extension of Sylvester's determinant theorem (41).

Let K be a matrix with block notation as above, and $K_{1,1}$ be invertible, then, (see e.g. [45], Theorem 2.1)

$$\det(K) = \det(K_{1,1}) \det(K_{2,2} - K_{2,1} K_{1,1}^{-1} K_{1,2}). \quad (62)$$

Let us also introduce the following notation for the blocks of $C_Z(0)$:

$$C_Z(0) = \begin{bmatrix} C_{Z \setminus p} & R^\top \\ \bar{R} & C_{Z \setminus p} \end{bmatrix} = \begin{bmatrix} C_{\mathbf{z}}(0) & \bar{R} \\ \bar{R}^\top & C_{Z \setminus p} \end{bmatrix},$$

where we define

$$R := [C_{\mathbf{z}}(p-1)^\top \ C_{\mathbf{z}}(p-2)^\top \ \dots \ C_{\mathbf{z}}(1)^\top] \in \mathbb{R}^{2 \times 2(p-1)}$$

$$\bar{R} := [C_{\mathbf{z}}(1) \ \dots \ C_{\mathbf{z}}(p-1)] \in \mathbb{R}^{2 \times 2(p-1)},$$

and

$$C_{Z \setminus p} := \begin{bmatrix} C_{\mathbf{z}}(0) & \dots & C_{\mathbf{z}}(p-2) \\ \vdots & \ddots & \vdots \\ C_{\mathbf{z}}(2-p) & \dots & C_{\mathbf{z}}(0) \end{bmatrix} \in \mathbb{R}^{2(p-1) \times 2(p-1)}.$$

Step 1: Analytic expression for $C_{\mathbf{z}}(0) - \bar{R} C_{Z \setminus p}^{-1} \bar{R}^\top$.

We first prove that

$$C_{\mathbf{z}}(0) - \bar{R} C_{Z \setminus p}^{-1} \bar{R}^\top = \Sigma + A_p Q_{pp}^{-1} A_p^\top, \quad (63)$$

that is, the residual variance when regressing z_t on $z_{t-1}, \dots, z_{t-(p-1)}$ given by $C_{\mathbf{z}}(0) - \bar{R} C_{Z \setminus p}^{-1} \bar{R}^\top$ can be expressed as the sum of $A_p Q_{pp}^{-1} A_p^\top$ and the residual variance when regressing z_t on $z_{t-1}, \dots, z_{t-(p-1)}, z_{t-p}$, given by Σ .

Recall the Yule-Walker equation (53)

$$C_Z(0) = \mathbf{A} \cdot C_Z(0) \cdot \mathbf{A}^\top + \Sigma_E.$$

and let us rewrite

$$\Sigma_E = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{A} = \begin{bmatrix} \mathbf{A}_{\setminus p} & A_p \\ I & 0 \end{bmatrix},$$

where we define

$$\mathbf{A}_{\setminus p} := [A_1 \ \dots \ A_{p-1}] \in \mathbb{R}^{2 \times 2(p-1)}.$$

The Yule-Walker equation can then be written in blocks as

$$\begin{bmatrix} C_{\mathbf{z}}(0) & \bar{R} \\ \bar{R}^\top & C_{Z \setminus p} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{\setminus p} & A_p \\ I & 0 \end{bmatrix} \begin{bmatrix} C_{Z \setminus p} & R^\top \\ \bar{R} & C_{\mathbf{z}}(0) \end{bmatrix} \begin{bmatrix} \mathbf{A}_{\setminus p}^\top & I \\ A_p^\top & 0 \end{bmatrix} + \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}.$$

We see from the top line that

$$\begin{aligned} \bar{R} &= \mathbf{A}_{\setminus p} C_{Z \setminus p} + A_p R \\ \Leftrightarrow A_p R &= \bar{R} - \mathbf{A}_{\setminus p} C_{Z \setminus p}, \end{aligned} \quad (64)$$

and that

$$\begin{aligned} C_{\mathbf{z}}(0) &= \mathbf{A}_{\setminus p} C_{Z \setminus p} \mathbf{A}_{\setminus p}^\top + \mathbf{A}_{\setminus p} R^\top A_p^\top + A_p R \mathbf{A}_{\setminus p}^\top + A_p C_{\mathbf{z}}(0) A_p^\top + \Sigma \\ &\stackrel{(64)}{=} -\mathbf{A}_{\setminus p} C_{Z \setminus p} \mathbf{A}_{\setminus p}^\top + \mathbf{A}_{\setminus p} \bar{R}^\top + \bar{R} \mathbf{A}_{\setminus p}^\top + A_p C_{\mathbf{z}}(0) A_p^\top + \Sigma \end{aligned} \quad (65)$$

from which we conclude that

$$\begin{aligned} \Sigma + A_p Q_{pp}^{-1} A_p^\top &\stackrel{(60)}{=} \Sigma + A_p [C_{\mathbf{z}}(0) - \bar{R} C_{Z \setminus p}^{-1} \bar{R}^\top] A_p^\top \\ &= \Sigma + A_p C_{\mathbf{z}}(0) A_p^\top - \bar{R} C_{Z \setminus p}^{-1} \bar{R}^\top A_p^\top \\ &\stackrel{(64)}{=} \Sigma + A_p C_{\mathbf{z}}(0) A_p^\top - [\bar{R} - \mathbf{A}_{\setminus p} C_{Z \setminus p}] C_{Z \setminus p}^{-1} [\bar{R} - \mathbf{A}_{\setminus p} C_{Z \setminus p}]^\top \\ &= \Sigma + A_p C_{\mathbf{z}}(0) A_p^\top - \bar{R} C_{Z \setminus p}^{-1} \bar{R}^\top + \mathbf{A}_{\setminus p} \bar{R}^\top + \bar{R} \mathbf{A}_{\setminus p}^\top - \mathbf{A}_{\setminus p} C_{Z \setminus p} \mathbf{A}_{\setminus p}^\top \\ &\stackrel{(65)}{=} C_{\mathbf{z}}(0) - \bar{R} C_{Z \setminus p}^{-1} \bar{R}^\top. \end{aligned}$$

Step 2. Derive $\det \Sigma = \det \tilde{\Sigma}$ from Andel and (63).

From Andel (59) we know that $\tilde{\Sigma} = (Q_{pp} + A_p^\top \Sigma^{-1} A_p)^{-1}$. As Q is positive definite, Q_{pp} is invertible such that

$$\frac{1}{\det(\tilde{\Sigma})} = \det(Q_{pp} + A_p^\top \Sigma^{-1} A_p) \stackrel{(61)}{=} \det(\Sigma + A_p Q_{pp}^{-1} A_p^\top) \frac{\det(Q_{pp})}{\det(\Sigma)}.$$

It therefore suffices to show that

$$\det(\Sigma + A_p Q_{pp}^{-1} A_p^\top) = \det(Q_{pp}^{-1}). \quad (66)$$

Drawing on Step 1, Equation (66) can be proven as follows:

$$\begin{aligned} \det(\Sigma + A_p Q_{pp}^{-1} A_p^\top) &= \det(Q_{pp}^{-1}) \\ \stackrel{(63),(60)}{\Leftrightarrow} \det(C_Z(0) - \bar{R} C_{Z \setminus p}^{-1} \bar{R}^\top) &= \det(C_Z(0) - R C_{Z \setminus p}^{-1} R^\top) \\ \stackrel{(62)}{\Leftrightarrow} \det \left(\begin{bmatrix} C_{Z \setminus p} & \bar{R}^\top \\ \bar{R} & C_Z(0) \end{bmatrix} \right) \det(C_{Z \setminus p}^{-1}) &= \det(C_Z(0)) \det(C_{Z \setminus p}^{-1}) \\ \Leftrightarrow \det \left(\begin{bmatrix} C_{Z \setminus p} & \bar{R}^\top \\ \bar{R} & C_Z(0) \end{bmatrix} \right) &= \det(C_Z(0)). \end{aligned}$$

Switching rows or columns of a matrix leaves its determinant invariant up to a factor of $(-1)^{i+j}$, where i and j are corresponding row or column indices. In the following, we perform block-wise rotation of a matrix block to the bottom, and to right, respectively. This is a concatenation of several row and column switches giving us a factor $(-1)^r$ for a given r . Note that r is the same for both operations due to their symmetric behavior. Therefore we have

$$\begin{aligned} \det \left(\begin{bmatrix} C_{Z \setminus p} & \bar{R}^\top \\ \bar{R} & C_Z(0) \end{bmatrix} \right) &= (-1)^r \det \left(\begin{bmatrix} \bar{R} & C_Z(0) \\ C_{Z \setminus p} & \bar{R}^\top \end{bmatrix} \right) \\ &= (-1)^{2r} \det \left(\begin{bmatrix} C_Z(0) & \bar{R} \\ \bar{R}^\top & C_{Z \setminus p} \end{bmatrix} \right) \\ &= \det \left(\begin{bmatrix} C_Z(0) & \bar{R} \\ \bar{R}^\top & C_{Z \setminus p} \end{bmatrix} \right) = \det(C_Z(0)), \end{aligned}$$

which completes the proof.

D. Proof of Theorem 2

Proof "⇐"

If A_1, \dots, A_p are diagonal, then under assumption (A2) of uncorrelated residuals, x_t and y_t are independent. It follows immediately from the argumentation in Section II-C that $\tilde{D}_{x \rightarrow y}^{(net)} = 0$. \square

Proof "⇒"

Let $\tilde{D}_{x \rightarrow y}^{(net)} = 0$. From (32) and (33) in Theorem 1 we necessarily have $\tilde{\Sigma}_{xx} = \Sigma_{xx}$ and $\tilde{\Sigma}_{yy} = \Sigma_{yy}$.

From the equality of determinants $\det(\Sigma) = \det(\tilde{\Sigma})$ in (37) and assumption (A2) of uncorrelated residuals, $\Sigma_{xy} = 0$, it follows:

$$\Sigma = \tilde{\Sigma} = \text{diag}.$$

Now, the induction proof of the following statement completes the whole proof.

Proof via induction:

Statement S(p): In any stable bivariate VAR(p) process (1) with the time-reversed representation (30) fulfilling (A1), (A2) and (A3), and

$$(AS): \quad \Sigma = \tilde{\Sigma} = \text{diag},$$

the A_1, \dots, A_p are diagonal.

Preliminaries.

First, note that (AS) implies that the time-reversed coefficient matrices $\tilde{A}_1, \tilde{A}_2, \dots, \tilde{A}_p$ are lower triangular, since

$$\tilde{\Sigma}_{xx} \stackrel{(AS)}{=} \Sigma_{xx} \stackrel{(A1)}{=} \Sigma_x \stackrel{(10)}{=} \tilde{\Sigma}_x. \quad (67)$$

From Andel (58) we know that

$$\underbrace{\tilde{A}_p}_{\text{lower diag}} \underbrace{\Sigma}_{\text{diag}} = \underbrace{\tilde{\Sigma}}_{\text{diag upper}} \underbrace{A_p^\top}_{\text{diag upper}}.$$

With (AS),

$$A_p = \tilde{A}_p = \text{diag} \quad (68)$$

immediately follows.

Basis: Show that the statement holds for $p = 1$.

Follows directly from (68).

Inductive step: Show that if $S(p-1)$ then $S(p)$.

From Andel (59):

$$\begin{aligned} \Sigma &\stackrel{(AS)}{=} \tilde{\Sigma} \stackrel{(59)}{=} (Q_{pp} + A_p^\top \Sigma^{-1} A_p)^{-1} = (Q_{pp} + A_p^2 \Sigma^{-1})^{-1} \\ \Rightarrow \Sigma^{-1} &= Q_{pp} + A_p^2 \Sigma^{-1} \\ \Rightarrow Q_{pp} &= (I - A_p^2) \Sigma^{-1} = \text{diag}. \end{aligned} \quad (69)$$

Denote with \tilde{Z}_t the VAR(1) representation of the time-reversed process \tilde{z}_t , and with $\tilde{Q} := C_{\tilde{Z}}(0)^{-1}$ its inverse covariance matrix with block notation $\tilde{Q} =: (\tilde{Q}_{lk})_{l,k=1}^p$. Then, due to symmetry,

$$\tilde{Q}_{pp} = (I - \tilde{A}_p^2) \tilde{\Sigma}^{-1} \stackrel{(68),(AS)}{=} (I - A_p^2) \Sigma^{-1} = Q_{pp}. \quad (70)$$

Let us now define the VAR($p-1$) process \mathbf{z}'_t by

$$\mathbf{z}'_t = \sum_{i=1}^{p-1} B_i \mathbf{z}'_{t-i} + \xi_t, \quad \langle \xi_t \xi_t^\top \rangle = \Sigma', \quad (71)$$

where

$$\begin{aligned} [B_1, \dots, B_{p-1}] &:= [C_Z(1), \dots, C_Z(p-1)] \cdot C_{Z \setminus p}(0)^{-1} \\ &= \bar{R} \cdot C_{Z \setminus p}(0)^{-1} \\ \Sigma' &:= C_Z(0) - \bar{R} C_{Z \setminus p}(0)^{-1} \bar{R}^\top \end{aligned}$$

arise from solving the Yule-Walker equations with respect to $C_Z(0), C_Z(1), \dots, C_Z(p-1)$. They are the least squares solution when regressing \mathbf{z}_t onto $\mathbf{z}_{t-1}, \dots, \mathbf{z}_{t-p+1}$.

Denote with $\tilde{B}_1, \dots, \tilde{B}_{p-1}$ and $\tilde{\Sigma}'$ the time-reversed coefficients and residual covariance matrix of $\tilde{\mathbf{z}}'_t := \mathbf{z}'_{-t}$. By definition of time reversal, the coefficients of $\tilde{\mathbf{z}}'_t$ are the solution of the Yule-Walker equations with respect to $C_Z(0), C_Z(-1), \dots, C_Z(1-p)$. They are the least squares solution when regressing $\tilde{\mathbf{z}}_t$ onto $\tilde{\mathbf{z}}_{t-1}, \dots, \tilde{\mathbf{z}}_{t-p+1}$.

We now show that all assumptions in $S(p-1)$ hold for \mathbf{z}'_t .

[A0] *Stability of \mathbf{z}'_t .*

Since \mathbf{z}_t is stable, performing an insufficient lag-order fit (which is the case for \mathbf{z}'_t) preserves stability (see [46]).

[A1] *Lower triangularity of B_1, \dots, B_{p-1} .*

Equations (8) and (10) in [47] state the following relations:

$$\begin{aligned} [A_1, \dots, A_{p-1}] &= [B_1, \dots, B_{p-1}] - A_p [\tilde{B}_{p-1}, \dots, \tilde{B}_1] \quad (72) \\ [\tilde{A}_1, \dots, \tilde{A}_{p-1}] &= [\tilde{B}_1, \dots, \tilde{B}_{p-1}] - \underbrace{\tilde{A}_p}_{A_p} [\tilde{B}_{p-1}, \dots, \tilde{B}_1] \end{aligned}$$

Since all A_i and \tilde{A}_i are lower triangular ((A1), (67)), and A_p is diagonal, we deduce:

$$\begin{aligned} [\tilde{B}_{p-k}]_{12} &= [A_p]_{11}^{-1} [B_k]_{12}, \quad \forall k \in \{1, \dots, p-1\}. \\ [\tilde{B}_{p-k}]_{12} &= [A_p]_{11} [B_k]_{12} \end{aligned}$$

This may only be fulfilled when either $[A_p]_{11} = 1$ or all B_k (and \tilde{B}_k) are lower triangular. Recalling that A_p is diagonal and $Q_{pp} = (I - A_p^2) \Sigma^{-1}$ is invertible, necessarily we have $[A_p]_{11} \neq 1$. Therefore, B_1, \dots, B_{p-1} are lower triangular.

[A2] *Diagonality of Σ'*

[AS] *and $\Sigma' = \tilde{\Sigma}'$.*

From [47], page 6, we have that

$$Q_{pp}^{-1} = \tilde{\Sigma}' \quad \text{and} \quad \tilde{Q}_{pp}^{-1} = \Sigma'.$$

From (69) and (70),

$$\Sigma' = \tilde{\Sigma}' = \text{diag}$$

immediately follows.

[A3] *Invertibility of $C_{\mathbf{z}'}(0)$.*

By construction \mathbf{z}'_t is the solution of the Yule-Walker equations with respect to the autocovariances $C_{\mathbf{z}}(0), \dots, C_{\mathbf{z}}(p-1)$. Thus, $C_{\mathbf{z}'}(0) = C_{\mathbf{z}}(0)$, which is invertible.

Therefore, we can apply the induction claim $S(p-1)$ for \mathbf{z}'_t saying that B_1, \dots, B_{p-1} are diagonal. With the same argument $\tilde{B}_1, \dots, \tilde{B}_{p-1}$ are also diagonal.

Using the factorization (72)

$$A_k = \underbrace{B_k}_{diag} - \underbrace{A_p}_{diag} \underbrace{\tilde{B}_{p-k}}_{diag}, \quad k \in \{1, \dots, p-1\}$$

we have that A_1, \dots, A_p are diagonal, which completes the proof. \square

REFERENCES

- [1] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Springer, 2007.
- [2] M. Winterhalder, B. Schelter, W. Hesse, K. Schwab, L. Leistriz, D. Klan, R. Bauer, J. Timmer, and H. Witte, "Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems," *Signal processing*, vol. 85, no. 11, pp. 2137–2160, 2005.
- [3] W. Mader, D. Feess, R. Lange, D. Saur, V. Glauche, C. Weiller, J. Timmer, and B. Schelter, "On the detection of direct directed information flow in fMRI," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 6, pp. 965–974, 2008.
- [4] S. L. Bressler and A. K. Seth, "Wiener-Granger Causality: A well established methodology," *NeuroImage*, vol. 58, no. 2, pp. 323–329, 2011.
- [5] A. Bolstad, B. D. Van Veen, and R. Nowak, "Causal network inference via group sparse regularization," *IEEE Transactions on Signal Processing*, vol. 59, no. 6, pp. 2628–2641, 2011.
- [6] R. K. Kaufmann and D. I. Stern, "Evidence for human influence on climate from hemispheric temperature relations," *Nature*, vol. 388, no. 6637, pp. 39–44, 1997.
- [7] U. Triacca, "On the use of Granger causality to investigate the human influence on climate," *Theoretical and Applied Climatology*, vol. 69, no. 3–4, pp. 137–138, 2001.
- [8] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.
- [9] J. D. Hamilton, *Time Series Analysis*. Princeton University Press, 1994.
- [10] H. Nalatore, M. Ding, and G. Rangarajan, "Mitigating the effects of measurement noise on Granger causality," *Phys. Rev. E*, vol. 75, p. 031123, 2007.
- [11] G. Nolte, A. Ziehe, V. V. Nikulin, A. Schlögl, N. Krämer, T. Brismar, and K.-R. Müller, "Robustly estimating the flow direction of information in complex physical systems," *Phys. Rev. Lett.*, vol. 100, p. 234101, 2008.
- [12] G. Nolte, O. Bai, L. Wheaton, Z. Mari, S. Vorbach, and M. Hallett, "Identifying true brain interaction from EEG data using the imaginary part of coherency," *Clinical neurophysiology*, vol. 115, no. 10, pp. 2292–2307, 2004.
- [13] G. Gómez-Herrero, M. Atienza, K. Egiazarian, and J. L. Cantero, "Measuring directional coupling between EEG sources," *NeuroImage*, vol. 43, no. 3, pp. 497 – 508, 2008.
- [14] J.-M. Schoffelen and J. Gross, "Source connectivity analysis with MEG and EEG," *Human Brain Mapping*, vol. 30, no. 6, pp. 1857–1865, 2009.
- [15] S. Haufe, R. Tomioka, G. Nolte, K.-R. Müller, and M. Kawanabe, "Modeling sparse connectivity between underlying brain sources for EEG/MEG," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 8, pp. 1954–1963, 2010.
- [16] S. Haufe, V. Nikulin, K.-R. Müller, and G. Nolte, "A critical assessment of connectivity measures for EEG data: A simulation study," *NeuroImage*, vol. 64, pp. 120–133, 2013.
- [17] S. Haufe, V. V. Nikulin, and G. Nolte, "Alleviating the influence of weak data asymmetries on granger-causal analyses," in *Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 25–33.
- [18] M. Vinck, L. Huurdeman, C. A. Bosman, P. Fries, F. P. Battaglia, C. M. Pennartz, and P. H. Tiesinga, "How to detect the Granger-causal flow direction in the presence of additive noise?" *NeuroImage*, vol. 108, pp. 301–318, 2015.
- [19] M. Kaminski and K. J. Blinowska, "A new method of the description of the information flow in the brain structures," *Biological cybernetics*, vol. 65, no. 3, pp. 203–210, 1991.
- [20] L. A. Baccalá and K. Sameshima, "Partial directed coherence: a new concept in neural structure determination," *Biological cybernetics*, vol. 84, no. 6, pp. 463–474, 2001.
- [21] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel method for nonlinear Granger causality," *Phys. Rev. Lett.*, vol. 100, p. 144103, 2008.
- [22] M. Grosse-Wentrup, "Understanding brain connectivity patterns during motor imagery for brain-computer interfacing," in *Advances in neural information processing systems*, 2009, pp. 561–568.
- [23] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, "Transfer entropy - a model-free measure of effective connectivity for the neurosciences," *Journal of Computational Neuroscience*, vol. 30, pp. 45–67, 2011.
- [24] M. Breakspear, M. J. Brammer, E. T. Bullmore, P. Das, and L. M. Williams, "Spatiotemporal wavelet resampling for functional neuroimaging data," *Hum Brain Mapp*, vol. 23, no. 1, pp. 1–25, Sep 2004.
- [25] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *Journal of the American Statistical Association*, vol. 77, no. 378, pp. 304–313, 1982.
- [26] G. Nolte, A. Ziehe, N. Krämer, F. Popescu, and K.-R. Müller, "Comparison of Granger causality and Phase Slope Index," in *NIPS Causality: Objectives and Assessment*, 2010, pp. 267–276.
- [27] G. Nolte, F. C. Meinecke, A. Ziehe, and K.-R. Müller, "Identifying interactions in mixed and noisy complex systems," *Phys. Rev. E*, vol. 73, p. 051913, 2006.
- [28] H. Wold, *A study in the analysis of stationary time series*. Almqvist & Wiksell, 1938.
- [29] J. Andel, "Symmetric and reversed multiple stationary autoregressive series," *The Annals of Mathematical Statistics*, vol. 43, no. 4, pp. 1197–1203, 1972.
- [30] M. S. Bartlett, *An Introduction to Stochastic Processes*. Cambridge University Press, 1955.
- [31] A. G. Akritas, E. K. Akritas, and G. I. Malaschonok, "Various proofs of Sylvester's (determinant) identity," *Mathematics and Computers in Simulation*, vol. 42, no. 4, pp. 585–593, 1996.
- [32] G. E. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, no. 2, p. 461464, 1978.
- [33] P. Newbold, "Feedback induced by measurement errors," *International Economic Review*, pp. 787–791, 1978.
- [34] L. Sommerlade, M. Thiel, B. Platt, A. Plano, G. Riedel, C. Grebogi, J. Timmer, and B. Schelter, "Inference of Granger causal time-dependent influences in noisy multivariate time series," *Journal of Neuroscience Methods*, vol. 203, no. 1, pp. 173 – 185, 2012.
- [35] G. C. Tiao and W. S. Wei, "Effect of temporal aggregation on the dynamic relationship of two time series variables," *Biometrika*, vol. 63, no. 3, pp. 513–523, 1976.
- [36] J. R. McCrorie and M. J. Chambers, "Granger causality and the sampling of economic processes," *Journal of Econometrics*, vol. 132, no. 2, pp. 311 – 336, 2006.
- [37] D. Zhou, Y. Zhang, Y. Xiao, and D. Cai, "Reliability of the Granger causality inference," *New Journal of Physics*, vol. 16, no. 4, p. 043016, 2014.
- [38] A. K. Seth, P. Chorley, and L. C. Barnett, "Granger causality analysis of fMRI BOLD signals is invariant to hemodynamic convolution but not downsampling," *NeuroImage*, vol. 65, pp. 540 – 555, 2013.
- [39] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, "Network modelling methods for fMRI," *Neuroimage*, vol. 54, no. 2, pp. 875–891, 2011.
- [40] A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer, "Estimation of a structural vector autoregression model using non-gaussianity," *The Journal of Machine Learning Research*, vol. 11, pp. 1709–1731, 2010.
- [41] A. Moneta, N. Chla, D. Entner, and P. Hoyer, "Causal search in structural vector autoregressive models," in *Causality in Time Series Challenges in Machine Learning*, 2011, vol. 5, pp. 95–118.
- [42] J. Chiang, Z. J. Wang, and M. J. McKeown, "A generalized multivariate autoregressive (GmAR)-based approach for EEG source connectivity analysis," *IEEE Transactions on Signal Processing*, vol. 60, pp. 453–465, 2012.
- [43] J. Peters, D. Janzing, and B. Schölkopf, "Causal inference on time series using restricted structural equation models," in *Advances in Neural Information Processing Systems*, 2013, pp. 154–162.
- [44] H. Nalatore, N. Sasikumar, and G. Rangarajan, "Effect of measurement noise on Granger causality," *Physical Review E*, vol. 90, no. 6, p. 062127, 2014.

- [45] D. V. Ouellette, "Schur complements and statistics," *Linear Algebra and its Applications*, vol. 36, pp. 187 – 295, 1981.
- [46] P. Whittle, "On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix," *Biometrika*, vol. 50, no. 1-2, pp. 129–134, 1963.
- [47] F. Sowell, "A decomposition of block toeplitz matrices with applications to vector time series," *Carnegie Mellon University*, 1989.