Assessing the significance of directed and multivariate measures of linear dependence between time series

Oliver M. Cliff,[1,2,3,*] Leonardo Novelli,[1,3] Ben D. Fulcher,[1,2] James M. Shine,[1,4] and Joseph T. Lizier[1,3]
[1]*Centre for Complex Systems, The University of Sydney, Sydney NSW 2006, Australia*
[2]*School of Physics, The University of Sydney, Sydney NSW 2006, Australia*
[3]*Faculty of Engineering, The University of Sydney, Sydney NSW 2006, Australia*
[4]*Brain and Mind Centre, School of Medical Sciences, The University of Sydney, Sydney NSW 2006, Australia*

(Received 20 March 2020; accepted 21 January 2021; published 12 February 2021)

Inferring linear dependence between time series is central to our understanding of natural and artificial systems. Unfortunately, the hypothesis tests that are used to determine statistically significant directed or multivariate relationships from time-series data often yield spurious associations (Type I errors) or omit causal relationships (Type II errors). This is due to the autocorrelation present in the analyzed time series—a property that is ubiquitous across diverse applications, from brain dynamics to climate change. Here we show that, for limited data, this issue cannot be mediated by fitting a time-series model alone (e.g., in Granger causality or prewhitening approaches), and instead that the degrees of freedom in statistical tests should be altered to account for the effective sample size induced by cross-correlations in the observations. This insight enabled us to derive modified hypothesis tests for any multivariate correlation-based measures of linear dependence between covariance-stationary time series, including Granger causality and mutual information with Gaussian marginals. We use both numerical simulations (generated by autoregressive models and digital filtering) as well as recorded fMRI-neuroimaging data to show that our tests are unbiased for a variety of stationary time series. Our experiments demonstrate that the commonly used F- and χ^2-tests can induce significant false-positive rates of up to 100% for both measures, with and without prewhitening of the signals. These findings suggest that many dependencies reported in the scientific literature may have been, and may continue to be, spuriously reported or missed if modified hypothesis tests are not used when analyzing time series.

DOI: 10.1103/PhysRevResearch.3.013145

I. INTRODUCTION

Linear dependence measures such as Pearson correlation, canonical correlation analysis, and Granger causality are used in a broad range of scientific domains to investigate the complex relationships in both natural and artificial processes. Despite their widespread use, concerns have been raised about the hypothesis tests typically used to assess the statistical significance of such measures from time series [1–6]. Specifically, the presence of autocorrelation in a signal—one of two defining properties of a stationary time series [7,8]—has been known to bias statistics since the beginning of time-series analysis [9]. If left unaccounted, this bias yields a greater number of both spurious correlations and missed causalities (Type I and Type II errors) due to size and power distortions of the hypothesis tests. With the recent findings [1–6] suggesting that existing techniques do not adequately address autocorrelation, the accuracy of many reported results across the empirical sciences may be called into question.

The notion that autocorrelation affects the sampling distribution of time-series properties has a long history in statistics, with research often focusing on the relationship between two univariate processes. Seminal work by Bartlett [10,11] revealed that autocorrelation can distort the degrees of freedom available to compute statistics such as Pearson correlation coefficients. In practical terms, this induces an "effective sample size," where the effective number of independent samples used in computing an estimate is different to the actual length of the dataset. Two opposing strategies have been proposed for handling autocorrelation: remove the autoregressive (AR) components of the time series before computing statistics, or modify the hypothesis tests that assess the distorted measurements. The former approach, known as prewhitening, involves filtering the time series in order to render the residuals serially independent [12]. Prewhitening is known to have many issues, such as reducing the size and power properties of hypothesis tests both in theory [13] and in practice, with simulated [14] and recorded [15,16] time-series data in a variety of domains. In contrast, the notion of modifying hypothesis tests remains relatively underused in practice, more often found in applications involving short time series and high autocorrelation, where the statistical bias of measures is most pronouced (with or without prewhitening), e.g., in

[*] oliver.cliff@sydney.edu.au

fMRI-based neuroimaging [1,2,17], as well as environmental and ecological studies [18,19]. Indeed, it was not until recently that the efficacy of a modified z-test for correlation analysis was demonstrated successfully on fMRI signals [2], which have been widely characterized using correlation coefficients [20]. Nevertheless, the theory of autocorrelation on the undirected relationship between bivariate time series is now well developed. However, the extension of this theory to multiple time series, and to directed relationships, remains incomplete.

Motivated to study directed dependencies in economics, Granger [21] introduced a measure of causal influence between AR models nearly 60 years ago. Since then, it has become exceedingly popular, exemplified by more than 100 000 works indexed by Google Scholar that contain the phrase "Granger causality" (as of June 2020). This impact is reflected in the measure's ubiquity in the scientific community beyond its origins in econometrics, generating highly influential results on phenomena ranging from brain dynamics [22–24] to climate change [25,26] and political relationships [27,28]. Granger causality controls for the confounding past of a process through linear regression, building statistics and hypothesis tests via residuals rather than the original process. However, researchers are becoming aware that certain preprocessing techniques that increase autocorrelation, such as filtering, raise the false-positive rate (FPR) of Granger causality tests when using the well-established χ^2- and F-distributions [3–5]. Even though these empirical studies have demonstrated that established Granger causality tests have distorted size and power properties (exhibiting Type I and II errors), it has remained unclear as to why and how to correct them. In this paper, we illustrate that these errors are due to an inflated variance of the null distribution as a function of autocorrelation remaining in the residuals, in the same way that bivariate correlation is affected.

In order to unify Bartlett's earlier investigations on correlation coefficients (under autocorrelation) with more complex measures such as Granger causality, we must expand the former body of work to account for multivariate relationships. One such multivariate generalization of Pearson correlation is referred to as Wilks' criterion [29], which quantifies the relationship between multiple sets of variables, and is Λ-distributed for independent observations [30]. In particular, we are interested in a special case of Wilks' criterion popularized by Hotelling [31] that focuses on two sets of variables, referred to as canonical correlation analysis. It was later established that, like Pearson correlation, estimates of canonical correlations are inefficient under autocorrelation [6], introducing Type I and II errors under hypothesis tests that assume independence (such as the Λ-distribution). Instead of deriving hypothesis tests directly for Wilks' criterion or canonical correlations, here we use the equivalent information-theoretic formulation.

Information theory's general applicability arises in simply requiring a probability distribution that can be either parametric or nonparametric [32,33]. When this probability distribution is modeled as a multivariate Gaussian, canonical correlation analysis and information theory overlap because mutual information can be decomposed into sums involving canonical variables [34]. Moreover, Granger causality is now understood as a special case of conditional mutual information, known as transfer entropy [35–37]. While this unification provides an elegant perspective, there remains a clear divide between the theoretical foundations of Bartlett (and others [38–40]) and the large family of multivariate linear dependence measures that information theory provides.

In this work we bridge this gap by leveraging the concept of the effective sample size to derive hypothesis tests for any correlation-based measure of linear dependence between covariance-stationary time series. This comprises a large family of well-known statistics based on ratios of generalized variance—such as Granger causality and mutual information—that we introduce in Sec. II. To achieve this, we first provide the one-tailed and two-tailed tests for the sample partial correlations between two univariate processes under autocorrelation in Sec. III. Although this result is important in its own right, in this work we primarily leverage it to construct the tests for more advanced inference procedures with multivariate and directed models of observed dynamics. Following this, we introduce the modified Λ-test (in Sec. IV), which we show is suitable for assessing the significance of any linear dependence measure that can be expressed as a ratio of generalized variances. Specifically, in Sec. V we use the two-tailed test to derive hypothesis tests for conditional mutual information estimates between bivariate time-series data. We then use the chain rule for mutual information to extend this result to multivariate time-series data. Finally, since Granger causality can be expressed as a conditional mutual information, in Sec. VI we extend our results further to derive Granger causality tests for both bivariate and multivariate time-series datasets. More broadly, the modified Λ-test can be used for any measure that can be expressed in terms of conditional mutual information (or, equivalently, Wilks' criterion or partial correlation), e.g., canonical correlations and partial autocorrelation [7,8] or information-theoretic measures (for linear-Gaussian processes) such as predictive information [41,42] and active information storage [43].

Using numerical simulations throughout Sec. VII, we validate the modified Λ-test and characterize the effect of autocorrelation on both the χ^2- and F-test. Our experiments involve generating samples from two first-order independent AR models and iteratively filtering the output signal such that the autocorrelation is increased for both time series; this simulates empirical analysis in practice, and allows for the process parameters to be modified while ensuring that the null hypothesis (of no interprocess dependence) is not violated. We perform these experiments for mutual information and Granger causality in their unconditional, conditional, and multivariate forms. Our results generally agree with the hypotheses that the FPR of F- and χ^2-tests can be inflated by either increasing the autocorrelation (through filtering) or, for the χ^2-test, the number of conditionals (through increasing the dimension of mutual information or the history length of Granger causality). These experiments mirror empirical applications where digital filtering is often used in preprocessing for many purposes, such as handling nonstationary effects, which inadvertently increases autocorrelation and therefore the FPRs of unmodified tests. Given minimally sufficient effective samples, however, we confirm that the modified Λ-tests remain unbiased for all scenarios. We thus show that, in contrast, the size (Type I errors) and power (Type II errors) of F- or χ^2-tests are arbitrarily low for a large class of multivariate linear dependence measures (approximately zero in certain instances) and overwhelmingly depends on the parameters of the underlying independent processes. We further demonstrate that the common approach of prewhitening a signal (in order to remove the effect of autocorrelation) does not suffice to control the FPR in almost all cases. Finally, by using a well-known brain-imaging dataset from the Human Connectome Project [44], we verify that our previous numerical simulations yield comparable results to experiments on commonly used datasets. For these experiments, the χ^2-tests of mutual information and Granger causality yield concerningly high FPRs of over 80% and 65% for a nominal significance of 5%—a 16- and a 13-fold increase—whereas our exact tests maintain the ideal FPR for all experiments. Open-source MATLAB code is made available to allow users to perform correct hypothesis testing for all dependence measures, as well as the above experiments, at Ref. [45].

Our theoretical and empirical findings suggest that this work presents the first statistically sound approach for testing the linear dependence between multivariate time-series data. Given that approaches such as prewhitening and Granger causality are specifically designed to account for autocorrelation, we conjecture that autocorrelation-induced statistical errors caused by F- or χ^2-tests (and others) may be even more prevalent in prior publications than previously suggested by several authors [1,3,4]. In particular, our case study of brain-imaging data is concerning, because the neuroscience community employs techniques such as correlation, mutual information, and Granger causality in order to infer pairwise dependence (known as "functional connectivity"). Implementation of our approach will enable correct inference of linear relationships within complex systems across myriad scientific applications.

II. MEASURES OF LINEAR DEPENDENCE

In this work, we focus on multivariate signals,

$$\{Z_1(t),\dots,Z_m(t)\},\ t=0,\pm1,\pm2,\dots, \tag{1}$$

that is, a collection of m series sampled at equally spaced time intervals. Writing

$$\boldsymbol{Z}(t)=(Z_1(t),\dots,Z_m(t))', \tag{2}$$

we shall refer to the m series as an m-dimensional vector of multiple time series such that $\boldsymbol{Z}(t)\in\mathbb{R}^m$.

For the purposes of inferring linear dependence, $\boldsymbol{Z}$ is partitioned into one k-variate and one l-variate subprocess [46]:

$$\boldsymbol{Z}=\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix}, \tag{3}$$

reflecting an interest in the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$. The linear dependence of $\boldsymbol{X}$ on $\boldsymbol{Y}$ (or vice versa) is measured by a scalar value that quantifies how much the outcomes of $\boldsymbol{Y}$ reduce uncertainty over outcomes of $\boldsymbol{X}$. Theoretically, in the absence of a linear relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ (the null hypothesis, $\mathcal{H}_0$) the reduction of uncertainty is exactly zero, meaning that $\boldsymbol{Y}$ does not linearly predict $\boldsymbol{X}$ at all. In practice, however, we have access only to a finite-length dataset with T observations over which to compute the measures, introducing a variation in statistical estimates and manifesting as nonzero values in the case of no relationship. Here we present this dataset as an $m\times T$ matrix $\boldsymbol{z}$ of consecutive real-valued samples $\boldsymbol{z}(t)\in\mathbb{R}^m$ of the process $\boldsymbol{Z}$ (again, this is partitioned into submatrices $\boldsymbol{x}$ and $\boldsymbol{y}$). To this end, the aim of linear-dependence tests is to infer whether there is a statistical dependence between $\boldsymbol{X}$ and $\boldsymbol{Y}$ based on the sample paths $\boldsymbol{x}$ and $\boldsymbol{y}$ alone.

We make the typical assumption that the underlying system, $\boldsymbol{Z}$, is a second-order stationary, purely nondeterministic process [7,8,47]. An important consequence of covariance-stationarity is that the time series may be represented, after appropriate mean removal and differencing [48], by the ARMA model:

$$\boldsymbol{Z}(t)=\boldsymbol{a}(t)+\sum_{u=1}^{p}\boldsymbol{\Phi}(u)\boldsymbol{Z}(t-u)+\sum_{u=1}^{q}\boldsymbol{\Theta}(u)\boldsymbol{a}(t-u), \tag{4}$$

where $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ are vectors of autoregressive (AR) and moving-average (MA) parameters, and $\boldsymbol{a}(t)$ is uncorrelated noise (the innovation process). We further assume that the noise is Gaussian, $\boldsymbol{a}(t)\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{\Sigma})$ for some arbitrary noise covariance $\boldsymbol{\Sigma}$, meaning that $\boldsymbol{Z}$ is a linear-Gaussian process.

A. Cross-correlation and autocorrelation

For covariance-stationary time series, the relationship between $Z_i(t)$ and $Z_j(t+u)$ depends only on the difference in times t and $t+u$ of the observation but not on t itself. Once the mean has been removed, such processes are fully defined by their cross-correlation,

$$\rho_{ij}(u)=\frac{\gamma_{ij}(u)}{\sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}}, \tag{5}$$

with $\gamma_{ij}(u)=\operatorname{cov}(Z_i(t),Z_j(t+u))$ the cross-covariance between $Z_i(t)$ and $Z_j(t+u)$. If $\rho_{ii}(u)\neq0$ for any $u>0$, then the univariate process Z_i exhibits autocorrelation, and the collection of $\rho_{ii}(u)$ for $u=0,\pm1,\pm2,\dots$ is generally called the autocorrelation function of Z_i. The sample cross-correlation coefficients are computed from time-series data $\boldsymbol{z}$ as

$$r_{ij}(u)=\frac{c_{ij}(u)}{\sqrt{c_{ii}(0)c_{jj}(0)}}, \tag{6}$$

with $c_{ij}(u)=N^{-1}\sum_{t=1}^{T}z_i(t)z_j(t+u)$ where $N=T-1$ as an unbiased estimate of the sample cross-covariance.

The first linear dependence measure we discuss is Pearson's product-moment correlation coefficient. For bivariate ($m=2$) processes, we shall write $\boldsymbol{X}=X$ and $\boldsymbol{Y}=Y$ and denote the cross-correlation between these variables [Eq. (5)] as $\rho_{XY}(u)$. Pearson's correlation coefficient is the lag-zero cross-correlation $\rho_{XY}=\rho_{XY}(0)$, and quantifies the (symmetric) association between paired observations of X and Y. The sample correlation coefficient is then given by

$$r_{xy}=\frac{c_{xy}}{c_{xx}c_{yy}}, \tag{7}$$

where $c_{xy}=c_{xy}(0)$ is the sample covariance, and $c_{xx}=c_{xx}(0)$ and $c_{yy}=c_{yy}(0)$ are sample variances [39].

In order to assess the statistical significance of linear dependence measures, such as the sample correlation coefficient, we must be able to compute their variance, i.e., $\sigma_r^2(x, y) = \mathrm{var}(r_{xy})$. For independent but autocorrelated stationary processes, the variance of the sample correlation coefficient can be estimated (to the first order) as [2,38,39]

$$\hat{\sigma}_r^2(x, y) \approx T^{-1}\left[1 + 2\sum_{u=1}^{T-1}\frac{T-u}{T}r_{xx}(u)r_{yy}(u)\right], \tag{8}$$

where $r_{xx}(u)$ and $r_{yy}(u)$ are the lag u sample autocorrelations [49]. Although we refer to Eq. (8) as Bartlett's formula, this first-order approximation is due to Clifford *et al.* [39], who presented the variance estimator for spatial autocorrelation and an estimate of the effective number of independent samples for correlation coefficients:

$$\hat{\eta}(x, y) = 1 + \hat{\sigma}_r^{-2}(x, y). \tag{9}$$

An important consequence of Eqs. (8) and (9) is that hypothesis tests, such as Student's t-test, should have degrees of freedom corresponding to the effective sample size of the analyzed time series (i.e., the effective degree of freedom [2,11]), rather than the original sample size [39]. That is, if both X and Y are autocorrelated, then the null distribution for r_{xy} follows a modified Student's t-test:

$$r_{xy}\sqrt{\frac{\hat{\eta}(x, y) - 2}{1 - r_{xy}^2}} \sim t[\hat{\eta}(x, y) - 2], \tag{10}$$

where $\hat{\eta}(x, y) - 2$ is the (estimated) effective degrees of freedom. An examination of this formula reveals that, when both x and y are positively autocorrelated, then there are, effectively, fewer independent observations than in the original dataset ($\hat{\eta} < T$); if only one process is negatively autocorrelated, then there appear to be more independent observations than in the original dataset ($\hat{\eta} > T$) [39]. Consequently, when the modified degree of freedom in Eq. (10) is neglected, inference procedures can either spuriously identify association (produce Type I errors) when $\hat{\eta} < T$ or miss actual correlations (Type II errors) when $\hat{\eta} > T$.

If either one (or both) of x or y are serially independent, then the sample correlation coefficients r_{xy} can be tested against Student's t-distribution with degrees of freedom $T - 2$. Thus, the textbook approach for minimizing the deleterious effects of autocorrelation is to whiten one of the time series by filtering any AR components (referred to as prewhitening). The idea is that, by filtering any AR components, the residuals become uncorrelated and so statistical tests that have been developed for independent variables can now be used without modifying the degree of freedom. In Sec. VIII we discuss this approach in more detail, showing that linear-dependence tests applied to signals that have been "whitened" in this way still exhibit significant statistical bias (in some cases worse than without prewhitening).

B. Partial correlation

Partial correlation $\rho_{XY\cdot W}$ measures the association between X and Y, whilst controlling for any concomitant effect of another c-variate process $\boldsymbol{W}$ [50,51]. Partial correlation is estimated by, first, computing the residual processes:

$$e_{x|\boldsymbol{w}} = x - \hat{x}(\boldsymbol{w}), \tag{11}$$

$$e_{y|\boldsymbol{w}} = y - \hat{y}(\boldsymbol{w}), \tag{12}$$

where $\hat{x}(\boldsymbol{w})$ denotes the linear prediction of x from $\boldsymbol{w}$ via ordinary least squares. Then an appropriate test statistic for the null hypothesis $\mathcal{H}_0{:}\rho_{XY\cdot W} = 0$ of no relation between X and Y, above any relationship with $\boldsymbol{W}$, is the sample partial correlation:

$$r_{xy\cdot\boldsymbol{w}} = \frac{\sum_t e_{x|\boldsymbol{w}}(t)\, e_{y|\boldsymbol{w}}(t)}{\sqrt{\sum_t e_{x|\boldsymbol{w}}^2(t)}\sqrt{\sum_t e_{y|\boldsymbol{w}}^2(t)}}. \tag{13}$$

By contrasting the formulas for sample partial correlation [Eq. (13)] with sample cross-correlation [Eq. (6)], it is evident that the former is equivalent to the bivariate correlation between the residuals, i.e., $r_{xy\cdot\boldsymbol{w}} = r_{e_{x|\boldsymbol{w}}e_{y|\boldsymbol{w}}}$.

Unlike Pearson correlation, there is a dearth of research into the null distribution of partial correlation coefficients for autocorrelated time-series data. As such, our first theoretical contribution (in Sec. III) is a derivation for the null distribution of sample partial correlations (13) under autocorrelation, i.e., extending the modified t-test [for bivariate correlation (10)] to facilitate residual processes.

C. Wilks' criterion and canonical correlations

Relating two or more sets of variables is achieved similarly to partial correlation (13), with the exception that the generalized variance is used, rather than the conditional variance [29,31]. Consider the relationship between the k-variate process $\boldsymbol{X}$ and the l-variate process $\boldsymbol{Y}$, in the context of a c-variate concomitant process $\boldsymbol{W}$. To measure their dependence, we use the same procedure as for univariate processes, except now the residuals $\boldsymbol{e}_{\boldsymbol{x}|\boldsymbol{w}}$ and $\boldsymbol{e}_{\boldsymbol{y}|\boldsymbol{w}}$ [from Eqs. (11) and (12)] are multivariate, making the sample covariance $\boldsymbol{s}_{\boldsymbol{xy}|\boldsymbol{w}}$ an $m \times m$ matrix, rather than a scalar value. The generalized sample variance is the determinant of these sample covariances $|\boldsymbol{s}_{\boldsymbol{xy}|\boldsymbol{w}}|$, and can be used to form a special case of (residual) Wilks' criterion [29]:

$$\frac{|\boldsymbol{s}_{\boldsymbol{xy}|\boldsymbol{w}}|}{|\boldsymbol{s}_{\boldsymbol{x}|\boldsymbol{w}}||\boldsymbol{s}_{\boldsymbol{y}|\boldsymbol{w}}|}. \tag{14}$$

Although in general Wilks' criterion facilitates any number of partitions of $\boldsymbol{Z}$, we will restrict our attention to two partitions [referring to the special case in Eq. (14) as Wilks' criterion when the meaning is clear]. The null distribution of the ratio of independent generalized variances (14) is known as Wilks' Λ-distribution, with its exact analytic form derived by a number of authors on the basis of no cross-correlation and under the hypothesis that each variable within $\boldsymbol{X}$, $\boldsymbol{Y}$, or $\boldsymbol{W}$ exhibits no autocorrelation [30,52]. Hotelling [31] extensively studied the case of Wilks' criterion with two sets of variables, showing invariance under any internal linear transformation of these sets and a decomposition into canonical correlations with an asymptotic (χ^2) null distribution. Much like their univariate counterparts, however, Hotelling's canonical correlations have been shown to be inefficient under autocorrelation

[6]. Consequently, neither approach is suitable for inferring linear dependence between the majority of time-series data due to the ubiquity of autocorrelation. In Sec. IV we address this issue by providing the null distribution to be used in the presence of autocorrelation [and of course when cross-correlations are present amongst any two variables, i.e., $X_i(t)$ and $X_j(t-u)$ may covary for any i, j, t, or u]. An application that is of particular interest is mutual information, which is equivalent to Wilks' criterion (14) for Gaussian marginals.

D. Mutual information

Mutual information $\mathcal{I}_{\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{W}}$ is a fundamental concept in information theory—and a building block of many other measures—that quantifies the amount of information about a process $\boldsymbol{X}$ obtained by observing another process $\boldsymbol{Y}$ (potentially in the context of a third process $\boldsymbol{W}$, making it a conditional mutual information) [33]. In general, information theory facilitates multivariate analysis by simply requiring well-defined probability distributions that can be either parametric or nonparametric.

When these are normally distributed, mutual information takes a form that is equivalent to Wilks' criterion [33,53]:

$$\hat{\mathcal{I}}_{\boldsymbol{x};\boldsymbol{y}|\boldsymbol{w}} = -\frac{1}{2}\log\left(\frac{|S_{\boldsymbol{xy}|\boldsymbol{w}}|}{|S_{\boldsymbol{x}|\boldsymbol{w}}||S_{\boldsymbol{y}|\boldsymbol{w}}|}\right). \tag{15}$$

This formula is asymptotically equivalent to the nested log-likelihood ratio (LR) of two models [54], and thus we can use a null distribution also provided by Wilks [55]. Following Wilks' theorem [55], under the null hypothesis $\mathcal{H}_0 : \mathcal{I}_{\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{W}} = 0$ and with normally distributed marginals, mutual information estimates are asymptotically chi-square distributed [35,53,54],

$$2T\,\hat{\mathcal{I}}_{\boldsymbol{x};\boldsymbol{y}|\boldsymbol{w}} \sim \chi^2(kl). \tag{16}$$

For limited data, a more precise null distribution can be derived from the standard F-test, albeit for the more specific case of no autocorrelation and with one of the processes being univariate [see Eq. (A4) and surrounding discussion in Appendix A 2]. That is, the mutual information between an i.i.d. variable X and an l-variate $\boldsymbol{Y}$, in the context of the concomitant $\boldsymbol{W}$, is, under the null hypothesis $\mathcal{H}_0 : \mathcal{I}_{X;\boldsymbol{Y}|\boldsymbol{W}} = 0$,

$$\frac{T-(l+c+1)}{l}[\exp(2\hat{\mathcal{I}}_{x;\boldsymbol{y}|\boldsymbol{w}}) - 1] \sim F(l, T-(l+c+1)). \tag{17}$$

The same arguments in Eqs. (15)–(17) hold for unconditional mutual information $\hat{\mathcal{I}}_{\boldsymbol{x};\boldsymbol{y}}$ by setting $\boldsymbol{w} = \emptyset$ (and $c = 0$).

Although we found no discussion on autocorrelation-induced biases of mutual information in literature, statistical tests will clearly be incorrect if autocorrelation is not taken into account. This is evident from the well-known result that mutual information reduces to a function of sample correlation coefficients when the dependent processes are univariate [56]:

$$2\,\hat{\mathcal{I}}_{x;y|\boldsymbol{w}} = -\log\left(1 - r^2_{xy\cdot\boldsymbol{w}}\right). \tag{18}$$

Moreover, it is clear by observing the equivalence between Eqs. (14) and (15), noting that mutual information estimates can be decomposed into sums involving canonical variables [53].

By deriving the exact hypothesis tests for mutual information in Sec. V, we provide a critical component of general-purpose techniques for measuring undirected relationships between sets of variables. However, mutual information was not originally intended to measure autocorrelated time-series dependencies, nor does it naturally model directed dependencies—this is the intended purpose of Granger causality.

E. Granger causality

Granger causality was explicitly designed to capture one-way dependence between stochastic processes by taking into account the confounding influence of their past (i.e., the autocorrelation). By considering $\boldsymbol{X}$ as a target (predictee) process and $\boldsymbol{Y}$ as a source (predictor) process, Granger causality $\mathcal{F}_{\boldsymbol{Y}\to\boldsymbol{X}|\boldsymbol{W}}$ explicitly aims to measure the causality (predictability) in $\boldsymbol{Y}$ about $\boldsymbol{X}$ in context of the relevant history of $\boldsymbol{X}$ (and, potentially, a concomitant process $\boldsymbol{W}$). Of course, the use of the term "causality" here refers to Wiener's definition (as a model of dependence based on prediction) rather than Pearl's (a mechanistic causal effect that can only be inferred using interventions); see Ref. [57] for a differentiation of these concepts.

The main assumption underlying Granger causality is that both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are (vector) AR processes [21,46]. That is, we assume that $\boldsymbol{X}(t)$ and $\boldsymbol{Y}(t)$ are causally dependent on the following states:

$$\boldsymbol{X}^{(p)}(t) = \begin{bmatrix} \boldsymbol{X}(t-1) \\ \vdots \\ \boldsymbol{X}(t-p) \end{bmatrix}, \quad \boldsymbol{Y}^{(q)}(t) = \begin{bmatrix} \boldsymbol{Y}(t-1) \\ \vdots \\ \boldsymbol{Y}(t-q) \end{bmatrix}. \tag{19}$$

Under this assumption, the (directed) influence from $\boldsymbol{Y}$ to $\boldsymbol{X}$ is quantified by conditional mutual information [35]:

$$\mathcal{F}_{\boldsymbol{Y}\to\boldsymbol{X}|\boldsymbol{W}}(p,q) = 2\,\mathcal{I}_{\boldsymbol{X};\boldsymbol{Y}^{(q)}|\boldsymbol{X}^{(p)}\boldsymbol{W}}. \tag{20}$$

Following Eq. (15), this measure can be estimated as a log-ratio of generalized variances [46]:

$$\hat{\mathcal{F}}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}(p,q) = \log\left(\frac{|S_{\boldsymbol{x}|\boldsymbol{x}^{(p)}\boldsymbol{w}}|}{|S_{\boldsymbol{x}|\boldsymbol{x}^{(p)}\boldsymbol{y}^{(q)}\boldsymbol{w}}|}\right). \tag{21}$$

Note that, except for the rather narrow case of $k = q = 1$, Granger causality is a multivariate measure (using generalized variances rather than conditional variances).

The AR orders, p and q, of each process are typically determined by statistical tests such as partial autocorrelation [8], Burg's method [58], the Akaike or Bayesian information criterion (AIC or BIC), cross-validation [4], or active information storage [43,59,60]. In this paper, we use Burg's method to infer the model order due to its efficiency and stability over the Yule-Walker equations [58]. Further, using results from this main text, we discuss the relationship between partial autocorrelation and active information storage in Appendix E.

In general, hypothesis tests for Granger causality can be derived from Wilks' theorem [55]. That is, under the null hypothesis $\mathcal{H}_0$:$\mathcal{F}_{\boldsymbol{Y}\to\boldsymbol{X}|\boldsymbol{W}}(p,q) = 0$, estimates of the Granger causality from $\boldsymbol{Y}$ to $\boldsymbol{X}$ [Eq. (21)] are asymptotically chi-square distributed [46]:

$$T\,\hat{\mathcal{F}}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}(p,q) \sim \chi^2(klq). \tag{22}$$

Alternatively, a finite-sample null distribution can be used if the predictee process is univariate ($k = 1$). Referring to Appendix A, the restricted model has $p + c + 1$ parameters, and the unrestricted model has $p + lq + c + 1$ parameters. Accordingly, we can build the statistic:

$$\frac{T - (p + lq + c + 1)}{lq}\{\exp\,[\hat{\mathcal{F}}_{y\to x|\boldsymbol{w}}(p, q)] - 1\}, \tag{23}$$

which, according to the standard F-test [Eq. (A4)], is distributed as

$$F(lq, T - (p + lq + c + 1)). \tag{24}$$

It should be emphasized, however, that the F-test is suitable only for serially independent observations. Thus, although the Granger causality measure explicitly accounts for autocorrelation in vector AR processes, the established hypothesis tests assume either a sequence of completely independent residuals (the F-test) or infinite data (the χ^2-test). The null distributions that we provide in Sec. VI overcome both of these issues, providing the first valid finite-sample tests for Granger causality.

III. MODIFIED TESTS FOR PARTIAL CORRELATION

In this section, we derive one-tailed and two-tailed tests for the null hypothesis, $\mathcal{H}_0 : \rho_{XY\cdot\boldsymbol{W}} = 0$, of no partial correlation between two univariate autocorrelated time series x and y, given a third (potentially multivariate) process $\boldsymbol{w}$. These tests are valid for any covariance-stationary time series X, Y, and $\boldsymbol{W}$ and sample size T.

A. Modified Student's *t*-test for partial correlation

Recall from Eq. (13) that the sample partial correlation is equivalent to the sample correlation between $e_{x|\boldsymbol{w}}$ and $e_{y|\boldsymbol{w}}$, i.e., $r_{xy\cdot\boldsymbol{w}} = r_{e_{x|\boldsymbol{w}}e_{y|\boldsymbol{w}}}$. Obtaining the null distribution for sample partial correlations between autocorrelated time series can thus be treated similarly to that of the correlation coefficients [see Eq. (10)].

A well-known result is that the sample partial correlation between independent observations is t-distributed $t(\nu)$, under the null hypothesis $\mathcal{H}_0$:$\rho_{XY\cdot\boldsymbol{W}} = 0$, with degrees of freedom $\nu = T - c - 2$ and $c = \dim(\boldsymbol{w}(t))$ [61]. As such, the modified statistic and null distribution is

$$r_{xy\cdot\boldsymbol{w}}\sqrt{\frac{\hat{\eta}(e_{x|\boldsymbol{w}}, e_{y|\boldsymbol{w}}) - c - 2}{1 - r^2_{xy\cdot\boldsymbol{w}}}} \sim t[\hat{\eta}(e_{x|\boldsymbol{w}}, e_{y|\boldsymbol{w}}) - c - 2]. \tag{25}$$

Here the (estimated) effective sample size $\hat{\eta}(e_{x|\boldsymbol{w}}, e_{y|\boldsymbol{w}})$ is still computed from Eq. (9) but with the autocorrelation functions of the residual vectors $e_{x|\boldsymbol{w}}$ and $e_{y|\boldsymbol{w}}$, rather than the original sample paths x and y. Intuitively, this is because $r_{xy\cdot\boldsymbol{w}}$ is itself a sample correlation of these residuals, so it is their autocorrelation—not that of the original time series—that directly determines the effective sample size. Another crucial addition is that the dimension of the conditional process $c = \dim(\boldsymbol{w}(t))$ further reduces the number of degrees of freedom [61], for the same reason as in standard F-tests [4]. When the residual vectors, $e_{x|\boldsymbol{w}}$ and $e_{y|\boldsymbol{w}}$, are (serially) independent, then Eq. (25) becomes equivalent to the standard Student's t-distribution for partial correlation.

B. Modified *F*-test for partial correlation

The Student's t-distribution in Eq. (25) allows for one-tailed (upper or lower) tests for the partial correlation by using the statistic (the LHS) as an input to the quantile function of the t-distribution. For two-tailed tests, another common approach is to square the statistic and, subsequently, the null distribution. The square of a random variable $Z \sim t(\nu)$ that follows Student's t-distribution (with parameter ν) follows an F-distribution with parameters 1 and ν, i.e., $Z^2 \sim F(1, \nu)$. Thus, under the null hypothesis $\mathcal{H}_0$:$\rho_{XY\cdot\boldsymbol{W}} = 0$, the square of the statistic in Eq. (25) (the LHS) follows an F-distribution:

$$n_{xy|\boldsymbol{w}}\frac{r^2_{xy\cdot\boldsymbol{w}}}{1 - r^2_{xy\cdot\boldsymbol{w}}} \sim F(1, n_{xy|\boldsymbol{w}}), \tag{26}$$

with an effective degree of freedom,

$$n_{xy|\boldsymbol{w}} = \hat{\eta}(e_{x|\boldsymbol{w}}, e_{y|\boldsymbol{w}}) - c - 2, \tag{27}$$

obtained from Eq. (9). We refer to the significance test that uses this distribution as the modified F-test. Note that a form of Eq. (26) without modifying the degree of freedom is commonly used for testing the coefficient of determination; thus this approach could also be used for constructing a finite-sample test of the coefficient of multiple correlation under autocorrelation.

IV. MODIFIED Λ-TESTS

Although the modified t- and F-tests introduced above are suitable for bivariate correlation-based measures, they are not appropriate for multivariate (and thus directed) null tests. Here we introduce the Λ^*-distribution, which can be used for hypothesis testing all linear dependence measures throughout this paper.

Recall that, for independent $\boldsymbol{X}$ and $\boldsymbol{Y}$ and $\boldsymbol{W}$, Wilks' criterion [Eq. (14)] is Λ-distributed. The exact form of the Λ-distribution has been extensively studied, with known relationships to the F- and beta-distributions [30,52]. The main purpose of this paper is to derive the finite sample distribution of such statistics under autocorrelation, i.e., where $\boldsymbol{X}(t) \not\perp\!\!\!\perp \boldsymbol{X}(t-u)$ and $\boldsymbol{Y}(t) \not\perp\!\!\!\perp \boldsymbol{Y}(t-v)$ for some $u, v > 0$.

As we will show throughout this work, the distribution for Wilks' criterion with two independent but serially correlated processes can be described by products of Λ-distributed variables with different effective degrees of freedom. We denote this distribution as $\Lambda^*(\boldsymbol{n})$, with the parameter $\boldsymbol{n} = (n_1, \ldots, n_b)'$ comprising the degrees of freedom of each independent Λ-distribution. That is, Wilks' criterion is, under the null hypothesis, Λ^*-distributed:

$$\frac{|s_{\boldsymbol{xy}|\boldsymbol{w}}|}{|s_{\boldsymbol{x}|\boldsymbol{w}}||s_{\boldsymbol{y}|\boldsymbol{w}}|} \sim \Lambda^*(\boldsymbol{n}), \tag{28}$$

where the $\Lambda^*(\boldsymbol{n})$ distribution itself can be described by a product of independent Λ-distributed variables:

$$\prod_{i=1}^{b} L_i, \quad \text{with } L_i \sim \Lambda(n_i, 1, 1). \tag{29}$$

Notice that this reduces to the Λ-distribution for two sets of independent variables [30]; however, with the Λ^*-distribution we are able to include the effective sample sizes.

Although the null distribution for the product of two independent Λ-distributed variates is known [30], deriving the exact distribution for the product of an arbitrary number of Λ-distributed variates is nontrivial. Fortunately, a relationship between the beta-, F-, and Λ-distributions [30,52] allows for simple numerical methods. To generate the distribution $\Lambda^*(\boldsymbol{n})$, we could sample beta-distributed variables:

$$\prod_{i=1}^{b} L_i = \prod_{i=1}^{b} V_i, \quad \text{with } V_i \sim B\left(\frac{n_i}{2}, \frac{1}{2}\right), \tag{30}$$

where $B(\alpha, \beta)$ is the beta distribution. Equivalently, one could sample independent F-distributed variables:

$$\prod_{i=1}^{b} L_i = \prod_{i=1}^{b} \frac{n_i}{U_i + n_i}, \quad \text{with } U_i \sim F(1, n_i). \tag{31}$$

In our experiments (and open-source code), we opt to sample independent beta-distributed variables and construct the Λ^*-distribution from their product as per Eq. (30). Throughout this work, we refer to hypothesis tests that use the Λ^*-distribution and modify the degree of freedom to account for autocorrelation as "modified Λ-tests."

From the relationship between the beta-, F-, and Λ-distributions [see Eqs. (29)–(31)], it is clear that the modified Λ-test is a generalization of the modified F-test, becoming equivalent for univariate statistics. For instance, returning to partial correlation, we have that

$$\frac{|s_{xy|\boldsymbol{w}}|}{|s_{x|\boldsymbol{w}}||s_{y|\boldsymbol{w}}|} = 1 - r^2_{xy\cdot\boldsymbol{w}} \sim \Lambda^*(n_{xy|\boldsymbol{w}}), \tag{32}$$

where $n_{xy|\boldsymbol{w}}$ is the effective degree of freedom. Thus, either the modified F-test or the modified Λ-test could be used for univariate statistics (i.e., ratios of conditional variances).

We can now derive explicit hypothesis tests for common directed and multivariate linear dependence measures using a similar approach. Note that, although the purpose of this work is explicitly for linear dependence measures between time-series data, the modified Λ-test can be easily extended to more general likelihood tests for ratios of generalized variances under spatial autocorrelation [39], which is also known to affect the sampling properties of statistics. We begin by deriving tests for the (conditional) mutual information between both univariate and multivariate linear-Gaussian processes.

V. MODIFIED TESTS FOR MUTUAL INFORMATION

In this section, we obtain hypothesis tests for the mutual information between multiple time series. We first present the hypothesis tests explicitly for conditional mutual information for bivariate time series and, by using the chain rule, obtain the null distribution for the mutual information between multivariate time series.

A. Two time series

The conditional mutual information for linear-Gaussian processes [Eq. (18)] is equivalent to the statistic in Eq. (32). Therefore, estimates of conditional mutual information under the null hypothesis, $\mathcal{H}_0 : \mathcal{I}_{X;Y|\boldsymbol{W}} = 0$, are Λ^*-distributed:

$$\exp\left(-2\hat{\mathcal{I}}_{x;y|\boldsymbol{w}}\right) = 1 - r^2_{xy\cdot\boldsymbol{w}} \sim \Lambda^*(n_{xy|\boldsymbol{w}}). \tag{33}$$

Of course, due to the relationship between the F- and Λ-distribution [noted in Eq. (32)], we can construct an equivalent modified F-test for conditional mutual information:

$$n_{xy|\boldsymbol{w}} \left[\exp\left(2\hat{\mathcal{I}}_{x;y|\boldsymbol{w}}\right) - 1\right] \sim F(1, n_{xy|\boldsymbol{w}}). \tag{34}$$

The null distributions we provide above explicitly account for autocorrelation via the effective degrees of freedom $n_{xy|\boldsymbol{w}}$ and also reduce to the F-distribution for information-theoretic quantities when observations of the analyzed time series are independent [cf. Eq. (17) by letting $\hat{\eta}(x, y) = T$]. Further, when x and y are serially uncorrelated, and in the limit $T \to \infty$, Eq. (34) becomes equivalent to the χ^2 null distribution for mutual information (see the discussion in Appendix A 2). Thus, the null distribution we present in Eq. (34) is a generalization of both the standard F-test (which is applicable only for i.i.d. variables) as well as the asymptotic distribution (which is applicable only for infinite data).

Although they are special cases of the modified Λ-test, there are important distinctions here from the χ^2-tests (16) and the standard F-tests (17) for conditional mutual information. The first is that we now have an effective sample size $\hat{\eta}$ that changes depending on the autocorrelation function of the residuals $e_{x|\boldsymbol{w}}$ and $e_{y|\boldsymbol{w}}$. The second is that the degrees of freedom $n_{xy|\boldsymbol{w}}$ is further reduced by $c = \dim(\boldsymbol{w}(t))$, the dimension of the conditional time series $\boldsymbol{w}$, which appears in the finite-sample F-tests but not the asymptotic χ^2-tests. Both of these differences introduce a significant bias in the estimation of linear dependence for many real-world applications, exemplified by the numerical simulations in Sec. VII A.

B. Multiple time series

Mutual information $\mathcal{I}_{\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{W}}$ can also be used to measure the dependence between multivariate processes $\boldsymbol{X}$ and $\boldsymbol{Y}$. Here we apply the chain rule and the results from the previous section to obtain a partial correlation decomposition that can be used for constructing a null distribution in the presence of autocorrelation.

The chain rule provides a decomposition of mutual information as a sum of conditional mutual information terms:

$$\hat{\mathcal{I}}_{\boldsymbol{x};\boldsymbol{y}|\boldsymbol{w}} = \sum_{g=1}^{k} \sum_{h=1}^{l} \hat{\mathcal{I}}^{\{gh\}}_{\boldsymbol{xy}|\boldsymbol{w}}. \tag{35}$$

That is, mutual information estimates $\hat{\mathcal{I}}_{\boldsymbol{x};\boldsymbol{y}|\boldsymbol{w}}$ between a k-variate process $\boldsymbol{x}$ and an l-variate process $\boldsymbol{y}$, in the context of the c-variate concomitant $\boldsymbol{w}$, can be computed by summing over conditional mutual information terms [33]. Each conditional mutual information term may be expressed as

$$\hat{\mathcal{I}}^{\{gh\}}_{\boldsymbol{xy}|\boldsymbol{w}} = \hat{\mathcal{I}}_{x_g;y_h|\boldsymbol{v}^{\{gh\}}_{\boldsymbol{xy}|\boldsymbol{w}}}, \tag{36}$$

where the conditional for the (g, h)-term is given by

$$\boldsymbol{v}^{\{gh\}}_{\boldsymbol{xy}|\boldsymbol{w}} = \begin{bmatrix} \boldsymbol{x}_{1:g-1} \\ \boldsymbol{y}_{1:h-1} \\ \boldsymbol{w} \end{bmatrix}, \tag{37}$$

with $\boldsymbol{x}_{1:g} = [x_1', \ldots, x_g']'$ a $g \times T$ matrix when $0 < g \leqslant k$, and the empty set, $\boldsymbol{x}_{1:g} = \emptyset$, when $g = 0$. Using this notation, we have an equivalent expression for mutual information as

$$\begin{aligned}\exp\left(-2\,\hat{\mathcal{I}}_{\boldsymbol{x};\boldsymbol{y}|\boldsymbol{w}}\right) &= \frac{|s_{\boldsymbol{xy}|\boldsymbol{w}}|}{|s_{\boldsymbol{x}|\boldsymbol{w}}||s_{\boldsymbol{y}|\boldsymbol{w}}|} \\ &= \prod_{g,h} \exp\left(-2\,\hat{\mathcal{I}}^{\{gh\}}_{\boldsymbol{xy|w}}\right) \\ &= \prod_{g,h}\left(1 - r^2_{x_g y_h \cdot \boldsymbol{v}^{\{gh\}}_{\boldsymbol{xy|w}}}\right). \qquad (38)\end{aligned}$$

Although this equation has a similar form to the well-known canonical correlation decomposition [30,53], the correlations are over different variables. More importantly for our purposes, and assuming independence of these partial correlations (see below), its null distribution can thus be obtained from the Λ^*-distribution:

$$\exp\left(-2\,\hat{\mathcal{I}}_{\boldsymbol{x};\boldsymbol{y}|\boldsymbol{w}}\right) \sim \Lambda^*(\boldsymbol{n}_{\boldsymbol{xy}|\boldsymbol{w}}), \qquad (39)$$

with parameter vector

$$\boldsymbol{n}_{\boldsymbol{xy}|\boldsymbol{w}} = \left(n^{\{11\}}_{\boldsymbol{xy}|\boldsymbol{w}}, \ldots, n^{\{kl\}}_{\boldsymbol{xy}|\boldsymbol{w}}\right)'. \qquad (40)$$

The remaining challenge is to compute the effective degree of freedom, $n^{\{gh\}}_{\boldsymbol{xy}|\boldsymbol{w}}$, for each independent partial correlation in Eq. (38). Recall that partial correlation can be computed from ordinary least squares. The residual vector for the (g, h)-term in Eq. (38) is

$$e^{\{gh\}}_{\boldsymbol{x}|\boldsymbol{v}} = x_g - \hat{x}_g\left(\boldsymbol{v}^{\{gh\}}_{\boldsymbol{xy}|\boldsymbol{w}}\right), \qquad (41)$$

$$e^{\{gh\}}_{\boldsymbol{y}|\boldsymbol{v}} = y_h - \hat{y}_h\left(\boldsymbol{v}^{\{gh\}}_{\boldsymbol{xy}|\boldsymbol{w}}\right), \qquad (42)$$

where the hat in $\hat{x}_g(\cdot)$ again denotes the linear prediction of x_g from the input argument. Then the degrees of freedom used in computing the (g, h) sample partial correlation are

$$n^{\{gh\}}_{\boldsymbol{xy}|\boldsymbol{w}} = \hat{\eta}\left(e^{\{gh\}}_{\boldsymbol{x}|\boldsymbol{v}}, e^{\{gh\}}_{\boldsymbol{y}|\boldsymbol{v}}\right) - \dim\left(\boldsymbol{v}^{\{gh\}}_{\boldsymbol{xy}|\boldsymbol{w}}(t)\right) - 2. \qquad (43)$$

Throughout this analysis, we ordered the summations first over the dimensions of $\boldsymbol{y}$, and then over the dimensions of $\boldsymbol{x}$. In practice, the order of these operations are arbitrary and was initially imposed solely for clarity in the chain-rule formula.

It should be noted that the modified Λ-test assumes both that the residuals are completely independent and that the estimated degrees of freedom is approximately correct. If, instead, the residuals become slightly correlated due to statistical errors in the regression, the Λ^*-distribution should be generated by sampling dependent beta- or F-distributed variables. Further, the effective degree of freedom is a first-order approximation, which may introduce biases in the hypothesis tests. Although our numerical simulations in Sec. VII show no such biases, we discuss the potential solutions in Appendix D.

VI. MODIFIED TESTS FOR GRANGER CAUSALITY

Recall that Granger causality can be expressed as a conditional mutual information [see Eq. (20)]. As such, we can leverage results from the previous section to introduce its null distribution. Many other information-theoretic and likelihood ratio-based measures could be similarly decomposed (from Wilks' criterion or conditional mutual information) in order to derive their finite-sample hypothesis tests.

A. Two time series

We shall first express the Granger causality for bivariate processes as a sum of conditional mutual information terms via the chain rule. Let upper indices (without parentheses) denote a backshifted variable, e.g., $\boldsymbol{X}^j(t) = \boldsymbol{X}(t - j)$ denotes the variable $\boldsymbol{X}(t)$ lagged by j time indices. Then, by applying the chain rule (35) to Granger causality (20), we can compute it as a sum of conditional mutual information estimates:

$$\hat{\mathcal{F}}_{y\to x|\boldsymbol{w}}(p, q) = 2\sum_{j=1}^{q} \hat{\mathcal{I}}_{x;y^j|\boldsymbol{v}^{\{j\}}_{y\to x|\boldsymbol{w}}}, \qquad (44)$$

with the jth conditional as the matrix

$$\boldsymbol{v}^{\{j\}}_{y\to x} = \begin{bmatrix} \boldsymbol{x}^{(p)} \\ \boldsymbol{y}^{(j-1)} \\ \boldsymbol{w} \end{bmatrix}, \qquad (45)$$

and the limiting case giving $\boldsymbol{y}^{(0)} = \emptyset$, i.e., the empty set. Again, following Eq. (39) we conclude that under the null hypothesis $\mathcal{H}_0 : \mathcal{F}_{Y\to X}(p, q) = 0$, Granger causality estimates are distributed as follows:

$$\exp\left[-\hat{\mathcal{F}}_{y\to x|\boldsymbol{w}}(p, q)\right] \sim \Lambda^*(\boldsymbol{n}_{y\to x|\boldsymbol{w}}), \qquad (46)$$

where

$$\boldsymbol{n}_{y\to x|\boldsymbol{w}} = \left(n^{\{1\}}_{y\to x|\boldsymbol{w}}, \ldots, n^{\{q\}}_{y\to x|\boldsymbol{w}}\right)'. \qquad (47)$$

Now, we can use the same approach from Sec. V to obtain the effective degrees of freedom used in computing Granger causality estimates. First, the residuals for the jth partial correlation in Eq. (44) are

$$e^{\{j\}}_{x|\boldsymbol{v}_{y\to x|\boldsymbol{w}}} = x - \hat{x}\left(\boldsymbol{v}^{\{j\}}_{y\to x|\boldsymbol{w}}\right), \qquad (48)$$

$$e^{\{j\}}_{y|\boldsymbol{v}_{y\to x|\boldsymbol{w}}} = y^j - \hat{y}^j\left(\boldsymbol{v}^{\{j\}}_{y\to x|\boldsymbol{w}}\right). \qquad (49)$$

Thus the number of degrees of freedom is different for each term, with the jth number computed as

$$n^{\{j\}}_{y\to x|\boldsymbol{w}} = \hat{\eta}\left(e^{\{j\}}_{x|\boldsymbol{v}_{y\to x|\boldsymbol{w}}}, e^{\{j\}}_{y|\boldsymbol{v}_{y\to x|\boldsymbol{w}}}\right) - \dim\left[\boldsymbol{v}^{\{j\}}_{y\to x|\boldsymbol{w}}(t)\right] - 2. \qquad (50)$$

Unlike the standard F-test for Granger causality [Eq. (24)], the modified Λ-test takes into account the effective number of degrees of freedom induced by autocorrelation in both the predictee and predictor processes, with the two approaches overlapping only when there is no autocorrelation in the residuals. This indicates that, with limited data, the F-test can be used only for assessing the significance of Granger causality estimates from *independent observations* (y) to univariate autocorrelated time series (x).

B. Multiple time series

Finally, we present hypothesis tests for the most complex linear dependence measure in the paper: the Granger causality from an l-variate predictor process $\boldsymbol{Y}$ to a k-variate predictee process $\boldsymbol{X}$, in the context of the c-variate concomitant process $\boldsymbol{W}$. By virtue of the chain rule (35), this general expression of

Granger causality (20) decomposes into three nested sums:

$$\hat{\mathcal{F}}_{\boldsymbol{x}\to\boldsymbol{y}|\boldsymbol{w}}(p,q) = 2\sum_{g=1}^{k}\sum_{h=1}^{l}\sum_{j=1}^{q}\hat{\mathcal{I}}_{x_g y_h^j|\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}^{\{ghj\}}}, \tag{51}$$

where the conditional for the (g, h, j) mutual information term is

$$\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}^{\{ghj\}} = \begin{bmatrix} \boldsymbol{x}_{1:g}^{(p)} \\ \boldsymbol{y}_{1:h-1}^{(q)} \\ \boldsymbol{y}_{h}^{(j-1)} \\ \boldsymbol{w} \end{bmatrix}. \tag{52}$$

That is, the (g, h, j) term is the conditional Granger causality for dimension g of $\boldsymbol{X}$ and the predictor observation j steps back of the hth subprocess of $\boldsymbol{Y}$. This is conditioned on all dimensions and predictor observations below g, h, and j, as well as on $\boldsymbol{W}$. Again, following Eq. (39) the distribution of Granger causality estimates, under the null hypothesis $\mathcal{H}_0\text{:}\mathcal{F}_{\boldsymbol{X}\to\boldsymbol{Y}}(p,q) = 0$, is given as

$$\exp\,[\hat{\mathcal{F}}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}(p,q)] \sim \Lambda^*(\boldsymbol{n}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}), \tag{53}$$

where

$$\boldsymbol{n}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}} = \big(n_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}^{\{111\}}, \dots, n_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}^{\{klq\}}\big)'.$$

The effective degrees of freedom are, again, computed from the residuals, where the residual processes used in computing the (g, h, j) partial correlation are

$$e_{\boldsymbol{x}|\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}}^{\{ghj\}} = x_g - \hat{x}_g\big(\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}^{\{ghj\}}\big), \tag{54}$$

$$e_{\boldsymbol{y}|\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}}^{\{ghj\}} = y_h^j - \hat{y}_h^j\big(\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}^{\{ghj\}}\big). \tag{55}$$

The number of degrees of freedom for each term in the chained sum can then be computed from these residuals:

$$n_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}^{\{ghj\}} = \hat{\eta}\big(e_{\boldsymbol{x}|\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}}^{\{ghj\}}, e_{\boldsymbol{y}|\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}}^{\{ghj\}}\big) - \dim\big(\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}^{\{ghj\}}(t)\big) - 2. \tag{56}$$

We note that, although the same p and q are used for each of the subprocesses of $\boldsymbol{x}$ and $\boldsymbol{y}$ in our presentation, our decomposition facilitates setting an individual history length for each term in the chained sum. The only difference would be to infer the optimal history length (p and q) for each residual vector in the chain.

VII. NUMERICAL VALIDATION

We perform numerical simulations in order to validate the modified Λ-test and characterize the effect of autocorrelation on the (unmodified) F- and χ^2-tests. The simulations, detailed in Appendix B 1, involve generating observations from two first-order independent AR processes and iteratively filtering the output signal such that the autocorrelation is increased for both time series.

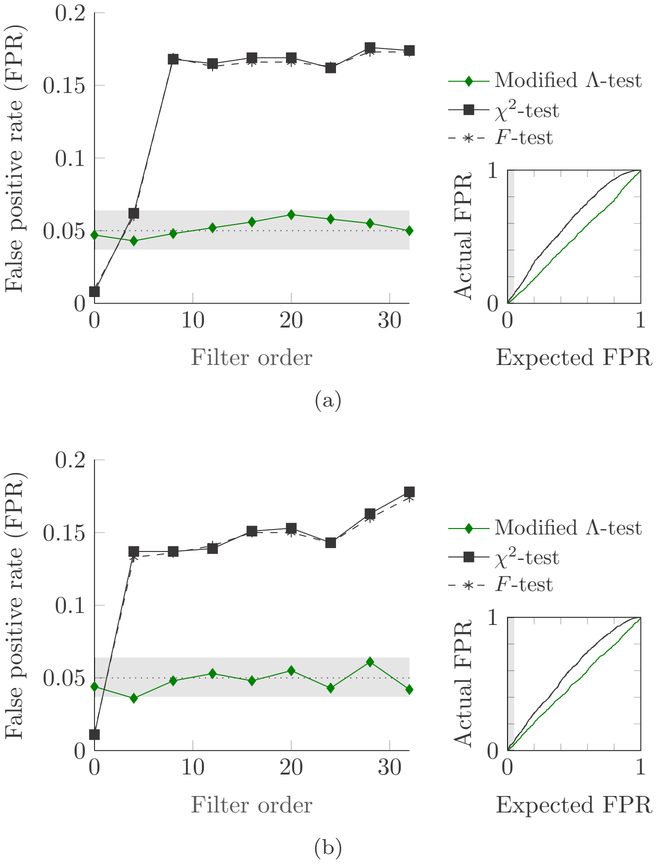
Following Barnett and Seth [4], we illustrate the finite-sample effects by generating relatively short stationary time series with $T = 2^9 = 512$ observations from the stochastic processes to obtain our dataset. We begin by sampling first-order AR processes to obtain our time-series data, $\boldsymbol{x}, \boldsymbol{y}$, and $\boldsymbol{w}$. Then, to illustrate the effect of higher autocorrelation on the FPR of both methods, we digitally filter each time series along the time dimension with two types of low-pass causal filters: a finite-impulse response (FIR) linear-phase least-squares filter and an infinite-impulse response (IIR) Butterworth filter. The filter order is variable, with higher filter orders generally increasing the autocorrelation of the time series. For each experiment, we perform 1000 trials and, using the statistical hypothesis-testing procedure described in Appendix A 4, we consider the FPR to be the proportion of p-values that are significant at the nominal level (typically 5% in this paper) in comparison to the relevant null distribution.

These experiments allow us to study how each test behaves under increasing levels of autocorrelation of both $\boldsymbol{x}$ and $\boldsymbol{y}$, whilst ensuring that the null hypothesis (of no dependence) is not violated. Rather than using a filter, it would of course be possible to increase the autocorrelation by selecting the ARMA parameters, $\boldsymbol{\Phi}(u)$ and $\boldsymbol{\Theta}(u)$, for each lag u. However, the formulations are equivalent: AR processes are all-pole IIR filters; MA processes are FIR filters; and ARMA processes are IIR filters with both poles and zeros. Thus, although these are identical formulations, we opt for digitally filtering processes to increase their autocorrelation as this is a common preprocessing step performed by practitioners to remove artifacts from time series (even differencing the signal is a type of filter). Moreover, as previously discussed, filtering the signals has been shown in the past to bias various dependence measures such as Granger causality [4]. Until this work, however, it has not been suggested that this bias is a function of autocorrelation nor has a valid hypothesis test been proposed based on the autocorrelation function.

A. Mutual information tests for bivariate time series

First, we use this approach to evaluate the performance of the hypothesis tests on assessing the significance of mutual information estimates between two independent (but serially correlated) time series. Our results are shown in Fig. 1, where the "F-tests" are from the finite-sample distribution [Eq. (17)], "χ^2-tests" refer to the asymptotic LR distributions [Eq. (16)], and the "Modified Λ-tests" refer to our hypothesis tests that account for autocorrelation [Eq. (33)]. As the plots illustrate, both the F-tests and the χ^2-tests overestimate the measures for higher filter orders (and therefore higher AR orders), yielding over 15% of false positives at the nominal significance of $\alpha = 0.05$—approximately three times the FPR expected from the test. The figures on the right illustrate the significance level α against the FPR for an eighth-order filter. From these figures, we can see that the FPR for the χ^2-tests is higher than nominal for all significance levels $\alpha \in (0, 1)$. In comparison, the modified Λ-test procedure yields the expected FPR for all filter orders.

A filter order of zero in Fig. 1 refers to generating the time-series data with the first-order AR model (B2) without any digital filtering. In this case, the χ^2- and F-tests yielded less than the nominal 5% FPR. This occurs when the number of effective samples becomes greater than the original sample size $\eta(x, y) > T$. Referring to Bartlett's formula (9), this is due to the product of negative autocorrelation exhibited by the Y process (induced by $\Phi_Y = -0.8$, indicating an effect of undersampling) and the positive autocorrelation exhibited by the X process (induced by $\Phi_X = 0.3$). Counterintuitively, this



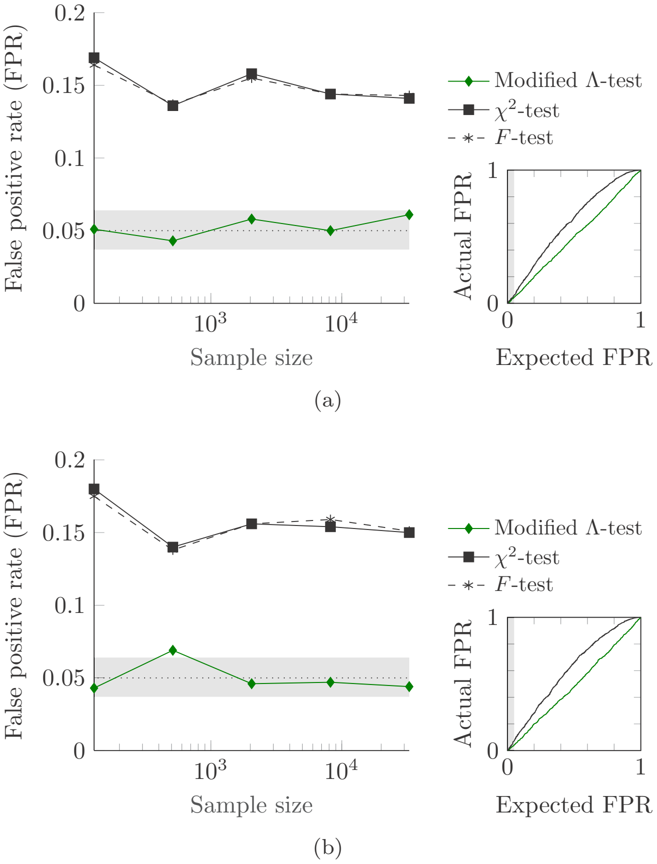
(a)

(b)

FIG. 1. Modified tests correctly assess the significance of the mutual information estimated between two univariate time series under FIR (a) and IIR (b) filtering. The mutual information was measured [using Eq. (18)] between independent univariate time series [generated by Eq. (B2)] after filtering, and tested using the χ^2-test [Eq. (16)], F-test [Eq. (17)], and the modified Λ-test [Eq. (33)]. The plots show the effect of increasing the filter order on the FPR for both an (a) FIR filter and (b) IIR filter. The shaded regions on the right indicate $\alpha = 0.05$, whilst the shaded regions on the left show the 95% confidence interval for the FPR (as defined in Appendix A 4). The subplots on the right capture the FPR for all potential significance levels α ("Expected FPR") with an eighth-order filtered signal. An ideal distribution is where the FPR equals α and thus sits perfectly on the diagonal, as per our tests.

would imply that the observations are anticorrelated in time. Such conservative results for the χ^2- and F-tests are likely to induce lower statistical power [i.e., a lower true-positive rate (TPR)] in scenarios when the effective sample size is greater than the original sample size. To verify this, we performed 1000 trials where there was a small dependence of X on Y (see Appendix B 1). The TPR was 0.049 (SE of 0.0068) for the χ^2-test and 0.1570 (SE of 0.0115) for the modified Λ-test. Thus, the power of our hypothesis test is three times greater than the χ^2-test in this scenario.

In Fig. 2 we show that increasing the sample size does not mediate the effect of autocorrelation on the χ^2- and F-tests. This is due to the fact that the effective degree of freedom is always a fraction of the degree of freedom. Thus, regardless of the sample size, both the asymptotic (χ^2) and finite (F) tests are invalid, unless modified to account for effective sample size.



(a)

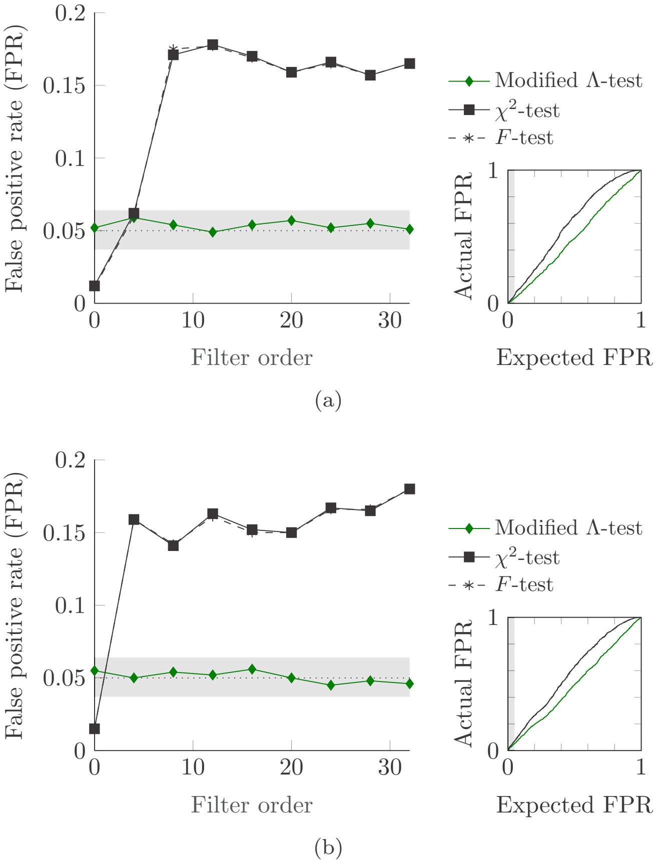
(b)

FIG. 2. Increasing the sample size does not mediate the effect of autocorrelation on the χ^2- and F-tests for mutual information. We perform the same tests as Fig. 1, except with an exponentially increasing sample size and a fixed eighth-order FIR (a) and IIR (b) filter. The subplots on the right show the FPR for $T = 2^{11}$ samples.

B. Conditional mutual information tests for bivariate time series

We now extend the previous results by evaluating the effect of conditioning mutual information between x and y on an independent, tertiary process $\boldsymbol{w}$. The FPRs for the χ^2-tests, F-tests, and the modified Λ-tests from these experiments are presented in Fig. 3, and exhibit similar characteristics to those of the mutual information tests in Fig. 1. Increasing the filter order generally increases the FPR for both unmodified tests, yet the modified Λ-test remains unbiased, maintaining the expected FPR of 5%.

As discussed in Sec. V, an important distinction between the null distributions for mutual information [Eq. (33)] is that the effective degree of freedom in the Λ*-distribution not only includes the effective sample size but also the dimension of the conditional c. To show the severity of the asymptotic approximation, we generate the x and y processes and filter the signal with an eighth-order FIR and IIR filter the same as before; however, we increase the number of independent processes $c = \dim(\boldsymbol{W}(t))$ in the multivariate conditional. The



(a)

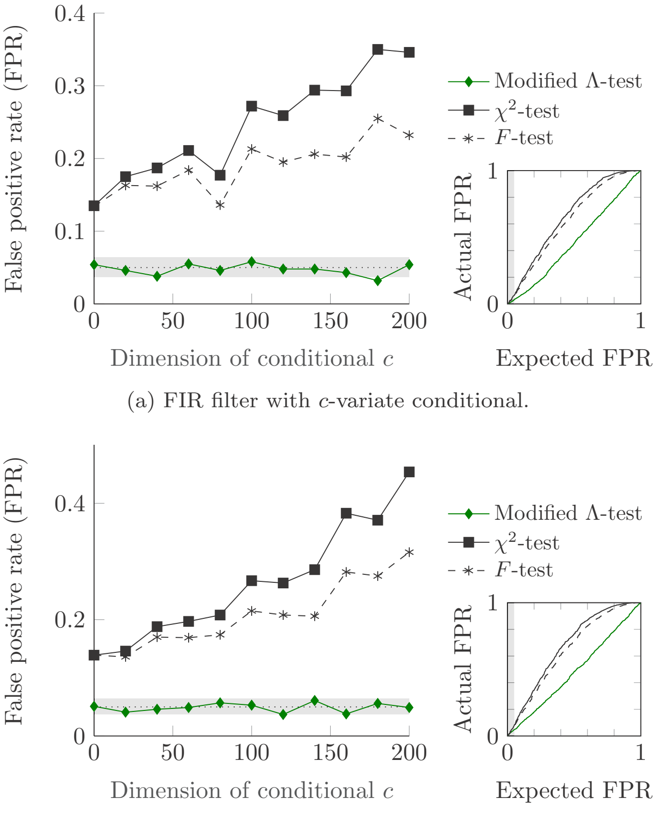
(b)

FIG. 3. Modified tests correctly assess the significance of the conditional mutual information estimated between two univariate time series, conditioned on a third univariate time series; each time series underwent FIR (a) or IIR (b) filtering. Conditional mutual information was measured [using Eq. (18)] between two univariate time series, given a third univariate process [generated by Eq. (B2) with $k = l = c = 1$], after filtering, and tested for significance using the χ^2-, F-, and modified Λ-tests. The subplots on the right show the FPR for each significance level α when the signal is filtered with an eighth-order filter.

result is shown in Fig. 4, where the FPR of the F- and χ^2-tests increases somewhat linearly with the dimension; however, the modified Λ-test remains unbiased. As evidenced by the improvement of the unmodified F-test over the χ^2-test, this experiment demonstrates that even when the autocorrelation function is the same, the dimension of the conditional must also be included in the hypothesis tests.

C. Mutual information tests for multivariate time series

In this section, we present results for the hypothesis tests of mutual information between multivariate ($m > 2$) time series. The multivariate time series are partitioned into two independent sets of processes, $\boldsymbol{X}$ and $\boldsymbol{Y}$, one with dimension k and one with dimension l. For each experiment, we let $k = l$ and use the state equations in Eq. (B2) to simulate $m = k + l$ independent AR processes for $m = \{2, 4, \dots, 10\}$. These signals are then filtered along the temporal dimension using eighth-order FIR and IIR filters. This signal generation process ensures that there is no correlation between signals within the same subprocess, i.e., $\rho_{ij} = 0$ for all $i, j \in [1, m]$. The results are shown in Fig. 5, where increasing the dimension approximately linearly increases the FPR of the original LR test to over 70% for both filters (continuing to increase for larger k and l), yet the modified Λ-test remains unbiased.



(a) FIR filter with c-variate conditional.

(b) IIR filter with c-variate conditional.

FIG. 4. The FPR of asymptotic tests increase with the dimension of the c-variate conditional process. The plots are the same as per Fig. 3, except as a function of c with a fixed eighth-order FIR (a) and IIR filter (b). The subplots on the right show the FPR for each significance level α when $c = 100$.

In the tests above, no correlations between subprocesses were included [e.g., $X_i(t)$ with $X_j(t-u)$ for $j \neq i$]; however, an internal cross-correlation between any of these subprocesses may further decrease the size and power of unmodified tests. Our experiments of Granger causality in the following sections naturally incorporate examples with correlated subprocesses in the mutual information calculation.

D. Granger causality tests for bivariate time series

This section examines the performance of each hypothesis test on estimates of Granger causality using the same (univariate) simulations from Sec. VII A. That is, the bivariate AR model [Eq. (B2) with $k = l = 1$ and no conditional $c = 0$] is simulated to generate $T = 512$ observations of the X and Y processes, which are then passed through FIR and IIR filters. Referring to Fig. 6, we perform this with each filter order and each filter type (FIR and IIR). After generating these sample paths, the AR order of the predictee, p, and the predictor, q, were inferred from Burg's method [58]. We then compute

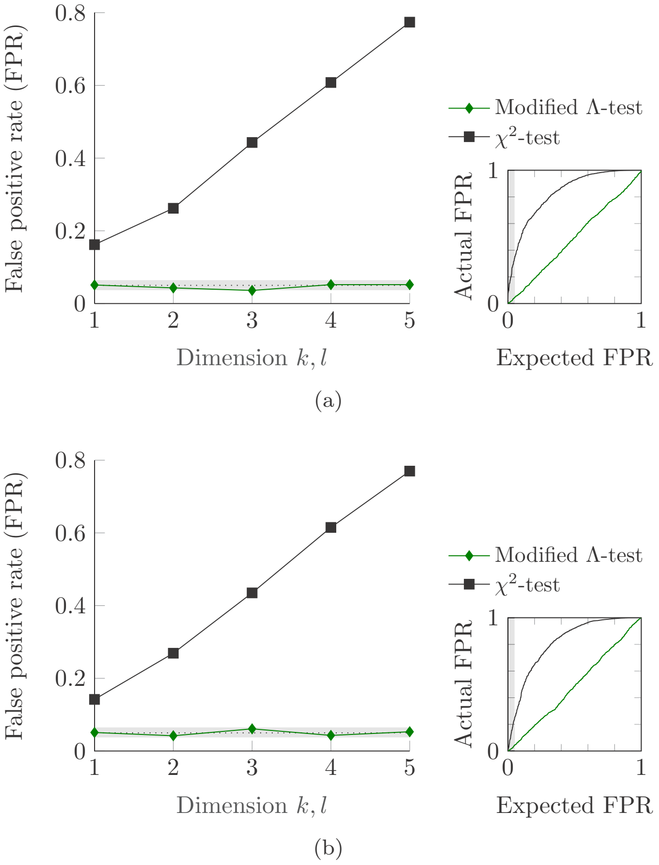


FIG. 5. Modified Λ-tests correctly assess the significance of the mutual information estimated between two multivariate time series. The mutual information was measured [using Eq. (38)] between the multivariate time series [generated by Eq. (B2) with an increasing dimension k and l of $\boldsymbol{X}$ and $\boldsymbol{Y}$], passed through eighth-order FIR (a) and IIR (b) filters, and tested for significance using the χ^2-test [Eq. (16)] and the modified Λ-test [Eq. (39)]. The subplots on the right show the FPR for each significance level α when $k = l = 3$.

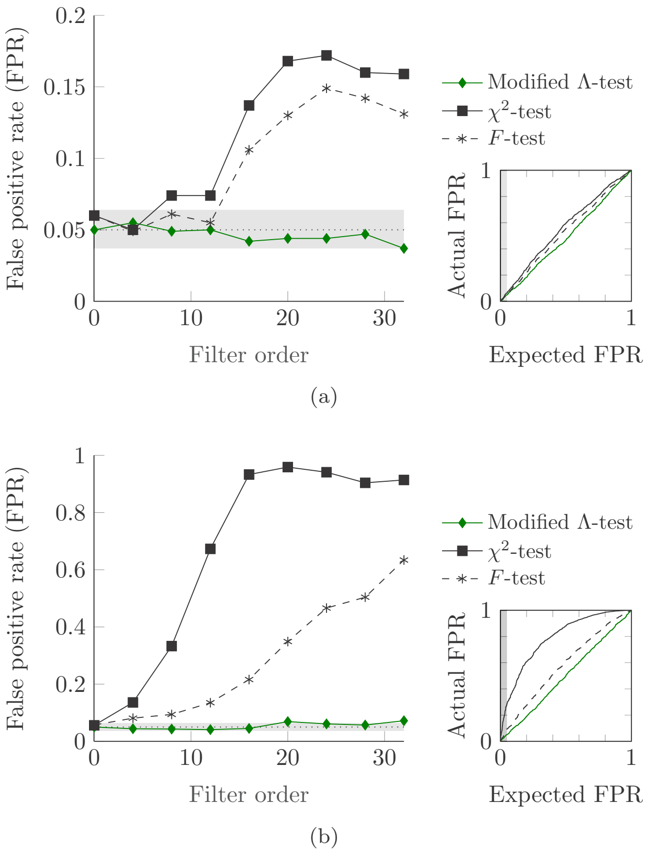


FIG. 6. Modified Λ-tests correctly assess the significance of the Granger causality estimated from one univariate time series to another. Granger causality, with history lengths p and q chosen via Burg's method, is estimated [using Eq. (44)] between univariate time series [generated by Eq. (B2)] after smoothing with an FIR (a) and an IIR (b) filter. Estimates are then tested using the χ^2-test [Eq. (22)], the F-test [Eq. (24)], and the modified Λ-test [Eq. (46)]. The subplots on the right show the FPR for each significance level α with an eighth-order filter.

Granger causality [via Eq. (44)] and use this estimate to obtain p-values from the CDFs of the F-distribution [Eq. (24)], the χ^2-distribution [Eq. (22)], and the Λ^*-distribution [Eq. (46)]. This is performed 1000 times in order to obtain a FPR of each approach. Whilst our experiment here is equivalent to a conditional mutual information, unlike the experiments in previous sections the autocorrelation within the time series naturally induces a cross-correlation amongst the variables within the predictor $\boldsymbol{Y}^{(q)}(t)$ and conditional $\boldsymbol{X}^{(p)}(t)$ processes. The results shown here illustrate that increasing the autocorrelation length via filtering increases the FPR of Granger causality under the F- and χ^2-tests, particularly when using an IIR filter. In contrast, the FPR of Granger causality using the modified Λ-tests remains mostly unbiased. It should be noted that, although within the confidence bounds, the FPR of the modified Λ-tests appear to be not exact for high-order filters; sources of error regarding this potential bias are discussed in Appendix D.

The model order for our experiments above was chosen using Burg's method; however, the are numerous approaches to inferring the "optimal" AR order as outlined in the introduction, all of which can result in vastly different model orders. To ensure our results are not a consequence of poor model identification, in Fig. 7 we illustrate the FPRs for increasing the predictor history length q from one to 200 (in increments of 20), whilst holding the autocorrelation length constant with an eighth-order filter. This effectively introduces more terms in Eq. (44), causing a larger divergence between the F-, χ^2- and modified Λ-tests. As expected, the FPR of the χ^2- and F-tests linearly increases in this range, whereas the modified Λ-test remains consistent with the 5% FPR. This linear increase of the FPR in χ^2- and F-tests is somewhat counterintuitive to the notion of Granger causality, where one may expect that accounting for more history would reduce spurious correlations. However, the opposite is true, simply due to a lack of correct finite-sample distributions (in the case of the unmodified tests).

In Fig. 8 we show the effect of increasing the sample size for tests on Granger causality estimates. Here we can see the χ^2- and F-tests converging for sufficient sample sizes. Unlike mutual information (from Fig. 2), estimating Granger causality involves regressing the autocorrelation of the

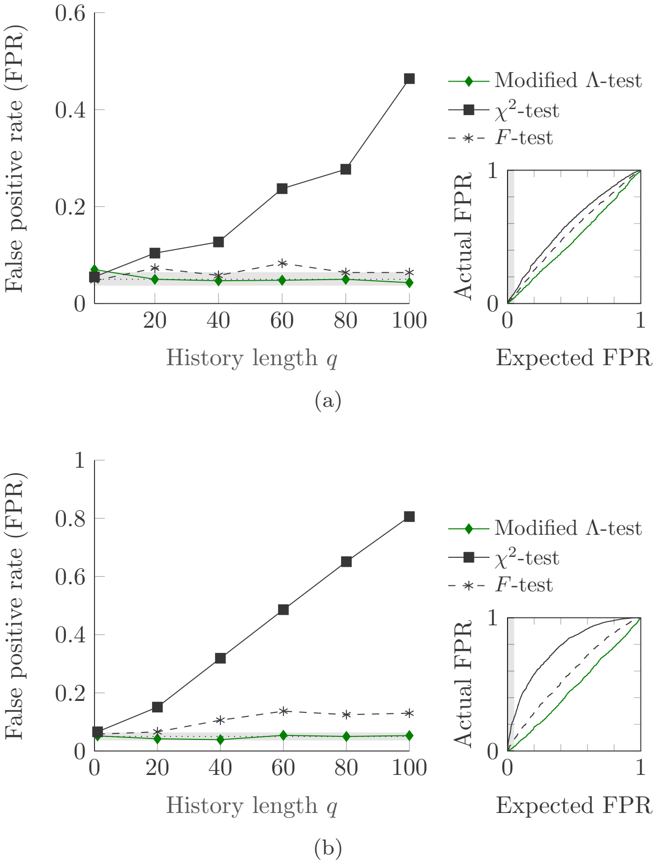


FIG. 7. The FPR of the F- and χ^2-tests for Granger causality increases with an increasing dependence on the past for time series generated by an eighth-order FIR (a) and IIR (b) filter. We illustrate this by varying the history length of the predictor process, q, from one to 200. The subplots on the right show the FPR for each significance level α when $q = 20$ (a) and $q = 40$ (b) to approximately match the orders chosen via Burg's method in Fig. 6.

predictee first, with the variance of these residuals reducing as the sample size grows. Thus, the effective sample size asymptotically approaches the sample size, however, the precise rate of this convergence depends the autocorrelation function and may change for every pair of time series. Even for this simple example, we see that on the order of 100 000 samples are required for convergence, which is not realistic in many empirical scenarios.

E. Granger causality tests for multivariate time series

Finally, we can evaluate the effect of increasing the dimensionality of both processes on Granger causality inference. In these experiments, we vary the dimension of the processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ from one to five. Recall from Eq. (51) that the number of terms involved in computing Granger causality (and its null distribution) is the product klq, of the dimensionality (k, l) and the history length of the predictor (q). Due to the relatively short time series length of $T = 512$ samples and high autocorrelation and dimensionality, allowing an arbitrary predictor history length of q results in the effective sample size approaching zero for the modified Λ-tests. Thus, for these experiments we fix the history length of the predictor to $q = 1$.

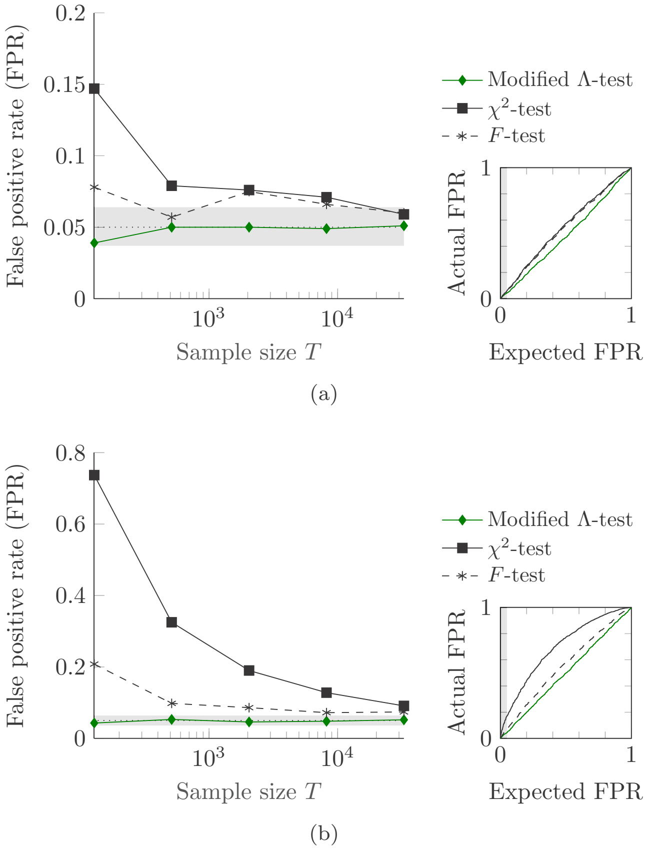


FIG. 8. The F- and χ^2-tests converge to the modified Λ-test for Granger causality with large sample sizes. This experiment is the same as Fig. 6, except with an exponentially increasing sample size and a fixed eighth-order FIR (a) and IIR (b) filter.

The results are shown in Fig. 9, where the χ^2-tests inflate the FPR close to 100% for with higher dimensional processes. Although our corrected tests perform well for moderate dimensionality, when $k, l > 3$ with the IIR filter, the FPR of our modified Λ-tests begin to have numerical issues. This is caused by the regression matrix not being well conditioned, i.e., the ratio of regressors to data points is too high. Nonetheless, we can see from the figure that our tests maintain a low FPR, becoming more conservative when the regression is ill-posed. Moreover, a poorly conditioned regression can be easily tested for in practice. So we conclude that with minimally sufficient observations, our tests maintain the desired FPR even for the most general case of multivariate Granger causality and, when the sample size is simply too small for reliable inference, our approach flags this as an issue.

VIII. EFFECT OF PREWHITENING

The rationale for applying prewhitening is to remove the autocorrelation in one time series such that the variance of computed statistics becomes equivalent to serially independent observations [12]. That is, instead of modifying the hypothesis tests, prewhitening modifies the input time series and thus the statistics themselves. Prewhitening is typically attempted by first inferring a model of one time series ($\boldsymbol{x}$), and transforming the process $\boldsymbol{x}$ to a residual process $\tilde{\boldsymbol{x}}$ through a

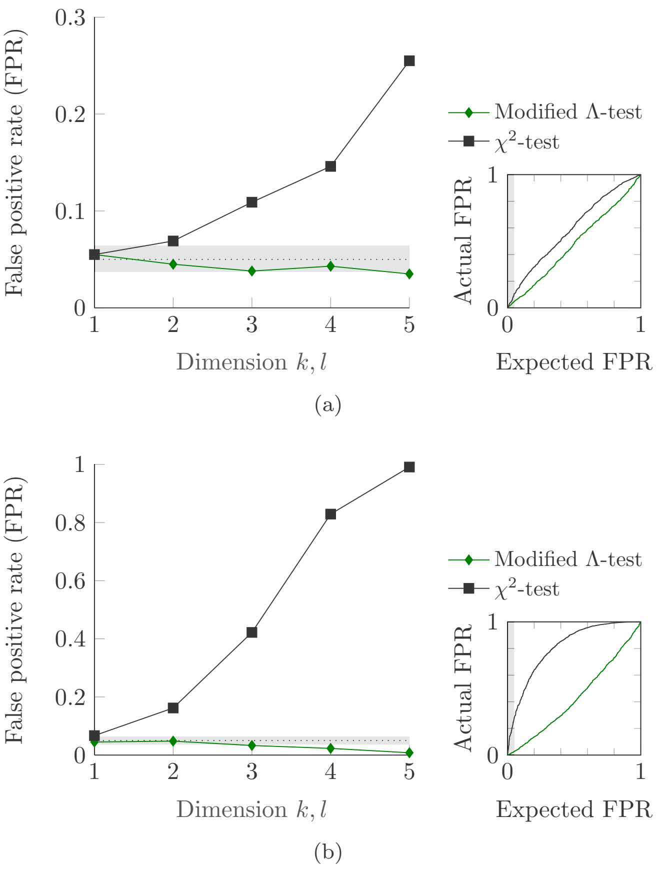


FIG. 9. Modified Λ-tests correctly assess the significance of the Granger causality estimated from one multivariate time series to another for increasing dimension k, l. Granger causality, with each predictee history length chosen optimally, is estimated [using Eq. (51)] between multivariate time series [generated by Eq. (B2)] after filtering via eighth-order FIR (a) and IIR (b) filters. Estimates are then tested using the χ^2-tests [Eq. (22)] and our modified Λ-test [Eq. (53)]. The subplots on the right show the FPR for each significance level α when $k = l = 3$.

filter constructed from its model. The same filter (with parameters inferred from $\boldsymbol{x}$) is then applied to the other time series $\boldsymbol{y}$ to create $\tilde{\boldsymbol{y}}$. Specifically, assuming any arbitrary ARMA(p, q) model for time series $\boldsymbol{x}$, prewhitening involves learning the parameter vectors $\hat{\boldsymbol{\Phi}}$ and $\hat{\boldsymbol{\Theta}}$ from $\boldsymbol{x}$, and then filtering the raw signals through the following equations:

$$\tilde{\boldsymbol{x}}(t) = \boldsymbol{x}(t) - \sum_{u=1}^{p} \hat{\boldsymbol{\Phi}}(u)\boldsymbol{x}(t-u) - \sum_{u=1}^{q} \hat{\boldsymbol{\Theta}}(u)\tilde{\boldsymbol{x}}(t-u),$$

$$\tilde{\boldsymbol{y}}(t) = \boldsymbol{y}(t) - \sum_{u=1}^{p} \hat{\boldsymbol{\Phi}}(u)\boldsymbol{y}(t-u) - \sum_{u=1}^{q} \hat{\boldsymbol{\Theta}}(u)\tilde{\boldsymbol{y}}(t-u).$$

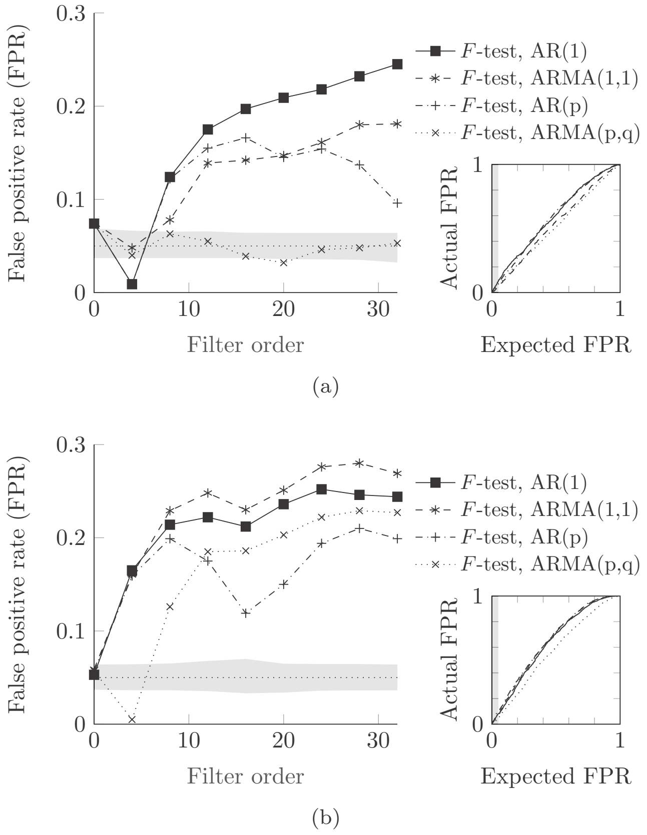
Using the same linear transformation (filter) for both time series renders their correlation theoretically invariant [12], however, the assumption is that $\tilde{\boldsymbol{x}}$ is now serially independent, and so the variance of sample correlations (for instance) converge to $\hat{\sigma}_r(x, y) = 1/T$.

A significant challenge in prewhitening signals is in selecting an appropriate model of the autocorrelation function. Of course, for the same arguments as presented in Sec. II, after mean removal and differencing, the most general model for covariance-stationary time series are ARMA models. However, inference procedures to learn both the order and parameters of ARMA models are computationally expensive. As such, many authors (and textbooks [12]) propose an AR(p) model would suffice to render the residuals, $\tilde{\boldsymbol{x}}$, independent, presuming that autocorrelations decay rapidly for stationary time series. Since this is the same assumption underlying Granger causality [regarding the residuals on the target after fitting an AR(p) model], our results from Sec. VII D suggest that statistics computed from signals prewhitened in this way will remain biased. In fMRI research, the most popular packages that are used for preprocessing time-series data (AFNI, SFL, and SPM) are similarly insufficient for handling autocorrelation due to their simplistic models [15]. Specifically, the package AFNI uses an ARMA(1,1) model learned from each voxel, whereas FSL uses Tukey tapering to smooth the data (see Appendix D), and SPM uses one global AR(1) model for all processes. Thus, each package assumes a fixed ARMA model can describe any arbitrary-order process. This is clearly insufficient, and if the wrong model is used, the residuals $\tilde{\boldsymbol{x}}$ remain dependent, resulting in an unknown variance of statistical estimates. This is evidenced by consistently high FPRs in empirical studies [15].

For completeness of this paper, however, we implement a number of prewhitening schemes in order to illustrate that such an approach is insufficient for assessing linear dependence between typical time series. Our experiments—on the same synthetic time series used in Sec. VII—show the effect of prewhitening univariate signals on the unmodified F-test for a number of different models: the AR(1) and ARMA$(1, 1)$ as well as the AR(p) and ARMA(p, q) with the optimal model orders (p and q) learned from data. For the AR(p) model inference, we allow for $p \in [1, 200]$ (where the sample size is $T = 512$), with the model order selected (and parameters inferred) using Burg's method [58]. Given the difficulty of learning higher-order ARMA(p, q) models, however, we restrict our search space, iterating through each potential $p, q \in \{1, \dots, 5\}$ and selecting the model with the lowest BIC score.

Our results for the performance of the F-test on mutual information estimates after prewhitening are shown in Fig. 10. Contrasting these results to Fig. 1, the only benefit of prewhitening appears to be for the FIR-filtered time series when an ARMA(p, q) model is used. In almost all other scenarios, the FPRs are either equivalent to, or worse than, the original tests. Figure 11 illustrates the effect of prewhitening on Granger causality F-tests (compare to Fig. 6 without prewhitening). Again, prewhitening appears to serve no benefit to tests for Granger causality, even for the relatively advanced ARMA(p, q) model. Concerningly, for IIR-filterd signals, the FPR increases to over 60% for all ARMA models with no scenario where the prewhitening approach results in an FPR within the expected range. In Appendix C we demonstrate similar results when using BIC and AICc scoring functions (as an alternative to Burg's method) to infer AR(p) models for prewhitening.

It is possible that, with an ideal model, the F- or χ^2-tests used for a prewhitened time series may be comparable (in size properties or FPR) to the modified Λ-test. However, even restricting our search space to an ARMA$(5, 5)$ model resulted



(a)

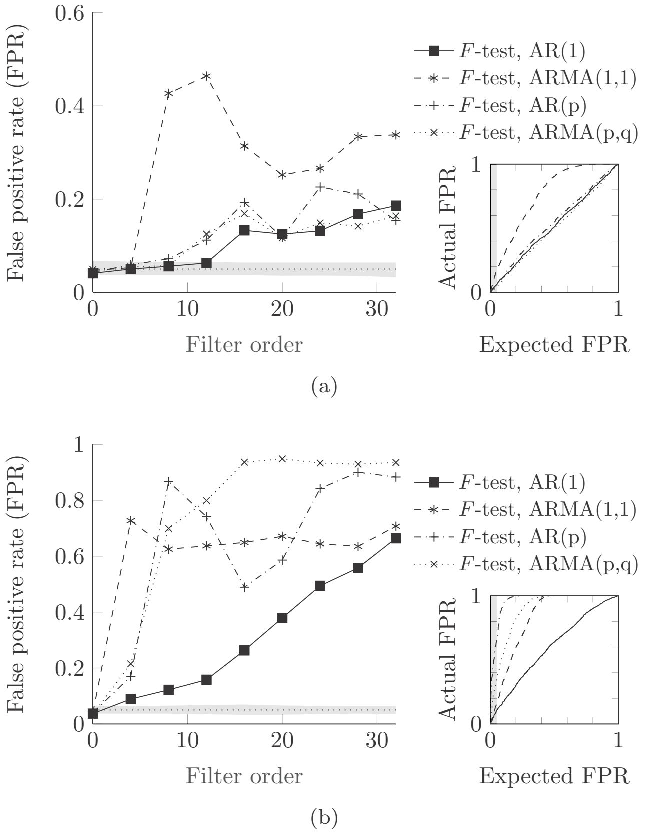
(b)

FIG. 10. Prewhitening the time series does not mediate the bias in F-tests for mutual information estimates after FIR (a) or IIR (b) filtering. The experiments are as per Fig. 1, with four prewhitening approaches used: the AR(1), ARMA(1, 1), AR(p), and ARMA(p, q). For the AR(p) and ARMA(p, q) models, the optimal order is inferred from Burg's method and the BIC score. In many cases for the IIR-filtered time series, either the AR(p) or the ARMA(p, q) failed to learn a stable model, and so these trials were removed, inducing nonuniform confidence intervals (reflected in the shaded region).

in an approximately 5000-fold increase in computational time over the modified Λ-test [62] and, moreover, remained biased in most scenarios. We conclude that, regardless of the model selected, the additional burden of prewhitening over using a modified hypothesis test is unjustified and that the outcome of unmodified hypothesis tests (with or without prewhitening) conveys inconsistent information about the underlying dependence structure between time series.

IX. CASE STUDY: HUMAN CONNECTOME PROJECT DATASET

Studies in computational neuroscience often leverage statistical analysis in order to formulate and test biologically plausible models. An important application in this field is the study of the human brain through fMRI, which is abundant with short, autocorrelated time series that have been studied using the measures of interest here [1–5]. As such, it is an archetypal real-world application to illustrate the issues of autocorrelation for time-series analysis.



(a)

(b)

FIG. 11. Prewhitening the time series does not mediate the bias in F-tests for Granger causality estimates after FIR (a) or IIR (b) filtering. The experiments are as per Fig. 6, with four prewhitening approaches used (as per Fig. 10): the AR(1), ARMA(1, 1), AR(p), and ARMA(p, q).

In fMRI research, the blood-oxygen level-dependent (BOLD) data are translated into a slowly varying (and thus highly autocorrelated) multivariate time series that traces the haemodynamic response of different locales (voxels) in the brain. Digital filtering is then commonly used as a preprocessing step to reduce line noise, nonstationarity and other artefacts in neuroimaging data. This induces an (either finite or infinite) impulse response that can increase autocorrelation, even if the original signals were not serially correlated. To characterize the FPR of linear dependence measures between empirical time series, we use completely independent time series from a widely accessed brain-imaging dataset known as the the Human Connectome Project (HCP) resting state fMRI (rsfMRI) dataset [44] (see Appendix B 2 for more detail; time series are selected from different random regions of interest from different random subjects to ensure independence, and then digitally filtered). This process is shown in Fig. 12, where a sample window of the raw and filtered data from two independent time series appears on the left panel.

First, we illustrate the effect of autocorrelation on mutual information by using a χ^2-test, with significance level 5%, computed with $T = 800$ samples of each independent time series. We begin by estimating mutual information $\hat{\mathcal{I}}_{x;y}$ between two unrelated time series x and y, sampled from the

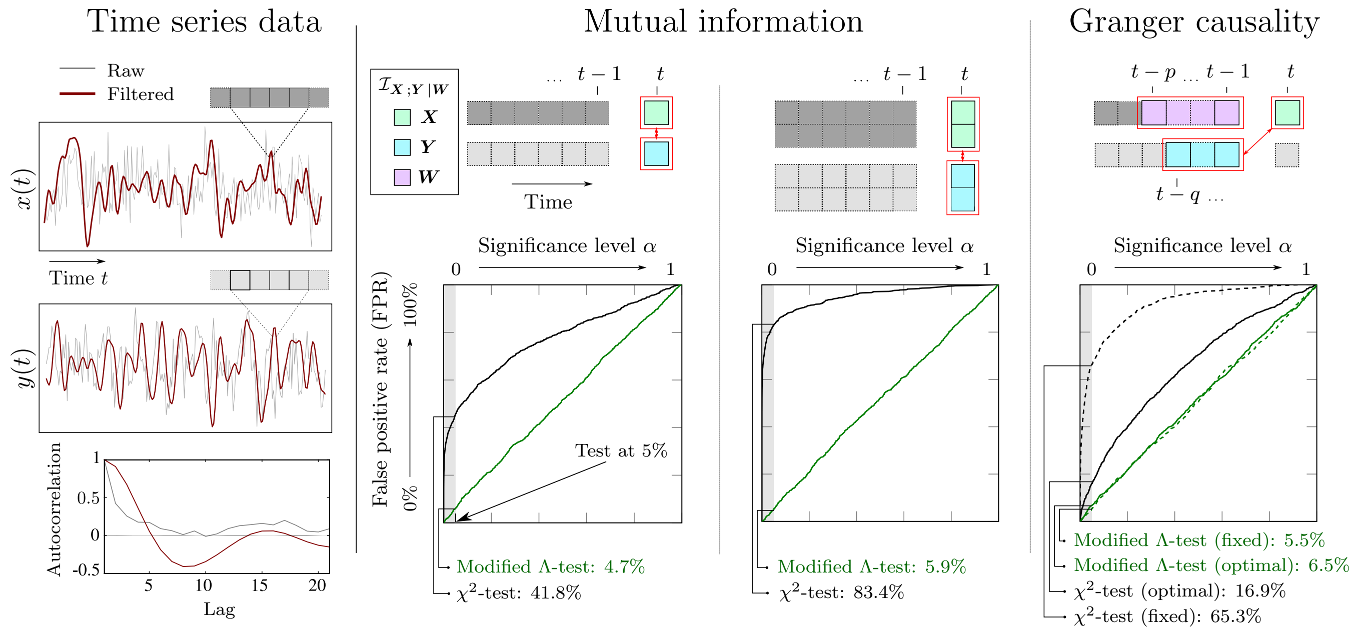


FIG. 12. Application to brain-imaging data, demonstrating our correction for the otherwise dramatic inflation of false-positive rates (FPRs) of classical hypothesis tests of dependence estimates in the presence of autocorrelation. We perform 1000 experiments with both the χ^2-tests as well as the modified Λ-tests for inferring the significance of two dependence measures: mutual information(between two univariate and two bivariate processes), and Granger causality. For each of these experiments, we randomly selected two uncorrelated fMRI time series, $\boldsymbol{X}$ and $\boldsymbol{Y}$, from the Human Connectome Project [44] (see Appendix B 2 for more details). A sample window of these univariate processes are shown in the left panel (gray lines), with the common preprocessing technique of band-pass digital filtering later applied to the signals (red lines). At the bottom of the left panel, we plot the sample autocorrelation function from both the raw and the filtered signal, illustrating the higher autocorrelation length (and lower effective sample size) induced by digital filtering. The panels on the right show the FPR of each test applied to the filtered signals as a function of the significance level α. For ideal hypothesis testing, we expect a line along the diagonal (i.e., the FPR equals the significance level). The χ^2-tests illustrate an increased FPR for all measures, as seen by both the plots and the FPR at 5% significance level. In contrast, our tests remain consistent with the expected FPR.

HCP dataset. We find that the test results are significantly biased, yielding an FPR of 41.8% (a ninefold increase of the expected rate). After prewhitening with an AR(p) filter, this increased to over 88% (not shown in the figure). The modified Λ-test demonstrated an FPR of 4.7%, which is well within the acceptable confidence interval. Next, we perform the same tests but with the mutual information between two sets of bivariate time series $\boldsymbol{x}$ and $\boldsymbol{y}$ ($k = l = 2$); this yields an 83% FPR (a 16-fold increase). When we correctly test these same measurements with the modified Λ-test [Eq. (39)], we find a FPR of 5.2%, matching the desired level within the confidence bounds.

For Granger causality, we perform two different tests: scenario (i) with the model orders, p and q, inferred for each trial, and scenario (ii) with a fixed $p = q = 100$ for all trials. For unrelated signal pairs, the FPR of Granger causality estimates using the χ^2-test is 16.9% (with optimal history length around 18), increasing further to 90.5% after prewhitening with an AR(p) filter. If a longer embedding length of 100 is chosen, the FPR is 66%—more than 13 times the expected value—increasing to 83.3% after prewhitening with an AR(p) model. When we test these same measurements with the modified Λ-test (46), we find a FPR of 5.5% and 6.5% for scenarios (i) and (ii), respectively, completely removing the false-positive bias exhibited by the χ^2-tests.

X. DISCUSSION

We have shown that the autocorrelation exhibited in covariance-stationary time-series data induces bias in the hypothesis tests of a broad class of linear dependence measures. By framing different dependence measures in unified theoretical terms, we provide the first demonstration of how Bartlett's formula can be applied to derive unbiased hypothesis tests, termed modified Λ-tests, for mutual information (and, consequently, Wilks' criterion and Granger causality) for both univariate and multivariate time-series data. These measures are used in a wide range of disciplines, modeling myriad important processes from anthropogenic climate change [26] to the brain dynamics of dementia patients [63]. The continued use of flawed testing procedures in empirical sciences is problematic, making it imperative that the corrections reported here be incorporated into future studies.

The effect of temporal autocorrelation on linear dependence has long been investigated in statistics; however, the majority of research has focused on simple linear correlations [1,2,10,11,40], which are restricted to measuring symmetric bivariate dependence structures. These studies, while representing important milestones in inferring the association between univariate time series, suffer from the inability to capture both multivariate and directed dynamical dependence. Even though measures such as mutual information and Granger causality can model a much richer set of dependence structures in temporal processes, the notion that their sampling properties could be altered in the presence of autocorrelation has thus far been largely overlooked. A major challenge of handling autocorrelation for more involved dependence measures was in extending Bartlett's formula to multivariate relationships. Crucially, our approach facilitates not only independent multivariate relationships but also correlated multivariate processes. Our results on the mutual information between multivariate time series used examples where the subprocesses were all independent. When extending this approach to Granger causality for arbitrarily large history lengths and dimensionality, the subprocesses become inherently correlated (since one subprocess is a time-lagged version of another, and they involve significant autocorrelation). This provides strong evidence that our approach is a generalization of Bartlett's dependence studies that is able to handle a much richer class of multivariate dependency structures beyond those already presented in this paper.

Granger causality is the *de facto* measure of directed dependence between stationary time series. The typical approach to assessing the significance of (potentially multivariate) estimates has been via the finite-sample F-tests [64] or the asymptotic χ^2-test [46]. In this study, we have shown that higher levels of autocorrelation (equivalently, a higher-order autoregression) in the signal inflates the variance of these statistical estimates, inducing significant bias for these traditional hypothesis tests. This means that using these tests induces errors when the predictor process is serially correlated—precisely the situation that Granger causality was designed to address. One might logically surmise that this issue could be mediated by accounting for a longer history of the process, i.e., conditioning on additional AR variables. Referring to Fig. 6, we have shown that this will result in the FPR being even further inflated. This is particularly concerning given that the autocorrelation function is one of the two properties that define a stationary process [7,8,47]—the other is its mean, which has no effect on scale-invariant dependence measures such as Granger causality, mutual information, and Pearson correlation. Similar to the thinking behind Granger causality, another common approach to remove autocorrelation is prewhitening, which intends to induce a serially independent process (of residuals) for hypothesis testing. Our experiments in Fig. 10 and Fig. 11 (as well as Appendix C) illustrated that many common approaches to prewhitening fail to control the FPR and, indeed, can further reduce the size of the unmodified (F and χ^2) hypothesis tests. We conclude that the unmodified hypothesis tests cannot be used to reliably infer the significance of Granger causality or mutual information estimates when applied to covariance-stationary time series and should be replaced with modified tests, such as the modified Λ-tests, particularly for limited time-series data.

Prior work [4] had already established that digital filtering led to biased Granger causality estimates for shorter time series, yet this effect was not understood nor able to be corrected until now. Due to the widespread use and influence of Granger causality across fields including neuroscience, ecology and economics, underlined by any examination of the literature (see Sec. I), this was a serious deficiency for directed inference of relationships in time-series analysis. Much like correlation coefficients, the issue was magnified in fields dealing with short, highly autocorrelated time series, as demonstrated in Fig. 12 for computational neuroscience using fMRI recordings. Many extensions to Granger causality have been proposed to explicitly model the autocorrelation function (such as ARMA [65] or state-space Granger causality [66]). However these approaches are also known to exhibit significant false-positive biases [67], aligning with our results in Sec. VIII on the ineffectiveness of prewhitening with ARMA models. In this paper, we showed that modifying the effective degrees of freedom of the null distribution suffices to eliminate the bias across all examined time series, without the additional burden of inferring complex models or prewhitening. More advanced methods (such as state-space Granger causality) may retain other empirical advantages, but their hypothesis tests are likely to require incorporation of similar modifications based on effective sample sizes. More broadly, our results strongly suggest that any hypothesis tests dealing with time-series analysis should be modified to account for autocorrelation, regardless of regressing or conditioning on AR components. The concerningly high FPRs exhibited in our experiments suggest that relationships established using previously tested Granger causality estimates should perhaps be revisited; particularly in fields that have high levels of autocorrelation and limited data.

Throughout this work, we have made the assumption that the time-series innovations are Gaussian and that all relationships are linear. Thus, we have only discussed linear-Gaussian probability distributions for information-theoretic measures. When instead applied to nonlinear time series, these probability distributions are often inferred using nonparametric density estimation techniques such as nearest neighbor or kernel methods [68]. Spurious estimates of the nearest-neighbor counts have previously been observed for autocorrelated signals by Theiler [69], who provided a solution by excluding observations that are close in time. This is now a popular approach to effectively account for autocorrelation in density estimation for nonlinear time-series analysis. In fact, in introducing transfer entropy—now understood as a model-free extension of Granger causality [35]—Schreiber explicitly recommended the use of a Theiler window (also known as serial- or dynamic-correlation exclusion) when kernel estimation methods are used [36]. The Theiler window approach has been demonstrated to control the FPR for such estimators in practice [70], yet remains a heuristic with no theoretical guarantees and, similar to Pearson correlation, is often neglected in practical estimation of transfer entropy [71]. We hypothesize that the methods outlined in our work could be extended in future to provide a more rigorous approach to handling autocorrelated nonlinear time series through, e.g., nonlinear versions of Bartlett's formula [72], facilitating a broader class of information-theoretic measures.

Finally, the dependence structure discussed in this work is assumed to be in the time domain, whereas many empirical studies are concerned with other forms of dependence that could similarly bias hypothesis tests. Future work will be required to consider handling such correlation structures in a similar fashion to that which we have presented, e.g., spectral models [21] or spatial autocorrelation [39]. Indeed, the formula for the effective sample size [Eq. (9)] was developed for spatial autocorrelation and can thus be easily extended to handle spatiotemporal autocorrelation, allowing for even broader class of null distributions that can be considered with the modified Λ-test.

ACKNOWLEDGMENTS

O.M.C. and B.D.F. were supported by NHMRC Ideas Grant 1183280. J.T.L. was supported through the Australian Research Council DECRA grant DE160100630. J.M.S. was supported through a University of Sydney Robinson Fellowship and NHMRC Project Grant 1156536. J.M.S. and J.T.L. were supported through The University of Sydney Research Accelerator (SOAR) Fellowship program. O.M.C., J.T.L., and J.M.S. were supported through The Centre for Translational Data Science at The University of Sydney's Research Incubator Funding Scheme. O.M.C., J.T.L., J.M.S., and B.D.F. were supported through the Centre for Complex Systems at The University of Sydney's Emerging Aspirations scheme. High-performance computing facilities provided by The University of Sydney (Artemis) have contributed to the research results reported within this paper.

APPENDIX A: STATISTICAL HYPOTHESIS TESTS

The linear dependence measures discussed in this paper are positive real-valued random variables $\hat{\Lambda} \in \mathbb{R}_{>0}$ that can be expressed as the ratio of the generalized variance of two models. In general, we consider two nested models, the "restricted" model with p_0 parameters (under which the null hypothesis $\mathcal{H}_0$ is true) and the "unrestricted" model with p_1 parameters (under which the alternate hypothesis $\mathcal{H}_1$ is true). These models are referred to as nested since $p_0 < p_1$ and the restricted model parameter space is a subset of the unrestricted model space. The statistics are expressed in terms of the generalized sample variance of these models:

$$\hat{\Lambda} = \frac{|\boldsymbol{s}_0|}{|\boldsymbol{s}_1|}, \tag{A1}$$

where $\boldsymbol{s}_i$ is the the residual sum-of-squares for model i and $|\boldsymbol{s}_i|$ is the generalized sample variance. The generalized sample variance is, asymptotically, inversely proportional to the likelihood of each model, and so taking the log of the ratio of generalized sample variances (A1) is equivalent to the LR between two models (for a large enough number of samples) [8,65]. For this reason, statistics of the form in Eq. (A1) appear in a number of linear dependence measures, such as mutual information (with Gaussian marginals) and Geweke's definition of Granger causality [46].

1. Asymptotic likelihood-ratio test

Wilks' theorem [55] is the basis of the χ^2-test, which states that a test statistic constructed from the LR of nested models will asymptotically follow a χ^2-distribution under the null hypothesis. Since all statistics used in this work fit this definition, a χ^2-test can be used. That is, if the true model is the restricted model, then as $T \to \infty$, the statistic is chi-square distributed:

$$T \, \log(\hat{\Lambda}) \xrightarrow{d} \chi^2(p_1 - p_0). \tag{A2}$$

However, as we show throughout the main text, the χ^2-test has a significant bias when applied to limited and autocorrelated time-series data, which results in a large number of false positives.

It is important to note that the LR test is but one of three classical procedures for hypothesis testing maximum likelihood estimates; the others are the Wald test and the Lagrange multiplier test [7,8,47]. The three tests overlap because the null distribution of each asymptotically follows the χ^2-distribution. Thus, the same issues hold if one were to use any test on linear dependence measures unless autocorrelation is considered in the null distributions.

2. Finite-sample F-test

In regression analysis, the F-test is used to infer the significance of nested models of independent observations with limited data, i.e., the finite-sample null distribution. Using the same notation as above, we obtain a distribution for the comparing the nested models:

$$\frac{T - p_1}{p_1 - p_0} \frac{S_0 - S_1}{S_1} \sim F(p_1 - p_0, T - p_1). \tag{A3}$$

F-statistics can be reformulated as nested ratios of sample variances through simply rearranging the LHS of (A3):

$$\frac{T - p_1}{p_1 - p_0} [\hat{\Lambda} - 1] \sim F(p_1 - p_0, T - p_1). \tag{A4}$$

Thus, the F-statistic is a function of the LR of two models and we can show its asymptotic distribution is chi-square. First, a Taylor expansion of the LHS of the F-statistic in Eq. (A4) gives $\log(\hat{\Lambda}) \approx \hat{\Lambda} - 1$. Moreover, for a random variable $X \sim F(\nu_1, \nu_2)$, then $Y = \lim_{\nu_2 \to \infty} \nu_1 X$ has the chi-square distribution $\chi^2(\nu_1)$. Thus, by this asymptotic relationship between the F- and χ^2-distributions, we have

$$\lim_{T \to \infty} (T - p_1) \log(\hat{\Lambda}) \xrightarrow{d} (p_1 - p_0) F(p_1 - p_0, T), \tag{A5}$$

$$T \, \log(\hat{\Lambda}) \sim \chi^2(p_1 - p_0). \tag{A6}$$

This result is discussed throughout the paper to explain the diverging behavior between the two tests.

3. Surrogate-distribution tests

Another established approach to empirically generating a null distribution involves permuting, redrawing or rotating the observations of one variable x or y and computing the relevant statistic for each surrogate dataset [4,50,68]. Naive approaches to permuting or redrawing will completely destroy the autocorrelation profile of that variable, making this empirical distribution representative of serially independent observations and similar to the analytic F-distribution.

Indeed, such empirical generation of the CDF via permutation testing was attempted (for Granger causality) by Barnett and Seth [4], and shown to incur the same inflated FPR issues as the F- and χ^2-tests. Alternatively, constrained realization approaches [73] can be used to generate surrogate time-series data that exhibit certain properties (such as the same power spectra) and have recently been shown effective for handling autocorrelation in EEG data [74]. Nonetheless, bootstrap tests are computationally inefficient and known to exhibit size and power distortions, however typically less so than asymptotic tests [75]. As such, we consider a comparison to these empirical approaches outside the scope of our paper.

4. Drawing inferences

Given an estimate $\hat{\Lambda}$ and null distribution (e.g., the F- or χ^2-distributions), we use the same general hypothesis testing procedure for all measures. Here we use the χ^2-test for $\hat{\Lambda}$ as an example, however the same applies to all linear dependence measures, including the Λ^*-distributions, which are numerically generated.

First, we set an arbitrary significance level α, which is (ideally) the probability of rejecting the null hypothesis even if it were true—this is set to 5% in this paper unless stated otherwise. Then the statistic $T \log(\hat{\Lambda})$ is input to the quantile function of the χ^2-distribution with $p_1 - p_0$ degrees of freedom. The output (the complement of the p-value) is the probability of measuring that value (or higher) under the null hypothesis $\mathcal{H}_0{:}\Lambda = 0$. If the p-value is below the significance level, then we deduce that the measured LR $\hat{\Lambda}$ is significant. A false positive occurs when the p-value is below the significance level α but the null hypothesis $\mathcal{H}_0$ is true, i.e., there is no actual dependence between the variables $\Lambda = 0$, yet $\hat{\Lambda}$ is considered significant. Ideally, we expect the proportion of false positives (the FPR) to match the significance level α, i.e., one would expect an FPR of 0.05 for a 5% significance level α. This same procedure is used for all linear dependence measures in this paper, with differing statistics and null distributions.

When the FPR is measured over R trials (usually $R = 1\,000$ in this paper), confidence intervals can be determined based on the binomial distribution of R draws of a random variable each with $\alpha = 0.05$ chance of success.

APPENDIX B: EXPERIMENTAL SETUP

1. Numerical simulations

For our experimental validation, we use an AR model similar to the example proposed in [4], with two processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ that have no interdependence, digitally filtered to increase their autocorrelation. We begin with the m-variate time-series data $\boldsymbol{z} = (\boldsymbol{z}(1), \dots, \boldsymbol{z}(T))$, i.e., an $m \times T$ matrix, where each realization is generated from a first-order vector AR model:

$$\boldsymbol{z}(t) = \boldsymbol{\Phi}(1)\boldsymbol{z}(t-1) + \boldsymbol{a}(t), \tag{B1}$$

with

$$\boldsymbol{z}(t) = \begin{bmatrix} \boldsymbol{x}(t) \\ \boldsymbol{y}(t) \end{bmatrix}, \quad \boldsymbol{\Phi}(1) = \begin{bmatrix} \boldsymbol{\Phi}_X(1) & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Phi}_Y(1) \end{bmatrix}, \tag{B2}$$

and AR parameters $\boldsymbol{\Phi}_X(1) = 0.3\boldsymbol{I}_k$ and $\boldsymbol{\Phi}_Y(1) = -0.8\boldsymbol{I}_l$. The innovations $\boldsymbol{a}(t) = (a_1(t), \dots, a_m(t))'$ are uncorrelated with mean $\boldsymbol{0}$ and unit variance matrix $\mathrm{var}(\boldsymbol{a}(t)) = \boldsymbol{I}_m$. The matrix $\boldsymbol{z}$ is then partitioned into $k \times T$ and $l \times T$ matrices denoted $\boldsymbol{x}(t)$ and $\boldsymbol{y}(t)$. For each measure, we are interested in either the mutual information between $\boldsymbol{x}$ and $\boldsymbol{y}$, or the Granger causality from $\boldsymbol{y}$ to $\boldsymbol{x}$. If a third (conditional) process $\boldsymbol{w}$ is required to contextualize these measures (in their conditional forms), we consider another first-order AR process $\boldsymbol{w}$ again using Eq. (B1), with $\boldsymbol{w}(t) \in \mathbb{R}^c$, $\boldsymbol{\Phi}_{\boldsymbol{W}}(1) = 0.4\boldsymbol{I}_c$ and unit variance innovation process. Each process, $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{w}$, are then independently filtered with either an FIR or an IIR filter. Both filters were low-pass, with their cutoff set to a normalized frequency of $\pi/2$ radians and a variable filter order.

In our study of the mutual information between bivariate processes, we inject a causal influence from the univariate process Y to X in order to test the TPR (i.e., statistical power of the test). In this scenario, we generate bivariate time-series data $\boldsymbol{z}$ from same state equations [Eq. (B1)] but with a small causal influence from Y to X in the AR parameters:

$$\boldsymbol{\Phi}(1) = \begin{bmatrix} \Phi_X(1) & \Phi_{XY}(1) \\ 0 & \Phi_Y(1) \end{bmatrix}, \tag{B3}$$

with $\Phi_{XY}(1) = 0.03$, whilst $\Phi_X(1) = 0.3$ and $\Phi_Y(1) = -0.8$, as per Eq. (B2).

2. Human Connectome Project

The HCP rsfMRI dataset [44] comprises 500 subjects imaged at a 0.72 s sampling rate for 15 min in the (relatively quiescent) resting state. This results in 1200 observations of spatially dense time-series data, which are then parcellated into 333 regions of interest in the brain. Thus, the dataset contains 500 subjects with 333 brain regions each, and each of these regions is associated with a stationary time series of 1200 observations. The raw (BOLD) data of each region was then preprocessed by removing the DC component, detrending, applying a third-order zero-phase (or forward and reverse) Butterworth bandpass filter (0.01–0.08 Hz). These are common techniques used to remove potential artefacts. We also removed 200 observations from the start and the end of the time series, in order to minimize filter initialization effects. This leaves $T = 800$ observations for the analysis. In order to build a scenario where the null hypothesis holds, we conduct experiments on 1000 time-series pairs, selecting different random regions of interest from different random subjects, making the corresponding time series completely independent of one another. The analysis was performed for mutual information (between both univariate and bivariate time series) and Granger causality in the same way as discussed for the simulated time-series experiments above. We use the same hypothesis testing procedure (discussed above) with a 5% significance level for the χ^2-test and our newly proposed modified Λ-test.

APPENDIX C: ADDITIONAL PREWHITENING TESTS

This Appendix provides further evidence that standard prewhitening techniques are insufficient for many covariance-stationary time series. In the main text, we show that prewhitening time series is ineffective when the time-series models are either AR(p) models that are inferred from Burg's method or ARMA(p, q) models that are inferred from the BIC score (up to a maximum order of $p = q = 5$). Here we extend these results to show that AR(p) models inferred via the AICc (AIC with small-sample correction) and BIC scoring functions are also insufficient.

In Fig. 13 we show the extended prewhitening results for mutual information tests. The algorithm iterates through all

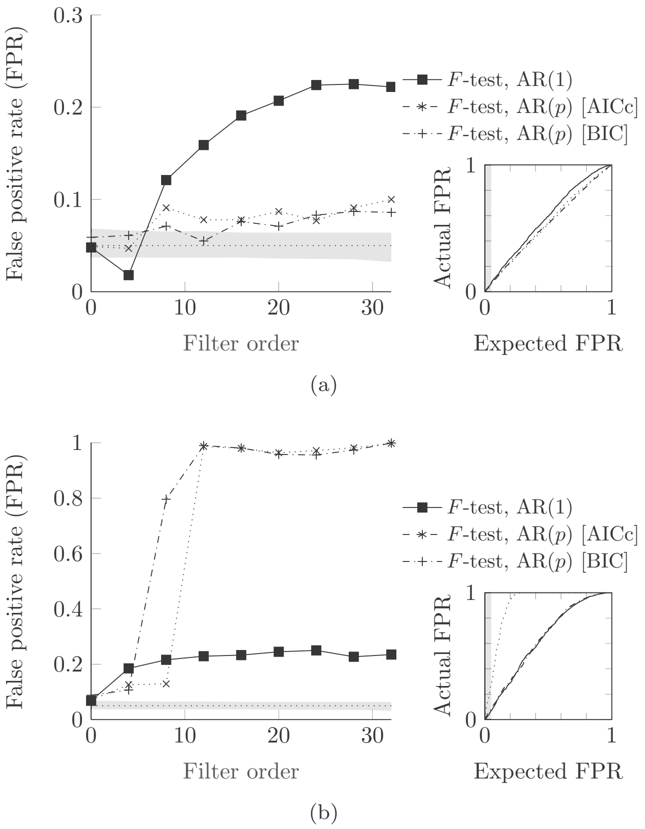


FIG. 13. Inferring the AR models via AICc and BIC does not consistently improve prewhitening results for mutual information tests with FIR (a) or IIR (a) filtering. The experiments are as per Fig. 1, with two additional prewhitening approaches used: the AR(p) model inferred from the AICc and BIC scores, respectively.

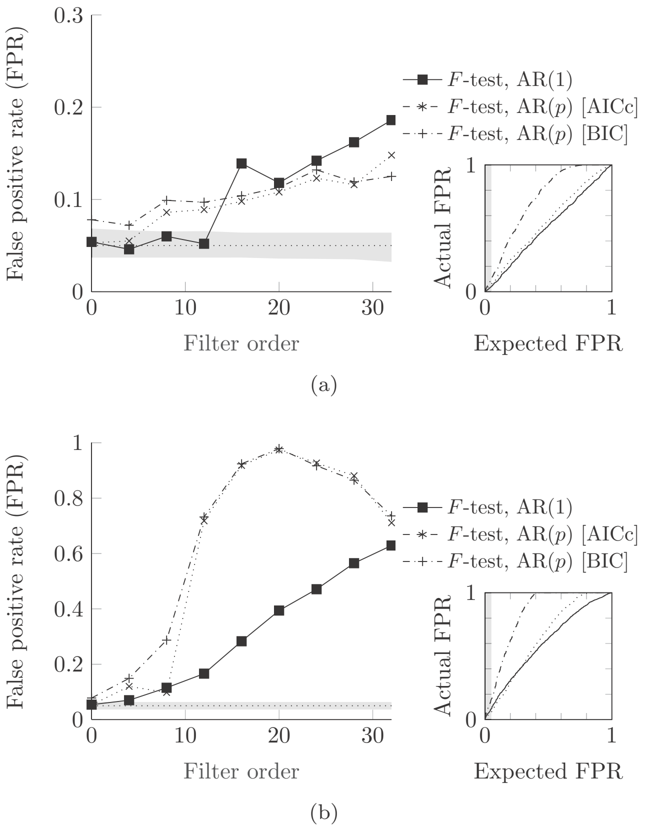


FIG. 14. Inferring the AR models via AICc and BIC does not improve prewhitening results for Granger causality tests with FIR (a) or IIR (a) filtering. The experiments are as per Fig. 6, with two additional prewhitening approaches used: the AR(p) model inferred from the AICc and BIC scores, respectively.

potential AR(p) models and selects the one that minimises the AICc and BIC scores (independently). For the FIR-filtered data, this approach shows a slightly increased FPR above the nominal value of 5%—this illustrates a marginal improvement over Burg's method from Fig. 10. However, for the IIR-filtered data, the FPR approaches 100% and is significantly worse than even methods with no or minimal prewhitening [cf. the AR(1) models or the standard F-test in Fig. 1]. Similarly, prewhitening is shown to be insufficient for Granger causality tests in Fig. 14, where equivalent experiments were performed with no major qualitative differences (compared to Fig. 11).

We do not show any further results for ARMA(p, q) models, e.g., by increasing the maximum order or via a different criterion because the procedure to learn the parameters of higher-order ARMA models was too computationally expensive using the standard functions. Given this constraint, all information criteria (AIC, AICc, or BIC) often chose the maximum order in practice, and so using alternative approaches was redundant. We conjecture that prewhitening with ARMA(p, q) models may perform favorably to AR(p) models given the ability to infer arbitrarily complex models. However, the constraints governed by the their inference procedures makes testing this currently intractable in practice.

APPENDIX D: CONSIDERATIONS AND EXTENSIONS OF BARTLETT'S FORMULA

In our derivations we use a first-order approximation of Bartlett's formula [Eq. (8)], that was originally described for spatially autocorrelated processes. However, since Bartlett's seminal work [10], there have been a number of other extensions made to his formula as well as techniques intended to overcome the issues of its empirical computation.

One of the more general cases of Bartlett's formula is due to Roy [40], who provided the large-sample distribution between pairs of sample cross-correlations at differing lags. Consider the four processes Z_i, Z_j, Z_k, Z_l. Let

$$\Delta_v(i, j, k, l) = \sum_{u=-\infty}^{\infty} \rho_{ij}(u)\rho_{kl}(u+v), \tag{D1}$$

where $\rho_{ij}(u)$ are the cross-correlation as per Eq. (5), and

$$s_{ab}(v) = T^{1/2}\,[r_{ab}(v) - \rho_{ab}(v)], \tag{D2}$$

as the standard error, with $r_{ab}(v)$ the sample cross-correlation in Eq. (6). In general, the asymptotic distribution of the standard error, $s_{ab}(v)$, is Gaussian with zero mean and covariance

$$\lim_{T\to\infty} \mathrm{cov}(s_{ab}(v), s_{de}(w))$$
$$\approx \Delta_{w-v}(a, d, b, e) + \Delta_{w+v}(b, d, a, e)$$

$$\begin{aligned}&-\rho_{ab}(v)[\Delta_w(a,d,a,e)+\Delta_w(b,d,b,e)]\\&-\rho_{de}(w)[\Delta_v(b,d,a,d)+\Delta_v(b,e,a,e)]\\&+\tfrac{1}{2}\rho_{ab}(w)\rho_{de}(w)[\Delta_0(a,d,a,d)+\Delta_0(a,e,a,e)\\&+\Delta_0(b,d,b,d)+\Delta_0(b,e,b,e)].\end{aligned}\tag{D3}$$

This derivation can be further generalized to the non-Gaussian case, for instance, by allowing for skewed distributions [76]. More recently, a first-order approximation of this formula was given by Afyouni *et al.* [2], which takes the same form as Eq. (D3), with Eq. (D1) slightly modified.

One consequence of knowing the full covariance structure [Eq. (D3)] is that such a distribution could further help in situations when the partial correlation terms that Wilks' statistic decomposes into are themselves correlated. That is, in the main text, we provided the variance for each Λ^*-distributed variable L_i (as a beta distribution by assuming independence). By assuming independence, we were able to obtain the sampling distribution of Wilks' criterion as a product of these Λ^*-distributions. However, if the random variables were correlated, then we must use the multivariate form of Bartlett's formula (D3), which provides the covariance of each variable L_i term with all other variables L_j (i.e., we would have n_{ij} for each i and j, rather than just n_i). Accounting for this covariance would require knowing the distribution of the general product of correlated beta-distributions that, to the best of our knowledge, is not an established result.

Another use case of Eq. (D3) is testing against an alternative hypothesis [$\mathcal{H}_1 : \rho_{ab}(0) \neq 0$], which is discussed at length by Afyouni *et al.* [2] for correlation coefficients. One could follow the same logic from this paper to provide the alternative hypothesis test for linear dependence measures based on Wilks' statistic.

There are a number of special cases of Roy's formula that are worth noting. In the event that we are interested in the covariance $\mathrm{cov}(s_{ab}(v), s_{ab}(w))$ between cross-correlation estimates of two univariate processes Z_a and Z_b at arbitrary lags v and w, this is obtained from Eq. (D3) by setting $d = a$ and $e = b$, reducing to the results reported in Refs. [77] and [7] (Theorem 11.2.3). Using this special case, the null distribution of Pearson (zero-lag) correlation between two univariate processes $\mathrm{var}(s_{ab}(0))$ can be obtained by setting $v = w = 0$. Finally, under the assumption that $\rho_{ab}(0) = 0$, most of these terms disappear and we are left with Bartlett's original formula [10]:

$$\begin{aligned}\lim_{T\to\infty}\mathrm{var}(s_{ab}(0)) &\approx \lim_{T\to\infty}\mathrm{var}(T^{1/2}[r_{ab}(0)])\\&=\sum_{u=-\infty}^{\infty}\rho_{ii}(u)\rho_{jj}(u)\end{aligned}\tag{D4}$$

with a similar form given by a first-order approximation [38]:

$$\mathrm{var}(s_{ab}(0)) \approx T^{-1}\sum_{u=-\infty}^{\infty}(T-|u|)\rho_{aa}(u)\rho_{bb}(u).\tag{D5}$$

Due to symmetry of the autocorrelation function about lag-zero for stationary processes, we can simply sum over the positive lags in Eq. (D5), $u > 0$, which was the form used throughout this paper [see Eq. (8)]. For the exact relationship between the large-sample approximations and the first-order approximations, we refer the reader to the discussions in Refs. [11,38,78]. Bartlett did indeed present a formula irrespective of sample size [11], which may yield an improvement for small sample distributions and give minor practical advantages, however, we did not find this necessary for any experiments and instead follow the approximations in Eq. (D5).

Another potential source of error in the sampling distributions come from incorrectly estimating the autocorrelation function $r_{aa}(u)$. Tapering (also known as data windowing) is commonly used in practice to regularize the autocorrelation samples to better estimate their true value [2,51]. These approaches involve scaling the autocorrelation samples by some factor, with the maximum lag truncated below the dataset length. Using this method, we can appropriate Bartlett's formula to

$$\mathrm{var}(s_{ab}(0)) \approx 1 + 2\sum_{u=1}^{U}\frac{T-u}{T}\lambda(u)r_{aa}(u)r_{bb}(u),\tag{D6}$$

where $\lambda(u)$ are a set of weights called the lag window and $U < T$ is the truncation point. The lag window comprises $\lambda(u)$ values that decrease with increasing u; two common approaches are the Parzen and the Tukey windows (see Ref. [51] for details). Numerous truncation points U have also been proposed, e.g., $T/4$, $T/5$, $\sqrt{T}$, and $2\sqrt{T}$ [2].

In the above few sections we outlined a number of potential factors that could introduce small size or power distortions in our hypothesis tests. To compare our approach with these more complex extensions, we ran experiments with the effective sample size computed from the full covariance matrix (D3), both with and without tapering. These were computed for the experiments from Fig. 1, however, as mentioned above, the product of correlated beta- or F-distributed variates is unknown and so our modified Λ-test could not be performed. Instead, we used the sums of z-transformed partial correlations (each of which make up the conditional mutual information term), rather than the products of squared partial correlations. That is, by transforming each partial correlation, we expect the sum of these correlation to be approximately Gaussian. In performing these tests, we found no notable difference in the FPRs for any of the validation experiments, suggesting that these additions made no significant difference towards reducing size or power distortions.

APPENDIX E: PARTIAL AUTOCORRELATION AND ACTIVE INFORMATION STORAGE

The partial autocorrelation function conveys important information regarding the dependence structure of an AR process [7]. For a univariate stationary time series Z, the partial autocorrelation $\alpha_Z(u)$ at lag u is the correlation between $Z(t)$ and $Z(t-u)$, adjusted for the intervening observations $\mathbf{Z}^{(u-1)}(t) = [Z(t-1)\,;\,\dots\,;\,Z(t-u+1)]$. Denote Z^u as the process of Z lagged by u time steps and $Z^{(u)}$ as the history up until that lag (inclusive) $\mathbf{Z}^{(u)} = [Z^1\,;\,\dots\,;\,Z^u]$. Then, for a stationary time series, the partial autocorrelation function is defined by [7]

$$\alpha_Z(1) = \rho_{ZZ^1}\tag{E1}$$

and

$$\alpha_Z(u) = \rho_{ZZ^u \cdot \mathbf{Z}^{(u-1)}}, \quad u > 1. \tag{E2}$$

Although we use Burg's method to identify the relevant history length p for an AR model of Z length for AR models in this paper, it is common practice to use the partial autocorrelation function instead, since $\alpha_Z(u) = 0$ for $u > p$ [7,8,47]. Again, this is a statistical estimate and thus the order p is inferred by testing each sample partial autocorrelation $\hat{\alpha}_Z(u)$ for significance against the null distribution.

Intriguingly, our work reveals a relationship between the partial autocorrelation function and active information storage [43]—a recently developed model-free measure for quantifying memory in a process—under the linear Gaussian assumption. The average active information storage $\mathcal{A}_X$ quantifies the information storage in a process. For a p-order Markov process X, this is quantified by the mutual information between the relevant history $\mathbf{X}^{(p)}(t)$ and variable $X(t)$:

$$\mathcal{A}_X(p) = \mathcal{I}_{X;\mathbf{X}^{(p)}}. \tag{E3}$$

Since the average active information storage is a specific type of mutual information, we can use the chain rule to decompose it into a sum of squared partial autocorrelations:

$$\begin{aligned} \mathcal{I}_{X;\mathbf{X}^{(p)}} &= -1/2 \sum_{u=1}^{p} \log\left(1 - \rho^2_{XX^u \cdot \mathbf{X}^{(u-1)}}\right) \\ &= -1/2 \sum_{u=1}^{p} \log\left\{1 - [\alpha_X(u)]^2\right\}. \end{aligned} \tag{E4}$$

This same logic can be straightforwardly applied to other measures such as excess entropy [41] and predictive information [42].

In additional to quantifying the memory within a process, active information storage is often used for inferring the optimal history length for both the Gaussian and non-Gaussian cases [68]. This is typically achieved by using the χ^2-test to infer the significance of increasing the embedding lengths p. For AR processes with Gaussian innovations, we infer the embedding length p for X by first taking the difference $\delta_x(u) = \mathcal{A}_x(u+1) - \mathcal{A}_X(u)$ and then generating a p-value by testing $2\,\delta_X(u)$ against a $\chi^2(1)$ distribution, which represents the null hypothesis of no increase in information storage. If the p-value is below a threshold (say, 5%), then the test is rejected and the lag is increased $u = u + 1$. This process is iterated until the null hypothesis is accepted, at which point we surmise that the optimal lag p is the one at which $\delta_X(p+1)$ is considered insignificant. This approach is similar to using the partial autocorrelation, as the difference $\delta_X(u)$ is equivalent to squared partial autocorrelation up to a factor of two. This can be seen from Eq. (E4):

$$\begin{aligned} \delta_X(u) &= \mathcal{A}_X(u+1) - \mathcal{A}_X(u) \\ &= -\tfrac{1}{2}\log\{1 - [\alpha_X(u+1)]^2\}. \end{aligned} \tag{E5}$$

In contrast to measures of dependence between multiple processes, the χ^2-test appears suitable here (without adjusting for an effective sample size) for testing $\delta_X(u)$ for $u > p$. This is because, after the full set of past variables is included in the regression, any higher order residuals $x^u - \hat{x}^u(\hat{x}^{(p)})$ with $u > p$ have statistically zero autocorrelation for every lag.

[1] C. E. Davey, D. B. Grayden, G. F. Egan, and L. A. Johnston, Filtering induces correlation in fMRI resting state data, NeuroImage **64**, 728 (2013).

[2] S. Afyouni, S. M. Smith, and T. E. Nichols, Effective degrees of freedom of the Pearson's correlation coefficient under autocorrelation, NeuroImage **199**, 609 (2019).

[3] E. Florin, J. Gross, J. Pfeifer, G. R. Fink, and L. Timmermann, The effect of filtering on Granger causality based multivariate causality measures, NeuroImage **50**, 577 (2010).

[4] L. Barnett and A. K. Seth, Behaviour of Granger causality under filtering: Theoretical invariance and practical application, J. Neurosci. Methods **201**, 404 (2011).

[5] A. K. Seth, A MATLAB toolbox for Granger causal connectivity analysis, J. Neurosci. Methods **186**, 262 (2010).

[6] P. M. Robinson, Generalized canonical analysis for time series, J. Multivariate Anal. **3**, 141 (1973).

[7] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods* (Springer Science & Business Media, 1991).

[8] G. C. Reinsel, *Elements of Multivariate Time Series Analysis* (Springer Science & Business Media, New York, 2003).

[9] G. U. Yule, Why do we sometimes get nonsense-correlations between time-series?—A study in sampling and the nature of time-series, J. R. Stat. Soc. **89**, 1 (1926).

[10] M. S. Bartlett, Some aspects of the time-correlation problem in regard to tests of significance, J. R. Stat. Soc. **98**, 536 (1935).

[11] M. S. Bartlett, On the theoretical specification and sampling properties of autocorrelated time-series, Suppl. J. R. Stat. Soc. **8**, 27 (1946).

[12] J. D. Cryer and K.-S. Chan, *Time Series Analysis: With Applications in R* (Springer Science & Business Media, New York, 2008).

[13] D. Sul, P. C. Phillips, and C.-Y. Choi, Prewhitening bias in HAC estimation, Oxford Bull. Econ. Stat. **67**, 517 (2005).

[14] M. Bayazit and B. Önöz, To prewhiten or not to prewhiten in trend analysis? Hydrol. Sci. J. **52**, 611 (2007).

[15] W. Olszowy, J. Aston, C. Rua, and G. B. Williams, Accurate autocorrelation modeling substantially improves fMRI reliability, Nat. Commun. **10**, 1220 (2019).

[16] S. Yue and C. Y. Wang, Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test, Water Resour. Res. **38**, 4 (2002).

[17] M. R. Arbabshirani, E. Damaraju, R. Phlypo, S. Plis, E. Allen, S. Ma, D. Mathalon, A. Preda, J. G. Vaidya, T. Adali, and V. D. Calhoun, Impact of autocorrelation on functional connectivity, NeuroImage **102**, 294 (2014).

[18] J. R. Bence, Analysis of short time series: Correcting for autocorrelation, Ecology **76**, 628 (1995).

[19] M. Macias-Fauria, A. Grinsted, S. Helama, and J. Holopainen, Persistence matters: Estimation of the statistical significance of paleoclimatic reconstruction statistics from autocorrelated time series, Dendrochronologia **30**, 179 (2012).

[20] K. J. Friston, Functional and effective connectivity: A review, Brain Connect. **1**, 13 (2011).

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

[22] A. Roebroeck, E. Formisano, and R. Goebel, Mapping directed influence over the brain using Granger causality and fMRI, NeuroImage **25**, 230 (2005).

[23] W. Liao, J. Ding, D. Marinazzo, Q. Xu, Z. Wang, C. Yuan, Z. Zhang, G. Lu, and H. Chen, Small-world directed networks in the human brain: Multivariate Granger causality analysis of resting-state fMRI, NeuroImage **54**, 2683 (2011).

[24] K. Friston, R. Moran, and A. K. Seth, Analysing connectivity with Granger causality and dynamic causal modelling, Curr. Opin. Neurobiol. **23**, 172 (2013).

[25] R. K. Kaufmann and D. I. Stern, Evidence for human influence on climate from hemispheric temperature relations, Nature (London) **388**, 39 (1997).

[26] D. D. Zhang, H. F. Lee, C. Wang, B. Li, Q. Pei, J. Zhang, and Y. An, The causality analysis of climate change and large-scale human crisis, Proc. Natl. Acad. Sci. USA **108**, 17296 (2011).

[27] J. R. Freeman, Granger causality and the times series analysis of political relationships, Am. J. Polit. Sci. **27**, 327 (1983).

[28] R. Reuveny and H. Kang, International trade, political conflict/cooperation, and Granger causality, Am. J. Polit. Sci. **40**, 943 (1996).

[29] S. S. Wilks, Certain generalizations in the analysis of variance, Biometrika **24**, 471 (1932).

[30] T. Pham-Gia, Exact distribution of the generalized Wilks's statistic and applications, J. Multivariate Anal. **99**, 1698 (2008).

[31] H. Hotelling, Relations between two sets of variates, Biometrika **28**, 321 (1936).

[32] D. J. MacKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, UK, 2003).

[33] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, 2012).

[34] J. Kay, Feature discovery under contextual supervision using mutual information, in *Proceedings of IEEE IJCNN*, (1992), Vol. 4, pp. 7984.

[35] L. Barnett, A. B. Barrett, and A. K. Seth, Granger Causality and Transfer Entropy are Equivalent for Gaussian Variables, Phys. Rev. Lett. **103**, 238701 (2009).

[36] T. Schreiber, Measuring Information Transfer, Phys. Rev. Lett. **85**, 461 (2000).

[37] T. Bossomaier, L. Barnett, M. Harré, and J. T. Lizier, *An Introduction to Transfer Entropy: Information Flow in Complex Systems* (Springer, Cham, Switzerland, 2016).

[38] G. Bayley and J. Hammersley, The "effective" number of independent observations in an autocorrelated time series, Suppl. J. R. Stat. Soc. **8**, 184 (1946).

[39] P. Clifford, S. Richardson, and D. Hémon, Assessing the significance of the correlation between two spatial processes, Biometrics **45**, 123 (1989).

[40] R. Roy, Asymptotic covariance structure of serial correlations in multivariate time series, Biometrika **76**, 824 (1989).

[41] W. Bialek, I. Nemenman, and N. Tishby, Complexity through nonextensivity, Phys. A **302**, 89 (2001).

[42] J. P. Crutchfield and D. P. Feldman, Regularities unseen, randomness observed: Levels of entropy convergence, Chaos **13**, 25 (2003).

[43] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, Local measures of information storage in complex distributed computation, Inf. Sci. **208**, 39 (2012).

[44] D. C. Van Essen, K. Ugurbil, E. Auerbach, D. Barch, T. Behrens, R. Bucholz, A. Chang, L. Chen, M. Corbetta, S. W. Curtiss *et al.*, The Human Connectome Project: A data acquisition perspective, NeuroImage **62**, 2222 (2012).

[45] https://github.com/olivercliff/exact-linear-dependence.

[46] J. Geweke, Measurement of linear dependence and feedback between multiple time series, J. Am. Stat. Assoc. **77**, 304 (1982).

[47] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control* (John Wiley & Sons, San Francisco, 2015).

[48] The most general form is the ARIMA model, where the integrative (I) component accounts for nonstationary time series. Here we assume that appropriate differencing and mean removal have already been performed in order to remove any integration of the time series.

[49] The problem of recovering the autocorrelation functions from data, and thus the effective sample size, is nontrivial and has resulted in procedures such as tapering to handle noisy estimates [2,51]; these notions are covered briefly in Appendix D.

[50] M. J. Anderson and J. Robinson, Permutation tests for linear models, Aust. NZ J. Stat. **43**, 75 (2001).

[51] C. Chatfield and H. Xing, *The Analysis of Time Series: An Introduction with R*, 7th ed. (Chapman and Hall/CRC, Boca Raton, 2019).

[52] A. M. Mathai and P. N. Rathie, The exact distribution of Wilks' criterion, Ann. Math. Stat. **42**, 1010 (1971).

[53] D. R. Brillinger, Some data analyses using mutual information, Braz. J. Probab. Stat. **18**, 163 (2004).

[54] L. Barnett and T. Bossomaier, Transfer Entropy as a Log-Likelihood Ratio, Phys. Rev. Lett. **109**, 138105 (2012).

[55] S. S. Wilks, The large-sample distribution of the likelihood ratio for testing composite hypotheses, Ann. Math. Stat. **9**, 60 (1938).

[56] C. E. Davey, D. B. Grayden, M. Gavrilescu, G. F. Egan, and L. A. Johnston, The equivalence of linear Gaussian connectivity techniques, Human Brain Mapping **34**, 1999 (2013).

[57] J. T. Lizier and M. Prokopenko, Differentiating information transfer and causal effect, Eur. Phys. J. B **73**, 605 (2010).

[58] M. De Hoon, T. Van der Hagen, H. Schoonewelle, and H. Van Dam, Why Yule-Walker should not be used for autoregressive modelling, Ann. Nucl. Energy **23**, 1219 (1996).

[59] J. Garland, R. G. James, and E. Bradley, Leveraging information storage to select forecast-optimal parameters for delay-coordinate reconstructions, Phys. Rev. E **93**, 022221 (2016).

[60] In general, the variables representing the past need not be a sequence of consecutive temporal indices, nor have the same history length p for each dimension of $\boldsymbol{X}$. For instance, one could use a Takens embedding [79,80] or any other statistically significant set of variables [81]. However, for simplicity, in this work we follow the vector AR model often used in Granger causality analysis [Eq. (B1)] and thus use a consecutive sequence of variables as the history of $\boldsymbol{X}(t)$.

[61] J. Wishart, The generalised product moment distribution in samples from a normal multivariate population, Biometrika **20A**, 32 (1928).

[62] Using the MATLAB `estimate` function in order to infer the parameters for each p and q; see our open-source toolkit [45] for details.

[63] R. Franciotti, N. Falasca, L. Bonanni, F. Anzellotti, V. Maruotti, S. Comani, A. Thomas, A. Tartaro, J. Taylor, and M. Onofrj, Default network is not hypoactive in dementia with fluctuating cognition: An Alzheimer disease-dementia with Lewy bodies comparison., Neurobiol. Aging **34**, 1148 (2013).

[64] L. Barnett and A. K. Seth, The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference, J. Neurosci. Methods **223**, 50 (2014).

[65] H. Boudjellaba, J.-M. Dufour, and R. Roy, Testing causality between two vectors in multivariate autoregressive moving average models, J. Am. Stat. Assoc. **87**, 1082 (1992).

[66] L. Barnett and A. K. Seth, Granger causality for state-space models, Phys. Rev. E **91**, 040101(R) (2015).

[67] A. Gutknecht and L. Barnett, Sampling distribution for single-regression Granger causality estimators, arXiv:1911.09625 (2019).

[68] J. T. Lizier, JIDT: An information-theoretic toolkit for studying the dynamics of complex systems, Front. Robotics AI **1**, 11 (2014).

[69] J. Theiler, Spurious dimension from correlation algorithms applied to limited time-series data, Phys. Rev. A **34**, 2427 (1986).

[70] See the Supporting Information of Novelli *et al.* [81] for a similar experiment with transfer entropy on the same HCP rsfMRI dataset used in our Fig. 12.

[71] Weber *et al.* [82] report a contrasting finding of increased FPR in transfer entropy inference due to filtering; however, the use of a Theiler window is not specified.

[72] C. Francq and J.-M. Zakoïan, Bartlett's formula for a general class of nonlinear processes, J. Time Ser. Anal. **30**, 449 (2009).

[73] T. Schreiber and A. Schmitz, Improved Surrogate Data for Nonlinearity Tests, Phys. Rev. Lett. **77**, 635 (1996).

[74] N. Schaworonkow, D. A. Blythe, J. Kegeles, G. Curio, and V. V. Nikulin, Power-law dynamics in neuronal and behavioral data introduce spurious correlations, Human Brain Mapping **36**, 2901 (2015).

[75] R. Davidson and J. G. MacKinnon, The size distortion of bootstrap tests, Econ. Theory **15**, 361 (1999).

[76] N. Su and R. Lund, Multivariate versions of Bartlett's formula, J. Multivariate Anal. **105**, 18 (2012).

[77] M. S. Bartlett, *An Introduction to Stochastic Processes: With Special Reference to Methods and Applications*, 3rd ed. (Cambridge University Press, Cambridge, 1978).

[78] M. H. Quenouille, Notes on the calculation of autocorrelations of linear autoregressive schemes, Biometrika **34**, 365 (1947).

[79] F. Takens, Detecting strange attractors in turbulence, in *Dynamical Systems and Turbulence*, Lecture Notes in Mathematics edited by D. Rand and L.-S. Young (Springer, Berlin, 1981), Vol. 898, pp. 366–381.

[80] O. M. Cliff, M. Prokopenko, and R. Fitch, Minimising the Kullback–Leibler divergence for model selection in distributed nonlinear systems, Entropy **20**, 51 (2018).

[81] L. Novelli, P. Wollstadt, P. Mediano, M. Wibral, and J. T. Lizier, Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing, Netw. Neurosci. **3**, 827 (2019).

[82] I. Weber, E. Florin, M. Von Papen, and L. Timmermann, The influence of filtering and downsampling on the estimation of transfer entropy, PLoS ONE **12**, e0188210 (2017).