Exact Inference of Linear Dependence Between Multiple Autocorrelated Time Series

Oliver M. Cliff,[1,2,*] Leonardo Novelli,[1,2] Ben D. Fulcher,[1,3] James M. Shine,[1,4] and Joseph T. Lizier[1,2]

[1] *Centre for Complex Systems, The University of Sydney, Sydney NSW 2006, Australia*
[2] *School of Civil Engineering, The University of Sydney, Sydney NSW 2006, Australia*
[3] *School of Physics, The University of Sydney, Sydney NSW 2006, Australia*
[4] *Brain and Mind Centre and Central Clinical School, The University of Sydney, Sydney NSW 2006, Australia*

The ability to quantify complex relationships within multivariate time series is a key component of modelling many physical systems, from the climate to brains and other biophysical phenomena. Unfortunately, even testing the significance of simple dependence measures, such as Pearson correlation, is complicated by altered sampling properties when autocorrelation is present in the individual time series. Moreover, it has been recently established that commonly used multivariate dependence measures—such as Granger causality—can produce substantially inaccurate results when applying classical hypothesis-testing procedures to digitally-filtered time series. Here, we suggest that the digital filtering-induced bias in Granger causality is an effect of autocorrelation, and we present a principled statistical framework for the hypothesis testing of a large family of linear-dependence measures between multiple autocorrelated time series. Our approach unifies the theoretical foundations established by Bartlett and others on variance estimators for autocorrelated signals with the more intricate multivariate measures of linear dependence. Specifically, we derive the sampling distributions and subsequent hypothesis tests for any measure that can be decomposed into terms that involve independent partial correlations, which we show includes Granger causality and mutual information under a multivariate linear-Gaussian model. In doing so, we provide the first exact tests for inferring linear dependence between vector autoregressive processes with limited data. Using numerical simulations and brain-imaging datasets, we demonstrate that our newly developed tests maintain the expected false-positive rate (FPR) with minimally-sufficient samples, while the classical log-likelihood ratio tests can yield an unbounded FPR depending on the parameters chosen. Due to the prevalence of autocorrelation in empirical datasets, many time-series dependencies in the scientific literature may have been, and may continue to be, spuriously reported or missed if our exact testing procedure is not widely adopted.

I. INTRODUCTION

Dependence measures such as Pearson correlation, mutual information, 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 validity of classical hypothesis tests typically used to infer the statistical significance of empirical calculations of such measures from data [1–5]. Specifically, the presence of autocorrelation (or serial correlation) in a signal—one of two defining properties of a stationary time series [6, 7]—has been found to bias the estimation of these linear-dependence measures. This bias yields a greater number of both spurious correlations and missed causalities (Type I and Type II errors) during hypothesis testing and calls into question the legitimacy of the many studies across the empirical sciences that have employed these techniques.

The effect of autocorrelation on estimating Pearson's product-moment correlation coefficient has been extensively studied in statistics. Seminal work on this subject is often attributed to Bartlett [8], who provided a variance estimator for sample correlation coefficients (among other statistics) between two identical first-order autoregressive (AR) processes under the null hypothesis of no correlation [8]. He later extended his analysis to account for arbitrary-order AR processes [9]; however, both time series were still assumed to be generated by the same model. Quenouille [10] and Bayley and Hammersley [11] independently took this work further by studying sample estimates when the two time series can exhibit different autocorrelation functions. Roy then provided a more comprehensive result on the asymptotic covariance for multivariate processes when there exists an underlying correlation between the processes [12], which provides the distribution under the alternate hypothesis. Variance estimators such as these, that explicitly account for autocorrelation, are commonly referred to as Bartlett's formula [2, 6, 7, 12].

Although the sampling properties established by Bartlett and others rigorously accounts for autocorrelation, until recently they have remained underused in practice, with the impact of the resulting false-positive bias clearest in applications involving short time series and high autocorrelation, e.g., in fMRI-based neuroimaging [1, 2], environmental and ecological studies [14, 15]. For instance, experiments on brain-imaging data have revealed that the related concept of digitally filtering a signal can induce spurious correlations [1]. The authors provided hypothesis tests based on the filter frequency response; however, since digital filtering is usu-

* oliver.cliff@sydney.edu.au

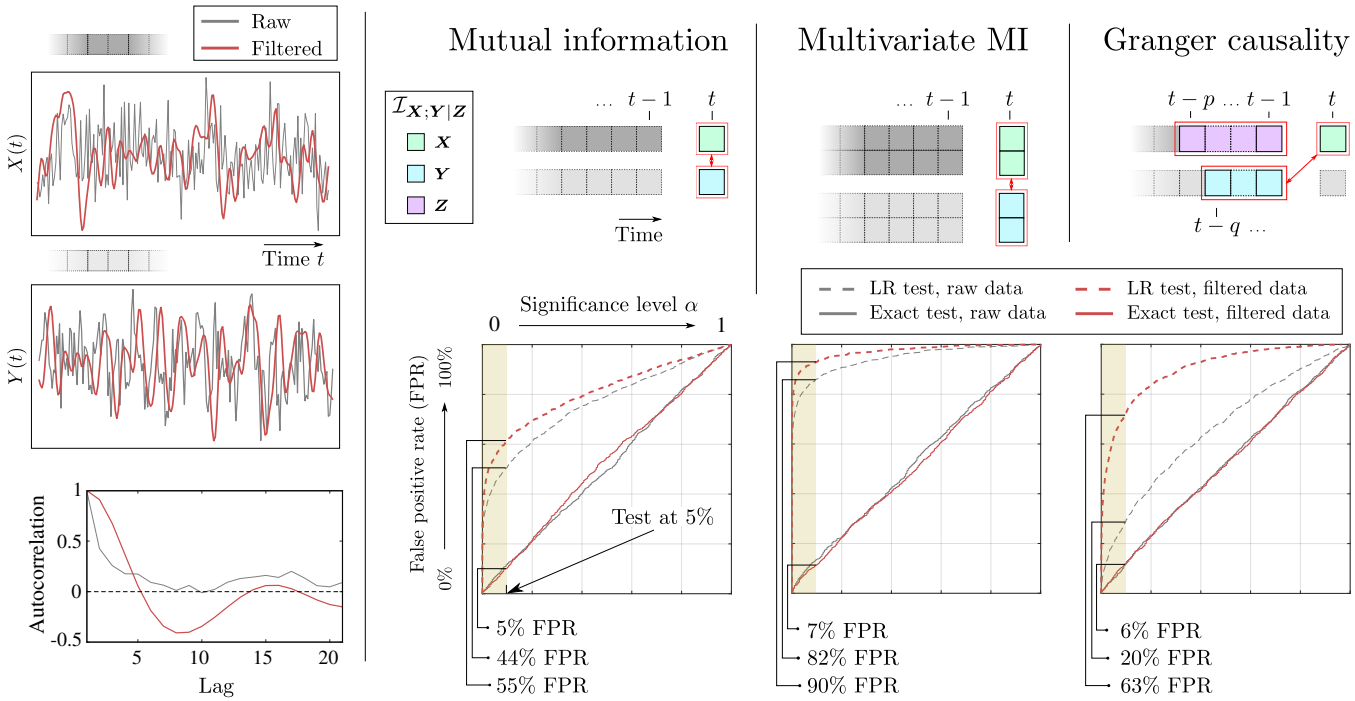


FIG. 1: Application to brain-imaging data, demonstrating our correction for the otherwise dramatic inflation of false-positive rates (FPRs) of classical hypothesis tests of dependence estimates in the presence of autocorrelation. We perform 1000 experiments with both the (uncorrected) log-likelihood ratio (LR) hypothesis tests as well as our (corrected) exact tests introduced in Sec. IV for inferring the significance of three dependence measures: mutual information, multivariate mutual information (with two variables in each process), 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 [13] (see Sec. V for more details). A sample window of these univariate processes are shown in the left panel (grey 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 $r_{ii}(u)$ of both the raw and the filtered signal $X(t)$, illustrating the higher autocorrelation length (and lower effective sample size) induced by digital filtering. The panels on the right show the FPR 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 LR 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.

ally not the only cause of autocorrelation, this signal characteristic is typically unknown in practice [2]. Following this result, Afyouni et al. [2] rediscovered Roy's asymptotic distribution, showing that it also holds as an approximation with finite samples. This distribution was then validated through minimal examples and fMRI brain-imaging time series, overcoming the issue of an unknown filter frequency response. The application to fMRI data was particularly important, since these typically only contain several hundred samples, are highly autocorrelated both before and after commonly-applied digital filtering, and have been widely studied with correlation coefficients [16]. These represent important milestones in inferring the relationship between univariate time series, however correlation coefficients are but one of many linear-dependence measures, with others in the family capturing both directed and multivariate dependence. Even though such measures can model a much richer set of dependence structures in physical processes, the notion that their sampling properties could be altered in the presence of autocorrelation has been so far overlooked; bridging this gap is the focus of this work.

Motivated to study directed dependencies in economics, Granger [17] introduced his measure of causal influence between autoregressive models nearly 60 years ago. Since then, it has become an exceedingly popular measure of linear dependence, exemplified by more than 99,000 works that mention the phrase "Granger causality", as indexed by Google Scholar (as of March, 2020). This impact is reflected in the measure's ubiquity in the scientific community beyond its origins in econometrics, yielding insights into phenomena ranging from brain dynamics [18–20] to climate change [21, 22] and political relationships [23, 24]. Although Granger causality was explicitly designed to capture one-way dependence between AR signals, researchers are becoming increasingly aware that preprocessing techniques that increase this autoregression, such as filtering, raises its FPR when us-

ing classical hypothesis tests such as the log-likelihood ratio (LR) and F-tests [3–5]. Following this line of inquiry, in Fig. 1 we use a widely-accessed brain-imaging dataset to illustrate the extent to which natural and artificial autocorrelation can inflate the FPRs over a nominated 5% rate (see Sec. V for more detail on this experiment). For unrelated signal pairs, the FPR of Granger causality using the LR test was computed to be 20%, increasing to 63% after bandpass filtering—more than twelve times the expected value. Even though previous studies have provided similar findings and clearly demonstrated increased Type I and II errors due to filtering, it remained unclear as to why this bias existed, nor how to correct it. We hypothesized that these errors were due to an inflated variance of the sampling distribution as a function of autocorrelation, in the same way that Pearson 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. A natural multivariate extension to Pearson correlation is canonical correlation analysis, which Hotelling first introduced as a way of relating more than two Gaussian variables [25]. Robinson [26] later established that, under serial correlation, estimates of canonical correlations are consistent and obey a central limit theorem. However, he also showed that, like Pearson correlation, canonical correlation estimates are inefficient under autocorrelation, which introduces Type I and II errors during inference. As an alternative, information theory naturally facilitates multivariate analysis [27, 28]. (Also, though not in scope for this study, information-theoretic measures can handle nonlinearity using model-free estimators, for which the issue of serial correlation is addressed using methods such as the Theiler window correction [29]; discussed in Sec. VI). Information theory's general applicability arises in simply requiring a probability distribution that can be either parametric or non-parametric. When this probability distribution is modeled as a multivariate Gaussian, canonical correlation analysis and information theory overlap because multivariate mutual information can be decomposed into sums involving canonical variables [30]. While this unification provides an elegant perspective, to this point it has not yet solved the issue of serial correlation. For instance, in Fig. 1 we characterise the effect of autocorrelation on the LR-based hypothesis tests—the *de facto* standard of significance testing linear-Gaussian model mutual information [31]. The results show that the estimates are significantly biased, yielding a FPR of 55% for mutual information (an 11-fold increase of the expected rate) and 90% for multivariate mutual information (an 18-fold increase) with linear-Gaussian models.

To overcome this issue, in this work we leverage the foundational knowledge of Bartlett et al. [8–11] to correct the sampling distributions for a large family of Gaussian-model linear-dependence measures based on the autocorrelation function of individual signals. After presenting preliminaries in Sec. II, we first show how a variety of univariate and multivariate measures can be decomposed into sums involving independent squared partial correlation coefficients in Sec. III. We then derive the Student's t-distribution of partial correlation coefficients between two autocorrelated signals in Sec. IV A and Sec. IV B, which is exact under the null hypothesis. Using this result, we are able to obtain the Bartlett-corrected F-distribution that each of the squared correlation terms follow through Sec. IV C-IV E; summing these F-distributions then gives the sampling distribution of the original dependence measure. We can then construct these distributions via Monte Carlo sampling, generate a p-value, and test the significance of the estimate. More specifically, we begin by providing the sampling distribution for mutual information (Sec. IV C) and conditional mutual information (Sec. IV D) under the Gaussian model. We then use the chain rule to construct sampling distributions for all other measures, such as multivariate mutual information (Sec. IV E), univariate (Sec. IV F) and multivariate Granger causality (Sec. IV G). Thus, our approach can be used to derive the sampling distribution of any measure that can be expressed in terms of conditional mutual information or partial correlation, e.g., canonical correlations and partial autocorrelation [6, 7] or information-theoretic measures such as predictive information [32, 33] and active information storage [34], under the linear-Gaussian assumption.

By expressing dependence as a sum of squared partial correlation terms, we identify two main sources of inflated Type I and II errors with the LR hypothesis tests [35] in Sec. IV. The first problem is that each independent term that goes into computing a dependence measure has an "effective sample size" (the number of samples in a dataset that are independent depending on the cross-correlation of the processes) that is potentially different to the original time-series length used in the calculation. When both processes are positively autocorrelated, the effective sample size is smaller than the original length and thus the LR test overestimates the statistic, potentially inducing spurious correlations (Type I errors). When instead one of the processes is negatively autocorrelated (and the other positively), the effective sample size is larger than the original dataset length and the LR test underestimates the statistic and may miss causalities (Type II errors). This effect is addressed for single calculations via the Bartlett correction, and extended for each squared partial correlation term in our sum here; notably, each term may be associated with a different effective sample size. The second source of inflated FPR comes from the exact distribution of partial correlation even if the samples are independent (i.e., not autocorrelated). When hypothesis testing partial correlation for an IID processes, the sample size is reduced by the dimension of the conditional process. Thus, for computing the null distribution, we must not only take into account the

Bartlett correction for each term that involves squared partial correlation but also the dimension of the conditional for that term in the chained sum [36]. Classical approaches to hypothesis testing, such as the LR and Wald tests, use null distributions that are only asymptotically valid and so do not account for these finite-sample effects. The severity of ignoring these corrections depends both on the time-series length and its autocorrelation and so we run a number of experiments with fixed sample size while changing other parameters to illustrate their effect.

Using numerical simulations throughout Sec. IV, and then on the brain-imaging data for Fig. 1 in Sec. V, we validate our corrected F-tests and characterize the effect of autocorrelation on the LR test. Our examples involve generating samples from two first-order independent AR processes and iteratively filtering the output signal such that the autocorrelation is increased for both time series; this allows for the process parameters to be modified while ensuring that the null hypothesis (of no inter-process 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 LR tests can be inflated by either increasing the autocorrelation (through filtering) or the number of conditionals (through increasing the dimension of mutual information or the history length of Granger causality). Given minimally-sufficient effective samples, however, we confirm that our Bartlett-corrected F-test remains consistent with the expected FPR of our hypothesis tests. We thus show that, in contrast, when using the LR test, the sensitivity and specificity for a large class of dependence measures can be arbitrarily low (approximately zero in certain instances) and overwhelmingly depends on the parameters of the underlying independent processes. Moreover, we provide the first exact approach for testing the significance of linear dependence between vector autoregressive processes. Given that measures such as Granger causality are specifically designed to account for autocorrelation, our theoretical and empirical findings suggest that autocorrelation-induced statistical errors caused by classical LR test (and others) may be even more prevalent in prior publications than previously suggested by a number of authors [1, 3, 4]. Our newly presented hypothesis tests enable statistically valid analysis of the linear dependence between multiple time series in future research. We expect our approach to be widely implemented in order to legitimise the inference of linear relationships within complex systems across myriad scientific applications, ranging from the macroevolution of climate dynamics to the microevolution of the brain.

II. PRELIMINARIES

We first introduce the definition, notation, and some standard assumptions for vector AR processes in Sec. II A. These can be obtained from a number of sources, e.g., see [7, 37, 38] or the chapters on multivariate time-series analysis in classic textbooks [6, 39]. Following this, we introduce some common measures that are used to infer dependence from information theory and econometrics in Sec. II B.

A. Multivariate time series

In this work, we focus on multivariate signals,

$$\{(Z_1(t)), \ldots, (Z_m(t))\}, \quad t = 0, \pm 1, \pm 2, \ldots, \tag{1}$$

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

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

we refer to the m series as an m-dimensional vector of multiple time series such that $\boldsymbol{Z}(t) \in \mathbb{R}^m$. Now, assume that $\mathbf{Z} = (\boldsymbol{Z}(t) : t \in \mathbb{Z})$ is a second-order stationary purely non-deterministic process with the mean removed, i.e., it is fully defined by its autocovariance. In many scenarios of practical interest, we assume that the time series $\boldsymbol{Z}$ can be represented by the causal state equations

$$\boldsymbol{Z}(t) - \boldsymbol{\Phi}(1)\boldsymbol{Z}(t-1) - \ldots - \boldsymbol{\Phi}(p)\boldsymbol{Z}(t-p) = \boldsymbol{a}(t), \tag{3}$$

with innovation process $(\boldsymbol{a}(t) : t \in \mathbb{Z})$ and a time-invariant matrix $\boldsymbol{\Phi}(u)$ that encodes the autoregression parameters. Thus, the observed dynamics are assumed to follow a p-order Markov process where each $\boldsymbol{a}(t) = (a_1(t), \ldots, a_m(t))'$ is uncorrelated with mean $\mathbf{0}$ and non-singular covariance matrix $\mathrm{Var}\,(\boldsymbol{a}(t)) = \boldsymbol{\Sigma}$. In this work we make the typical assumption that the innovation process is Gaussian such that $\boldsymbol{a}(t) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The form of (3) is known as the vector AR (VAR) model [7], and is one of the most common discrete-time stochastic process models. In digital filtering literature, it is equivalent to an all-pole infinite impulse response (IIR) filter [6], whose input is the white noise innovation process $\boldsymbol{a}$.

Following Geweke [40], we partition the $m \times 1$ column vector $\boldsymbol{Z}(t)$ into $k \times 1$ and $l \times 1$ subvectors denoted $\boldsymbol{X}(t) \in \mathbb{R}^k$ and $\boldsymbol{Y}(t) \in \mathbb{R}^l$, i.e., $\boldsymbol{Z}(t) = [\boldsymbol{X}(t)\,; \boldsymbol{Y}(t)]$ where the semi-colon denotes vertical (row-wise) concatenation. This reflects an interest in the relationship between the processes $\mathbf{X} = (\boldsymbol{X}(t) : t \in \mathbb{Z})$ and $\mathbf{Y} = (\boldsymbol{Y}(t) : t \in \mathbb{Z})$. As a consequence, the matrix $\boldsymbol{\Phi}(u)$ in (3) can be represented by block submatrices:

$$\boldsymbol{\Phi}(u) = \begin{bmatrix} \boldsymbol{\Phi}_{\boldsymbol{X}}(u) & \boldsymbol{\Phi}_{\boldsymbol{XY}}(u) \\ \boldsymbol{\Phi}_{\boldsymbol{YX}}(u) & \boldsymbol{\Phi}_{\boldsymbol{Y}}(u) \end{bmatrix},$$

where $\boldsymbol{\Phi}_{\boldsymbol{XY}}(u)$ is the causal influence of $\boldsymbol{Y}(t-u)$ on $\boldsymbol{X}(t)$.

B. Measures of linear dependence

In this section, we introduce two types of measures of linear dependence: correlation coefficients and

information-theoretic quantities under a linear Gaussian model. The former often deal with univariate ensembles or processes, and the latter more naturally extend to multivariate random vectors. Although these are known quantities, we have collated definitions from a variety of sources here in order to present these measures with consistent notation.

Each of the measures below are a function of the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$. Notationally, the covariance matrix $\boldsymbol{K_X}$ is the matrix whose (i,j) entry is the covariance $\mathrm{Cov}(X_i, X_j) = \mathrm{E}\left[(X_i - \mathrm{E}\left[X_i\right])(X_j - \mathrm{E}\left[X_j\right])\right]$. The partial covariance matrix $\boldsymbol{K_{XY|W}}$ is the covariance between X and Y if a third variable $\boldsymbol{W}$ were to be held constant, i.e., a matrix where the (i,j) entry is the partial covariance $\mathrm{Cov}(X_i, Y_j \mid \boldsymbol{W})$. In matrix notation, $\boldsymbol{K_{XY|W}}$ is defined as

$$\boldsymbol{K_{XY|W}} = \boldsymbol{K_{XY}} - \boldsymbol{K_{XW}}\boldsymbol{K_W}^{-1}\boldsymbol{K_{WY}}. \tag{4}$$

For a univariate process, the variance of X is $\mathrm{Cov}(X,X)$ and expressed as $\mathrm{Var}\left(X\right) = \sigma_X^2$.

1. Correlation coefficients

For a stationary process $\boldsymbol{Z}$, the covariance 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. Let $\gamma_{ij}(u) = \mathrm{Cov}(Z_i(t), Z_j(t+u))$ be the cross-covariance between $Z_i(t)$ and $Z_j(t+u)$. The cross-correlation is then defined as [6, 7]

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

If $\rho_{ii}(u) \neq 0$ for any $u > 0$, then the stochastic process Z_i exhibits autocorrelation, and the collection of $\rho_{ii}(u)$ for $u = 0, \pm 1, \pm 2, \ldots$ is generally called the autocorrelation function of Z_i. We denote the zero-lag correlation between Z_i and Z_j as $\rho_{ij} = \rho_{ij}(0)$. If X and Y are univariate ($k = l = 1$), the zero-lag correlation between these variables simplifies to Pearson correlation:

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}\left(X\right)\mathrm{Var}\left(Y\right)}}. \tag{6}$$

When conditioning the correlation on another (possibly multivariate) process $\boldsymbol{W}$, we obtain partial correlation, which is computed as [41]

$$\rho_{XY\cdot\boldsymbol{W}} = \frac{\mathrm{Cov}(X,Y|\boldsymbol{W})}{\sqrt{\mathrm{Var}\left(X|\boldsymbol{W}\right)\mathrm{Var}\left(Y|\boldsymbol{W}\right)}}. \tag{7}$$

A special case of this is partial autocorrelation, which is commonly used to infer an AR order [6, 7], as discussed in Appendix A.

2. Mutual information

Mutual information $\mathcal{I}_{\boldsymbol{X};\boldsymbol{Y}}$ quantifies the amount of information in a process $\boldsymbol{X}$ obtained by observing another process $\boldsymbol{Y}$. For the linearly-coupled Gaussian variables considered here, this quantity can be computed from the joint $(\boldsymbol{X}, \boldsymbol{Y})$ covariance matrix $\boldsymbol{K_{XY}}$ and the independent covariance matrices $\boldsymbol{K_X}$ and $\boldsymbol{K_Y}$ [42]:

$$\mathcal{I}_{\boldsymbol{X};\boldsymbol{Y}} = -\frac{1}{2}\log\left(\frac{|\boldsymbol{K_{XY}}|}{|\boldsymbol{K_X}||\boldsymbol{K_Y}|}\right) \tag{8}$$

Similar to Pearson correlation, we can quantify the mutual information between $\boldsymbol{X}$ and $\boldsymbol{Y}$ when taking into account the outcome of a third process $\boldsymbol{W}$. This is known as the conditional mutual information of $\boldsymbol{X}$ and $\boldsymbol{Y}$, given $\boldsymbol{W}$. For linear Gaussian variables, we can express the conditional mutual information $\mathcal{I}_{\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{W}}$ in terms of three partial covariance matrices [42]:

$$\mathcal{I}_{\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{W}} = -\frac{1}{2}\log\left(\frac{|\boldsymbol{K_{XY|W}}|}{|\boldsymbol{K_{X|W}}||\boldsymbol{K_{Y|W}}|}\right), \tag{9}$$

3. Granger causality and transfer entropy

Granger causality was introduced as a directed measure of causal dependence between two AR models. It measures the dependence of a predictee process $\boldsymbol{X}$ on a predictor $\boldsymbol{Y}$, in the context of the past of $\boldsymbol{X}$ [17]. Thus, unlike correlation coefficients or mutual information, Granger causality was derived for the purpose of measuring dependence between processes that exhibit autocorrelation. 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 is often inferred using interventions); see [43] for a differentiation of these concepts.

To present this notion, we require some new notation. Let $\boldsymbol{X}^{(p)}(t) = [\boldsymbol{X}(t-1)\,;\ldots;\,\boldsymbol{X}(t-p)]$ represent the past p states of $\boldsymbol{X}(t)$, i.e., $\boldsymbol{X}^{(p)}(t)$ is a $kp \times 1$ column vector (since $\boldsymbol{X}(t) \in \mathbb{R}^k$). Now, consider two additional multivariate processes: one that represents the relevant history of the predictee $\boldsymbol{X}^{(p)} = (\boldsymbol{X}^{(p)}(t) : t \in \mathbb{Z})$ and one that captures relevant history of the predictor $\boldsymbol{Y}^{(q)} = (\boldsymbol{Y}^{(q)}(t) : t \in \mathbb{Z})$. Granger causality is then typically computed as a LR of partial covariance matrices [40, 42]:

$$\mathcal{F}_{\boldsymbol{Y}\to\boldsymbol{X}}(p,q) = \log\left(\frac{\left|\boldsymbol{K}_{\boldsymbol{X}|\boldsymbol{X}^{(p)}}\right|}{\left|\boldsymbol{K}_{\boldsymbol{X}|\boldsymbol{X}^{(p)}\boldsymbol{Y}^{(q)}}\right|}\right), \tag{10}$$

where the numerator is the covariance of the "reduced" model (i.e., with the past of $\boldsymbol{X}$ but without $\boldsymbol{Y}$) and the denominator is the covariance of the "full" model (i.e., taking into account both the past of the predictee $\boldsymbol{X}$ and the past of the predictor $\boldsymbol{Y}$). The optimal orders p and q are often inferred by statistical tests such

as partial autocorrelation [7] or active information storage (discussed in Appendix A). Further, in general, the variables representing the past $\boldsymbol{X}^{(p)}(t)$ 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 [44–46], where a time delay $\tau \in \mathbb{Z}$ is included such that the relevant past states are now contained in the variable $\boldsymbol{X}^{(p,\tau)}(t) = [\boldsymbol{X}(t-\tau)\,;\,\ldots\,;\,\boldsymbol{X}(t-p\tau)]$, or any other statistically significant set of variables [47, 48]. However, for simplicity, in this work we assume that the data are generated by a causal model such as in Eq. (3) and thus use a consecutive sequence of variables as the history of $\boldsymbol{X}(t)$.

Granger causality can be conditioned on another multivariate process $\boldsymbol{W}$ by including it in both covariance matrices [4, 42, 49]:

$$\mathcal{F}_{\boldsymbol{Y}\to\boldsymbol{X}|\boldsymbol{W}}(p,q) = \log\left(\frac{\left|\boldsymbol{K}_{\boldsymbol{X}|\boldsymbol{X}^{(p)}\boldsymbol{W}}\right|}{\left|\boldsymbol{K}_{\boldsymbol{X}|\boldsymbol{X}^{(p)}\boldsymbol{Y}^{(q)}\boldsymbol{W}}\right|}\right), \tag{11}$$

which allows us to compute directional dependence of $\boldsymbol{X}$ on $\boldsymbol{Y}$ in the context of some tertiary process $\boldsymbol{W}$.

The information-theoretic quantity known as transfer entropy [50, 51] was independently established as the mutual information between the current state $\boldsymbol{X}$ and the past of $\boldsymbol{Y}$, in the context of the past of $\boldsymbol{X}$:

$$\mathcal{T}_{\boldsymbol{Y}\to\boldsymbol{X}}(p,q) = \mathcal{I}_{\boldsymbol{X};\boldsymbol{Y}^{(q)}|\boldsymbol{X}^{(p)}}. \tag{12}$$

Transfer entropy has gained popularity in measuring the dependence between nonlinear time series, from robotics [52] to neuroscience [53]. As shown by Barnett et al. [42], the transfer entropy and Granger causality are equivalent for Gaussian innovation terms, i.e., $\mathcal{F}_{\boldsymbol{Y}\to\boldsymbol{X}}(p,q) = 2\,\mathcal{T}_{\boldsymbol{Y}\to\boldsymbol{X}}(p,q)$. This can be seen by comparing Eqs. (9) and (10) and, crucially, means that Granger causality can be interpreted as a conditional mutual information.

III. DEPENDENCE AS SQUARED CORRELATION

In this section, we will show that the above information-theoretic quantities reduce to expressions involving squared correlation coefficients, under the linear Gaussian assumption. This is perhaps unsurprising because linear dependence is implied by a reduction in uncertainty when adding more variables into a model, and a reduction in variance is quantified by squared correlation [54]. However, this representation not only unifies the relationship between typically disconnected dependence measures but it is a crucial intermediate step in deriving the sample distributions for linear-dependence measures under temporal autocorrelation. Specifically, later in Sec. IV we present the sampling distribution for partial correlation between autocorrelated processes. Thus, by formulating any measure as a function of partial correlation we can construct its sampling distribution. This equivalence between information-theoretic measures and correlation coefficients has previously been shown for both mutual information and conditional mutual information; however, these formulations, while elegant, remained without without practical consequences. In this section we extend these results to the multivariate case, which can then be used to derive a novel decomposition of Granger causality in terms of partial correlation.

A. Mutual information

A well-known result in the literature is that mutual information for univariate processes is a function of the squared correlation coefficient [27, 28]. That is, for univariate X and Y, we have the following covariance matrices:

$$K_X = \mathrm{Var}\,(X)\,,\quad K_Y = \mathrm{Var}\,(Y)\,, \tag{13}$$

$$\boldsymbol{K}_{XY} = \begin{bmatrix} \mathrm{Var}\,(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{Var}\,(Y) \end{bmatrix}, \tag{14}$$

Taking the determinant of the joint covariance matrix $|\boldsymbol{K}_{XY}| = \sigma_X^2\sigma_Y^2(1-\rho_{XY}^2)$, we obtain the established result via Eq. (8) that, for linear Gaussian variables [28],

$$\begin{aligned}\mathcal{I}_{X;Y} &= -\frac{1}{2}\log\left(\frac{|\boldsymbol{K}_{XY}|}{|K_X||K_Y|}\right) \\ &= -1/2\,\log(1-\rho_{XY}^2).\end{aligned} \tag{15}$$

Similar formulas relating squared correlation to uncertainty reduction exist ubiquitously throughout literature. For instance, ρ_{XY}^2 is the coefficient of determination, $1-\rho_{XY}^2$ is the unexplained variance resulting from the prediction [54], and $\sqrt{1-\rho_{XY}^2}$ is T.L. Kelley's alienation coefficient [55], which is equivalent to mutual information when taken inside the log.

B. Conditional mutual information

It has further been established that the conditional mutual information between two univariate processes X and Y, conditioned on a third (potentially multivariate) process $\boldsymbol{W}$, can be expressed as a function of partial correlation [56].

In this scenario, we have the following covariance matrices:

$$K_{X|\boldsymbol{W}} = \mathrm{Var}\,(X|\boldsymbol{W})\,,\quad K_{Y|\boldsymbol{W}} = \mathrm{Var}\,(Y|\boldsymbol{W})\,, \tag{16}$$

$$\boldsymbol{K}_{XY|\boldsymbol{W}} = \begin{bmatrix} \mathrm{Var}\,(X|\boldsymbol{W}) & \mathrm{Cov}(X,Y|\boldsymbol{W}) \\ \mathrm{Cov}(Y,X|\boldsymbol{W}) & \mathrm{Var}\,(Y|\boldsymbol{W}) \end{bmatrix}. \tag{17}$$

Taking the determinant of the joint conditional covariance matrix yields

$$\left|\boldsymbol{K}_{XY|\boldsymbol{W}}\right| = \sigma_{X|\boldsymbol{W}}^2\sigma_{Y|\boldsymbol{W}}^2(1-\rho_{XY|\boldsymbol{W}}^2),$$

from which we obtain an expression for conditional mutual information as [56]

$$\begin{aligned}\mathcal{I}_{X;Y|\boldsymbol{W}} &= -\frac{1}{2}\log\left(\frac{|\boldsymbol{K}_{XY|\boldsymbol{W}}|}{|K_{X|\boldsymbol{W}}||K_{Y|\boldsymbol{W}}|}\right) \\ &= -1/2\,\log(1-\rho^2_{XY\cdot\boldsymbol{W}}).\end{aligned} \tag{18}$$

C. Multivariate mutual information

As a novel contribution, we can now use the known expressions from the previous section and apply the chain rule for conditional mutual information to represent multivariate measures of linear dependence in terms of squared partial correlations.

The chain rule provides a decomposition of multivariate mutual information as a sum of conditional mutual information terms. That is, mutual information between multivariate $\boldsymbol{X}$ and $\boldsymbol{Y}$ can be expressed as

$$\begin{aligned}\mathcal{I}_{\boldsymbol{X};\boldsymbol{Y}} &= \sum_{g=1}^{k}\sum_{h=1}^{l}\mathcal{I}_{X_g;Y_h|\boldsymbol{V}^{\{gh\}}_{\boldsymbol{XY}}} \\ &= -1/2\sum_{g=1}^{k}\sum_{h=1}^{l}\log\left(1-\rho^2_{X_iY_j\cdot\boldsymbol{V}^{\{gh\}}_{\boldsymbol{XY}}}\right),\end{aligned} \tag{19}$$

with the conditional for the (g,h) term given by

$$\boldsymbol{V}^{\{gh\}}_{\boldsymbol{XY}} = \Big[\boldsymbol{X}_{1:g-1}\,;\,\boldsymbol{Y}_{1:h-1}\Big], \tag{20}$$

where $\boldsymbol{X}_{1:g} = [X_1\,;\,\ldots\,;\,X_g]$ with $g \le k$, and $\boldsymbol{X}_{1:g} = \emptyset$ with $g = 0$. This approach can be trivially extended to conditioning on another tertiary variable $\boldsymbol{W}$. Now, the decomposition is

$$\mathcal{I}_{\boldsymbol{X};\boldsymbol{Y}|\boldsymbol{W}} = -1/2\sum_{g=1}^{k}\sum_{h=1}^{l}\log\left(1-\rho^2_{X_iY_j\cdot\boldsymbol{V}^{\{gh\}}_{\boldsymbol{XY}|\boldsymbol{W}}}\right), \tag{21}$$

with conditional

$$\boldsymbol{V}^{\{gh\}}_{\boldsymbol{XY}|\boldsymbol{W}} = \Big[\boldsymbol{V}^{\{gh\}}_{\boldsymbol{XY}}\,;\,\boldsymbol{W}\Big]. \tag{22}$$

Since conditional mutual information is a basis for a number of other dependence measures, we can build on these results to provide similar decompositions of a large family of linear-dependence measures based on this result. One measure that is fundamental to studying the dependence between autoregressive processes is Granger causality.

D. Granger causality

Granger causality is equivalent to the mutual information between $\boldsymbol{X}$ and $\boldsymbol{Y}^{(q)}$, conditioned on $\boldsymbol{X}^{(p)}$ (see Eq. (12)). As such, we can leverage the previous section and introduce a decomposition of Granger causality into a sum of conditional mutual information terms, and thus present it as a sum of log terms involving partial correlations. Many other information-theoretic and log-likelihood measures could be similarly decomposed in order to derive their exact hypothesis tests.

We first present the decomposition of Granger causality between univariate X and Y, with a fixed predictor history of $q = 1$ and an arbitrary predictee history length p. For this case, the relevant history of X is described in the multivariate process $\boldsymbol{X}^{(p)}(t) = \big[X(t-1)\,;\,\ldots\,;\,X(t-p)\big]$ and the relevant history of Y is the original process lagged by one time-step: $Y^1(t) = Y(t-1)$. From Eq. (18), this can be expressed as a log function of partial correlation as

$$\mathcal{F}_{Y\to X}(p,1) = -\log\left(1-\rho^2_{XY^1\cdot\boldsymbol{X}^{(p)}}\right). \tag{23}$$

For an arbitrary predictor history q, Granger causality becomes akin to (conditional) multivariate mutual information, and is thus a sum of log terms, similar to Eq. (21). To present this, we denote a backshifted (lagged) variable by upper indices without parenthesis, e.g., $Y^j(t) = Y(t-j)$ is the same variable lagged by j time steps. Further, recall that for an arbitrary history length j, we have $\boldsymbol{Y}^{(j)}(t) = [Y(t-1)\,;\,\ldots\,;\,Y(t-j)]$. We can now express Granger causality as a sum involving independent partial correlation terms:

$$\mathcal{F}_{Y\to X}(p,q) = -\sum_{j=1}^{q}\log\left(1-\rho^2_{XY^j\cdot\boldsymbol{V}^{\{j\}}_{Y\to X}}\right), \tag{24}$$

with the jth conditional as the matrix

$$\boldsymbol{V}^{\{j\}}_{Y\to X} = \Big[\boldsymbol{X}^{(p)}\,;\,\boldsymbol{Y}^{(j-1)}\Big], \tag{25}$$

and, again, $\boldsymbol{Y}^{(0)}_h = \emptyset$.

With higher dimensional processes $\boldsymbol{X}$ and $\boldsymbol{Y}$, this decomposes into three nested sums. That is, multivariate Granger causality can be expressed as:

$$\mathcal{F}_{\boldsymbol{Y}\to\boldsymbol{X}}(p,q) = -\sum_{g=1}^{k}\sum_{h=1}^{l}\sum_{j=1}^{q}\log\left(1-\rho^2_{X_gY^j_h\cdot\boldsymbol{V}^{\{ghj\}}_{\boldsymbol{X}\to\boldsymbol{Y}}}\right), \tag{26}$$

with the (g,h,j) term using the conditional matrix

$$\boldsymbol{V}^{\{ghj\}}_{\boldsymbol{Y}\to\boldsymbol{X}} = \Big[\boldsymbol{X}^{(p)}_{1:g}\,;\,\boldsymbol{Y}^{(q)}_{1:h-1}\,;\,\boldsymbol{Y}^{(j-1)}_{h}\Big]. \tag{27}$$

Similar to mutual information, Granger causality can be conditioned on another process $\boldsymbol{W}$ as:

$$\mathcal{F}_{\boldsymbol{Y}\to\boldsymbol{X}|\boldsymbol{W}}(p,q) = -\sum_{g=1}^{k}\sum_{h=1}^{l}\sum_{j=1}^{q}\log\left(1-\rho^2_{X_gY^j_h\cdot\boldsymbol{V}^{\{ghj\}}_{\boldsymbol{X}\to\boldsymbol{Y}|\boldsymbol{W}}}\right), \tag{28}$$

with

$$\boldsymbol{V}^{\{ghj\}}_{\boldsymbol{Y}\to\boldsymbol{X}|\boldsymbol{W}} = \Big[\boldsymbol{V}^{\{ghj\}}_{\boldsymbol{Y}\to\boldsymbol{X}}\,;\,\boldsymbol{W}\Big]. \tag{29}$$

The decomposition of a large family of information-theoretic (and econometric) measures into terms involving partial correlations is not only a novel unification of two previously disjoint fields; it also facilitates our understanding of their sampling distributions when applied to multiple autocorrelated time series in the following section.

IV. INFERRING LINEAR DEPENDENCE

Having decomposed the linear-dependence measures in terms of squared correlation coefficients, we are now able to describe how their estimates would be distributed if there were no actual dependence between the processes, i.e., under the null hypothesis. With these distributions, we can then infer whether there is a statistical dependence between the processes at a given level of confidence; this procedure is known as hypothesis testing.

Hypothesis tests for linear-dependence measures are often derived from one of the three classical procedures for maximum likelihood estimates: the Wald test, the score test, and the LR test [6, 7, 39]. The sample distribution of estimates under any of these tests are only known asymptotically and so, as we will show, they will fail under serial correlation with less data. Due to the strong overlap between all three approaches, we only discuss the LR test here, which is a commonly used test for testing the significance of information-theoretic measures and is based on Wilks' theorem. His theorem states that a test statistic that is twice the LR of nested models will follow a χ^2-distribution. However, under serial correlation, the assumptions of independence of samples underscoring this result fail and these sampling distributions can be shown to be biased [4]. In the previous section, we expressed a number of these measures in terms of squared correlation coefficients. We derived these expressions because the sampling distributions of correlation coefficients for autocorrelated processes have been extensively studied by Bartlett [8, 9] and others. In the next few sections, we leverage this knowledge to obtain large-sample [57] distributions of these estimates.

Since we are now dealing with statistical estimates, we assume access to a finite-length dataset over which to compute the measures. Without loss of generality, the dataset comprises T observations of the sample path of stochastic process $\boldsymbol{Z}$, i.e., $\boldsymbol{z} = (\boldsymbol{z}(t) : t \in [1, T])$. The estimate of a quantity is indicated by a hat, with any subscripted random variable X replaced by its realisation x, e.g., $\hat{\rho}_{xy}$ is the estimate of ρ_{XY}. The cross-covariance for a sample path is computed as

$$\hat{\gamma}_{ij}(u) = c_{ij}(u) = \frac{1}{T}\sum_{t=1}^{T} z_i(t) z_j(t+u). \tag{30}$$

This can be used to compute the sample cross-correlation coefficients:

$$\hat{\rho}_{ij}(u) = r_{ij}(u) = \frac{c_{ij}(u)}{\sqrt{c_{ii}(0)c_{jj}(0)}}. \tag{31}$$

In the next section we discuss known results on the inefficiency of sample cross-correlations (under autocorrelation). We then build on these results to derive large-sample distributions for the more advanced linear-dependence measures introduced above.

A. Linear correlation coefficient

Correlation coefficients measure the dependence between two univariate processes where each sample $\boldsymbol{z}(t) = (x(t), y(t))'$ of the dataset is generated from a bivariate normal distribution. The null hypothesis is that there is no correlation $\rho_{XY} = 0$ between the processes. Under this assumption, the sample correlation coefficient for white variables X and Y follows Student's t-distribution:

$$r_{xy}\sqrt{\frac{T-2}{1-r_{xy}^2}} \sim t(T-2). \tag{32}$$

In this work, however, we are concerned with time series data and thus the assumptions underlying the distribution in (32) often fail because the processes X and Y are serially correlated. Instead, when they exhibit autocorrelation, there are a smaller number of "effective samples" than T in the dataset due to each sample's inherent correlation with the next. This effective sample size can be inferred from a number of formulae, the earliest of which is due to Bartlett [8, 9]. Bartlett's formula states that the large-sample null distribution of correlation between linear Gaussian processes (or any zero-skew noise process) is normal with mean zero and variance

$$\mathrm{Var}\,(r_{ij}) \approx [\eta(Z_i, Z_j)]^{-1},$$

which depends on the autocorrelation function of $\boldsymbol{Z}$:

$$\eta(Z_i, Z_j) = T\Big[\sum_{u=\infty}^{\infty} \rho_{ii}(u)\rho_{jj}(u)\Big]^{-1}. \tag{33}$$

In practice, the unknown autocorrelations $\rho_{ii}(u)$ are replaced the estimates $r_{ii}(u)$ obtained from the sample path $\boldsymbol{z}$, which gives us an estimate of the variance via η. Moreover, in a stationary process, these autocorrelations are symmetric, and so we can replace this infinite sum with only the positive lags over the dataset [7]:

$$\eta(z_i, z_j) = T\Big[1 + 2\sum_{u=1}^{T-1} r_{ii}(u) r_{jj}(u)\Big]^{-1}. \tag{34}$$

This problem of recovering autocorrelation functions from data is often considered nontrivial and has resulted

in procedures such as tapering to handle noisy estimates [2, 41]. We cover this concept and some other considerations for Bartlett's formula briefly in Appendix B.

For testing the hypothesis of no correlation between time series, therefore, the number of effective samples N_{xy} is computed from Bartlett's formula:

$$N_{xy} = \eta(x, y), \tag{35}$$

and the null distribution of sample correlation r_{xy} between autocorrelated paths x and y is given as by a Bartlett-corrected version of Eq. (32):

$$r_{xy}\sqrt{\frac{N_{xy}-2}{1-r_{xy}^2}} \sim t(N_{xy}-2). \tag{36}$$

Naturally, when x and y are white, $N_{xy} = T$ and this reduces to the standard t-test for Pearson correlation.

By using the LHS of Eq. (36) as a statistic, we can construct a p-value from the distribution to test against a nominated level of significance. Specifically, this statistic is input to the cumulative distribution function (CDF) of Student's t-distribution with $T-2$ degrees of freedom, and the output (the complement of the p-value) is the probability that such a correlation or higher would be observed under the null hypothesis. More precisely, we define an arbitrary significance level α, which is the probability of rejecting the null hypothesis even if it were true—this is often set to 5% or 1%. If the p-value is below the significance level, then we deduce that there is a correlation. A false positive occurs when the p-value is below the significance level α but there is no actual correlation between the variables, i.e., $\rho_{XY} = 0$. Ideally, we expect the proportion of false positives (the FPR) to match the significance level α, i.e., one would expect an FPR of 5% for $\alpha = 0.05$ significance level. Similar procedures can be used for all linear-dependence measures below.

B. Partial correlation

In order to use the above logic in uncovering sample distributions for the measures in Sec. II B, we need to extend the results to partial correlation. For partial correlation, the degrees of freedom in the t-test are reduced by the number of variables upon which are conditioned [58]. That is, consider a partial correlation $r_{xy\cdot\boldsymbol{w}}$, with the dimensionality of the conditional $\boldsymbol{w}(t)$ denoted $\dim(\boldsymbol{w}(t)) = c$. A common approach to computing partial correlation is through a Pearson correlation of the residuals from linear regression of x and y on $\boldsymbol{w}$ [41]. Specifically, one first computes the residual vectors

$$e_{x|\boldsymbol{w}} = x - \bar{x}(\boldsymbol{w})$$
$$e_{y|\boldsymbol{w}} = y - \bar{y}(\boldsymbol{w}),$$

where $\bar{x}(\boldsymbol{w})$ denotes the projection of $\boldsymbol{w}$ onto x. These fitted values are obtained by ordinary least squares $\bar{x}(\boldsymbol{w}) = (\boldsymbol{P}x)'$ with projection matrix $\boldsymbol{P} = \boldsymbol{w}'(\boldsymbol{w}\boldsymbol{w}')^{-1}\boldsymbol{w}$, and so $e_{x|\boldsymbol{w}}$ and $e_{y|\boldsymbol{w}}$ are simply the variables x and y after their linear dependence on $\boldsymbol{w}$ has been removed. Then, the sample partial correlation is equivalent to the sample correlation between $e_{x|\boldsymbol{w}}$ and $e_{y|\boldsymbol{w}}$ [41], i.e., $r_{xy\cdot\boldsymbol{w}} = r_{e_{x|\boldsymbol{w}}e_{y|\boldsymbol{w}}}$. Thus, the sampling distribution under the null hypothesis can be treated equivalently to that of the correlation coefficients above:

$$r_{xy\cdot\boldsymbol{w}}\sqrt{\frac{N_{xy|\boldsymbol{w}}-2}{1-r_{xy\cdot\boldsymbol{w}}^2}} \sim t(N_{xy|\boldsymbol{w}}-2), \tag{37}$$

with effective sample size

$$N_{xy|\boldsymbol{w}} = \eta\big(e_{x|\boldsymbol{w}}, e_{y|\boldsymbol{w}}\big) - c. \tag{38}$$

Here, the effective sample size $\eta(e_{x|\boldsymbol{w}}, e_{y|\boldsymbol{w}})$ is computed from Eq. (34) 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. Another crucial addition is that the dimension of the conditional process $\dim(\boldsymbol{w}(t)) = c$ further reduces the effective sample size [58], for the same reason regressors reduce the degrees of freedom in standard F-tests [4].

C. Mutual information

We can obtain the sampling distribution of mutual information (under autocorrelation) from its expression as a function of the correlation coefficient in Eq. (15). The sample mutual information can be computed as

$$\hat{\mathcal{I}}_{x;y} = -1/2\,\log(1-r_{xy}^2). \tag{39}$$

The classical LR hypothesis test assumes that mutual information between univariate Gaussian processes under the null hypothesis of no relationship follows the distribution [31, 59]:

$$2T\,\hat{\mathcal{I}}_{x;y} \sim \chi^2(1). \tag{40}$$

Using our expression of mutual information as a correlation coefficient, we obtain a different large-sample distribution to Eq. (40). Our approach is to use the sampling distribution of r_{xy}^2 in order to obtain the null distribution for $\mathcal{I}_{x;y}$ when x and y are autocorrelated. First, if a random variable Z has Student's t-distribution with degrees of freedom N, then Z^2 has an F-distribution with parameters 1 and N, i.e., $Z^2 \sim F(1, N)$. Thus, the square of the statistic in Eq. (36) (the LHS) follows an F-distribution:

$$(N_{xy}-2)\frac{r_{xy}^2}{1-r_{xy}^2} \sim F(1, N_{xy}-2), \tag{41}$$

with the effective sample size N_{xy} defined earlier in Eq. (35). From Eq. (41), we can now obtain the approximate distribution for mutual information between

autocorrelated processes. Rearranging (39) gives squared correlation as a function of sample mutual information

$$r_{xy}^2 = 1 - \exp\left(-2\,\hat{\mathcal{I}}_{x;y}\right), \tag{42}$$

which we can substitute into (41) in order to construct an F-distributed statistic under the null:

$$(N_{xy} - 2)[\exp\left(2\,\hat{\mathcal{I}}_{x;y}\right) - 1] \sim F(1, N_{xy} - 2). \tag{43}$$

This equation is similar to the well-known F-test for information-theoretic quantities [4, 5], however with a Bartlett-corrected effective sample size N_{xy}, rather than the original sample size T, making it exact for limited time-series data.

Although Eq. (43) follows an analytic distribution, later in the paper we present more complex statistics that are no longer analytically tractable. For consistency then, here we present the general Monte Carlo approach used to generate an empirical CDF of the null distribution. Let the tilde denote a sample from a distribution, e.g., $\widetilde{F}(1, N_{xy} - 2)$ is a sample from the $F(1, N_{xy} - 2)$ distribution. Then, a random sample $\widetilde{\mathcal{I}}_{x;y}$ of the distribution under the null hypothesis that x and y are independent (but themselves may be autocorrelated) is then generated by:

$$\widetilde{\mathcal{I}}_{x;y} \coloneqq \frac{1}{2}\log\left(\frac{\widetilde{F}(1, N_{xy} - 2)}{N_{xy} - 2} + 1\right). \tag{44}$$

Using a large number of independent samples $\widetilde{\mathcal{I}}_{x;y}$ obtained in this way allows us to generate an empirical CDF of the null distribution from which we can obtain the p-value for testing the significance of the measured value $\hat{\mathcal{I}}_{x;y}$.

Another established approach to empirical generation of a CDF for the null distribution involves permuting, re-drawing or rotating the samples of one variable x or y [4, 31, 60]. Clearly, however, permuting or re-drawing will completely destroy the autocorrelation profile of that variable, making this empirical distribution approach the analytic LR distribution here. Even rotating, while seemingly only breaking the time series at one point, will have strong impacts across the autocorrelation profile for multiple lags. 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 LR test.

Finally, we can see how the large-sample distribution we presented in Eq. (43) becomes equivalent to the classical LR distribution (40) by taking the limit $T \to \infty$. First, the LHS of Eq. (43) can be approximated by the LHS of Eq. (40) through a first-order Taylor expansion; as the number of samples T increases, this approximation improves (i.e., the LHS converges). Second, it is known that if a random variable $X \sim F(d_1, d_2)$, then the variable $Y = \lim_{d_2\to\infty} d_1 X$ has a $\chi^2(d_1)$ distribution. Since the second parameter d_2 of the F-distribution in Eq. (43) is a function of T, then as $T \to \infty$, $d_2 \to \infty$ as well. Thus, given enough samples T, the distributions (the RHS) converge.

1. Numerical evaluation

We performed a number of Monte Carlo simulations that illustrate how our new hypothesis test based on (43) and (44) performs against the classical LR test in (40). In particular, we are interested in studying how each test behaves under increasing levels of autocorrelation of both X and Y.

For these tests, we use an AR model similar to the example proposed in [4], with two processes X and Y that have no interdependence. We begin with the first-order bivariate AR model:

$$\boldsymbol{Z}(t) - \boldsymbol{\Phi}(1)\boldsymbol{Z}(t-1) = \boldsymbol{a}(t), \tag{45}$$

with

$$\boldsymbol{Z}(t) = \begin{bmatrix} X(t) \\ Y(t) \end{bmatrix}, \quad \boldsymbol{\Phi}(1) = \begin{bmatrix} \Phi_X(1) & 0 \\ 0 & \Phi_Y(1) \end{bmatrix}, \tag{46}$$

where $\Phi_X(1) = 0.3$, $\Phi_Y(1) = -0.8$, and the innovation process has unit variance $\boldsymbol{\Sigma} = \boldsymbol{I}_2$. To illustrate the affect of higher autocorrelation on the FPR of both methods, we digitally filter the original processes X and Y with two types of filters: a finite-impulse response (FIR) linear-phase least-squares filter and an infinite-impulse response (IIR) Butterworth filter. Both filters were low-pass, with their cutoff set to a normalised frequency of $\pi/2$ radians. The filter order is variable such that increasing the order increases the autocorrelation of the signals. It would, of course, be possible to manually choose the parameters $\boldsymbol{\Phi}(u)$ for increasing AR order u; however, filtering artefacts from time series is a common preprocessing step for practitioners (even differencing the signal is a form of an FIR filter), and so it is important to show the effect this can have on linear-dependence measures. Moreover, 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 this autocorrelation.

Following [4], to illustrate the finite-sample effects, we generate relatively short stationary time series with $T = 2^9 = 512$ samples to obtain our dataset $\boldsymbol{z} = [x\,;\,y]$. After simulating these processes, we estimate the mutual information using Eq. (39), and obtain p-values by sampling the CDF of both the asymptotic distribution (40) and our proposed large-sample distribution (44). We then consider the FPR to be the proportion of p-values that are significant at $\alpha = 0.05$ over 1000 trials.

The results of these tests are shown in Fig. 2, where the uncorrected FPR refers to the LR χ^2 distributions (40) and the corrected tests refer to our newly derived distributions (44). The plots illustrate that the LR tests

overestimate the estimates for 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 8th order filter. From these figures, we can see that the FPR for the LR hypothesis tests is higher than nominal for all significance levels $\alpha \in (0, 1)$. In comparison, our Bartlett-corrected F-test procedure yields an FPR that approaches the expected value for all filter orders. Although the FPR for the IIR filtered signals appear marginally biased, none of the FPRs for individual filter orders are statistically different from 5%. One potential source of this increased estimation variance is from poor estimation of the autocorrelation functions, another is from the approximation made in Eq. (33); both of these are discussed further in Appendix B.

A filter order of zero in Fig. 2 refers to using the causal model (45) without any digital filtering. Interestingly, in this case, the LR yielded less than the nominal 0.05 FPR. This occurs when the number of effective samples becomes greater than the original sample size $N_{xy} > T$. Referring to Bartlett's formula (34), this is due to the product of negative autocorrelation exhibited by the Y process (induced by $\Phi_Y = -0.8$) and the positive autocorrelation exhibited by the X process (induced by $\Phi_X = 0.3$). Counter-intuitively, this would imply that the samples are less autocorrelated than white variables. Although the reduction of false positives appears to be a good result for the LR test, it implies that the statistical power (i.e., the true-positive rate (TPR)) will now be lessened. To verify this, we performed 1000 trials where there was a small dependence of X on Y. That is, the X process is now generated from state equation (45) with AR matrix

$$\mathbf{\Phi}(1) = \begin{bmatrix} \Phi_X(1) & \Phi_{XY}(1) \\ 0 & \Phi_Y(1) \end{bmatrix}$$

and $\Phi_{XY}(1) = 0.03$, and $\Phi_X(1)$ and $\Phi_Y(1)$ unchanged. The TPR was 0.049 (SE of 0.0068) for the LR test and 0.1570 (SE of 0.0115) for our F-test. Thus, the sensitivity of our hypothesis test is three times higher than the LR test in this scenario.

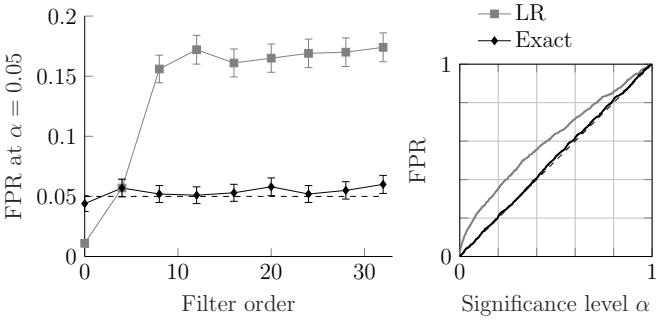
D. Conditional mutual information

Conditional mutual information can be treated similarly to mutual information. From Eq. (18), we have that

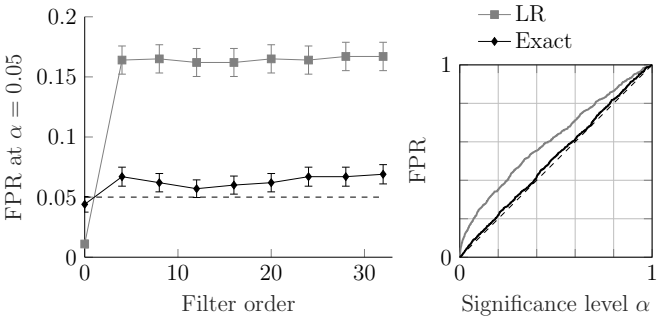
$$\hat{\mathcal{I}}_{x;y|\boldsymbol{w}} = -1/2 \log(1 - r^2_{xy\cdot\boldsymbol{w}}) \tag{47}$$

with $r_{xy\cdot\boldsymbol{w}}$ the sample partial correlation between x and y, given $\boldsymbol{w}$. From Wilk's theorem, the LR distribution of this statistic under the null hypothesis of no relationship is [31, 59]

$$2T\,\hat{\mathcal{I}}_{x;y|\boldsymbol{w}} \sim \chi^2(1). \tag{48}$$



(a) FIR filter.



(b) IIR filter.

FIG. 2: Correction of otherwise inflated FPRs with our technique: for mutual information estimates under autocorrelation, as a function of filter type and order. The mutual information was measured (using Eq. (39)) between independent univariate time series (generated by Eq. (46)) after filtering, and tested using the LR test (Eq. (40)) and our exact test (Eq. (44)). The plots show the effect of increasing the filter order on the FPR for both an (a) FIR filter and (b) IIR filter. The errorbars represent the standard error of the FPR over trials at a given filter order. The subplots on the right capture the FPR for all potential significance levels α with an 8th order filtered signal. An ideal distribution is where the FPR equals α and thus sits perfectly on the diagonal, as per our tests.

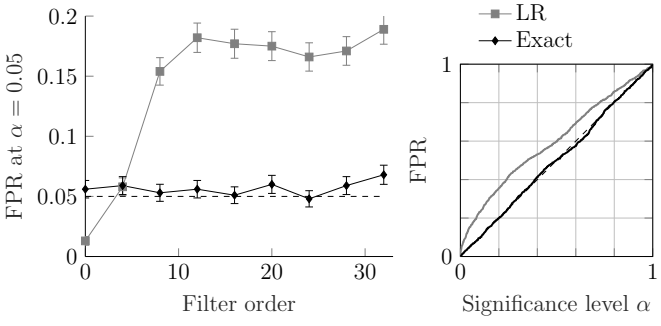
Following the same logic as above, however, under the null hypothesis we can construct the following F-distributed statistic for conditional mutual information:

$$(N_{xy|\boldsymbol{w}} - 2)[\exp(2\,\hat{\mathcal{I}}_{x;y|\boldsymbol{w}}) - 1] \sim F(1, N_{xy|\boldsymbol{w}} - 2), \tag{49}$$

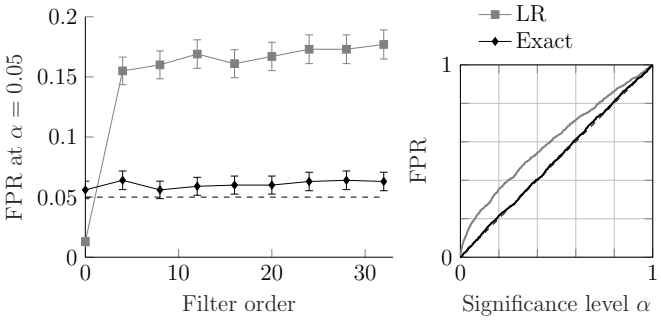
with the effective sample size $N_{xy|\boldsymbol{w}}$ from Eq. (38). This enables us to generate samples $\widetilde{\mathcal{I}}_{x;y|\boldsymbol{w}}$ for an empirical distribution under the null, as per Eq. (44), for conditional mutual information as:

$$\widetilde{\mathcal{I}}_{x;y|\boldsymbol{w}} \coloneqq \frac{1}{2} \log\left(\frac{\widetilde{F}(1, N_{xy|\boldsymbol{w}} - 2)}{N_{xy|\boldsymbol{w}} - 2} + 1\right). \tag{50}$$

There are two important distinctions here from the original sampling distribution. The first is that we now have an effective sample size that changes depending on the

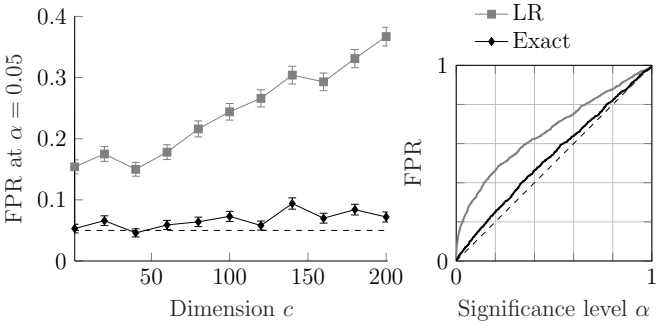


(a) FIR filter.

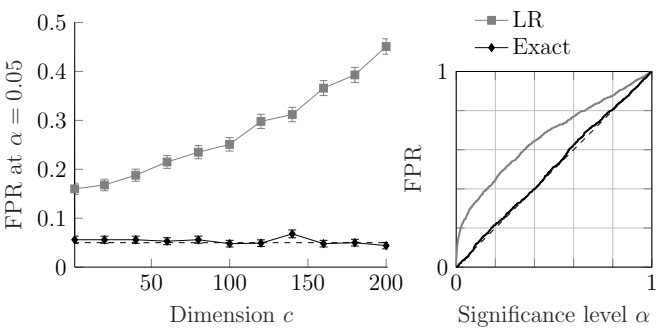


(b) IIR filter.

FIG. 3: Correction of otherwise inflated FPRs with our technique: for conditional mutual information estimates under autocorrelation, as a function of filter type and order. Conditional mutual information was measured (using Eq. (47)) between two univariate time series, given a third univariate process (generated by Eq. (51) with $k = l = c = 1$), after filtering, and tested for significance using the LR test (Eq. (48)) our exact test (Eq. (50)). The subplots on the right show the FPR for each significance level α when the signal is filtered with an 8th order filter.



(a) FIR filter.



(b) IIR filter.

FIG. 4: Correction of otherwise inflated FPRs with our technique: for conditional mutual information estimates under autocorrelation, as a function of dimension of the conditional processes. Time series are generated and analysed as per Fig. 3, aside from an increasing dimension of conditional variable $\dim(\boldsymbol{W}(t)) = c$ and use of 8th order FIR and IIR filters. Estimates are significance tested via the LR (Eq. (48)) and our (Eq. (50)) tests. The subplots on the right show the FPR for each significance level α when $c = 100$.

autocorrelation function of signals X and Y. The second is that the sample size is further reduced by the dimension of the conditional term $\boldsymbol{W}$. Although the asymptotic results still hold (see Sec. IV C), both of these reductions introduce a significant bias in the estimation of linear dependence for many real-world applications.

1. *Numerical evaluation*

For validating our hypothesis tests with conditional mutual information, we extend the model (46) to include a third process with arbitrary dimension $\dim(\boldsymbol{W}(t)) = c$. That is, the process is generated from (45) with

$$\boldsymbol{Z}(t) = \begin{bmatrix} X(t) \\ Y(t) \\ \boldsymbol{W}(t) \end{bmatrix}, \quad \boldsymbol{\Phi}(1) = \begin{bmatrix} \Phi_X(1) & 0 & \mathbf{0} \\ 0 & \Phi_Y(1) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Phi}_{\boldsymbol{W}}(1) \end{bmatrix} \tag{51}$$

where the AR parameter is $\boldsymbol{\Phi}_{\boldsymbol{W}}(1) = 0.4\boldsymbol{I}_c$ and the innovation process has unit variance $a_{\boldsymbol{W}}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$. The three processes X, Y, and $\boldsymbol{W}$ are then digitally filtered (independently, along the time dimension only) with increasing filter orders in the same way as discussed in Sec. IV C.

The results, presented in Fig. 3 are similar to those of mutual information tests in Fig. 2. Again, increasing the filter order increases the FPR for the LR test, yet our approach remains consistent with the expected FPR for a given significance level α.

As discussed above, an important distinction between the distributions for mutual information (44) and conditional mutual information (50) is that the second parameter of the F-distribution not only includes the Bartlett corrected sample size but also the dimension of the conditional c. To show this effect, we generate the X and Y processes and filter the signal with an 8th order FIR and IIR filter the same as before; however, we increase the number of independent processes $\dim(\boldsymbol{W}(t)) = c$ in the multivariate conditional. The result is shown in Fig. 4, where the FPR of the LR test increases linearly with the dimension, however our new approach is consistent with

the expected FPR of 0.05. Hence, even when the autocorrelation function is the same, the dimension of the conditional must also be included in the analysis.

E. Multivariate mutual information

The sampling distribution of multivariate measures can be obtained by extending the results from the previous sections using the chain rule. From (19), the estimate for conditional mutual information with multivariate processes is computed as

$$\hat{\mathcal{I}}_{\boldsymbol{x};\boldsymbol{y}|\boldsymbol{w}} = -\frac{1}{2}\sum_{g=1}^{k}\sum_{h=1}^{l}\log\Big(1 - r^2_{x_g y_h \cdot \boldsymbol{v}^{\{gh\}}_{\boldsymbol{x}\boldsymbol{y}|\boldsymbol{w}}}\Big), \tag{52}$$

with design matrix for the (g, h) term given by

$$\boldsymbol{v}^{\{gh\}}_{\boldsymbol{x}\boldsymbol{y}|\boldsymbol{w}} = \Big[\boldsymbol{x}_{1:g-1}\,;\,\boldsymbol{y}_{1:h-1}\,;\,\boldsymbol{w}\Big]. \tag{53}$$

Following Wilk's theorem, again, this statistic under the null hypothesis of no relationship follows the asymptotic distribution [31, 59]:

$$2T\,\hat{\mathcal{I}}_{\boldsymbol{x};\boldsymbol{y}|\boldsymbol{w}} \sim \chi^2(kl). \tag{54}$$

However, using our results, we can obtain the sampling distribution for measurements between autocorrelated processes. Each partial correlation in (52) is obtained through ordinary least squares. The residual processes for the (g, h) term in (52) is

$$\begin{aligned} e^{\{gh\}}_{\boldsymbol{x}|\boldsymbol{v}} &= x_g - \bar{x}_g\Big(\boldsymbol{v}^{\{gh\}}_{\boldsymbol{x}\boldsymbol{y}|\boldsymbol{w}}\Big) \\ e^{\{gh\}}_{\boldsymbol{y}|\boldsymbol{v}} &= y_h - \bar{y}_h\Big(\boldsymbol{v}^{\{gh\}}_{\boldsymbol{x}\boldsymbol{y}|\boldsymbol{w}}\Big). \end{aligned} \tag{55}$$

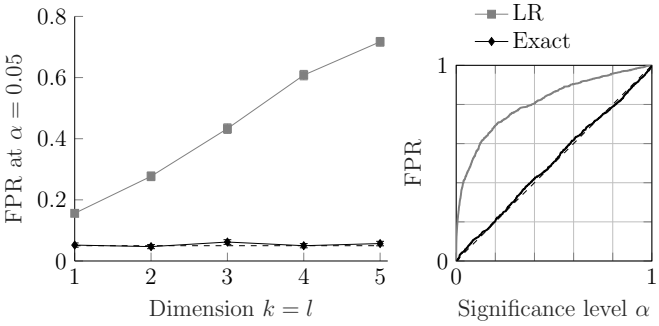
Then, the number of effective samples for each term is given by:

$$N^{\{gh\}}_{\boldsymbol{x}\boldsymbol{y}|\boldsymbol{w}} = \eta\Big(e^{\{gh\}}_{\boldsymbol{x}|\boldsymbol{v}_{hg}}, e^{\{gh\}}_{\boldsymbol{y}|\boldsymbol{v}_{hg}}\Big) - \dim\Big(\boldsymbol{v}^{\{gh\}}(t)\Big). \tag{56}$$

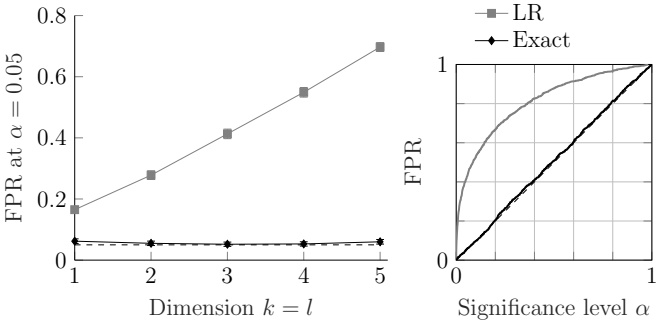
While we can form an F-distributed statistic for each term in the sum of Eq. (52) (as per Eq. (49)), we do not have an analytic form for the distribution of the sum. As such, we obtain the sampling distribution of the null for multivariate conditional mutual information empirically by summing over individual samples drawn as described by Eq. (50):

$$\widetilde{\mathcal{I}}_{\boldsymbol{x};\boldsymbol{y}|\boldsymbol{w}} \coloneqq \frac{1}{2}\sum_{g=1}^{k}\sum_{h=1}^{l}\log\left(\frac{\widetilde{F}(1, N^{\{gh\}}_{\boldsymbol{x}\boldsymbol{y}|\boldsymbol{w}} - 2)}{N^{\{gh\}}_{\boldsymbol{x}\boldsymbol{y}|\boldsymbol{w}} - 2} + 1\right). \tag{57}$$

The sums in (57) have been ordered first over the dimensions of $\boldsymbol{Y}$, and then over the dimensions of $\boldsymbol{X}$. The order of these summations are arbitrary, however we imposed an order solely for the effective sample size formula.



(a) FIR filter.



(b) IIR filter.

FIG. 5: Correction of otherwise inflated FPRs with our technique: for multivariate mutual information estimates under autocorrelation, as a function of dimension of the processes. The mutual information was measured (using Eq. (52)) between the multivariate time series (generated by Eq. (58) with an increasing dimension k and l of $\boldsymbol{X}$ and $\boldsymbol{Y}$) passed through 8th order FIR and IIR filters, and tested for significance using the LR test (Eq. (54)) and our exact test (Eq. (57)). The subplots on the right show the FPR for each significance level α when $k = l = 3$.

The above derivation assumes that each F-distributed random variable is uncorrelated to all others. This may present an issue if the residuals were correlated. This could be mediated by using multivariate versions of Bartlett's formula that take into account covariance; unfortunately, this requires generating sums of correlated F-distributed variables. This is discussed in more depth in Appendix B.

1. *Numerical evaluation*

In order to demonstrate the effect of autocorrelation on multivariate mutual information, we extend the univariate state equations (46) to allow for multivariate subprocesses:

$$\boldsymbol{Z}(t) = \begin{bmatrix}\boldsymbol{X}(t)\\ \boldsymbol{Y}(t)\end{bmatrix}, \quad \boldsymbol{\Phi}(1) = \begin{bmatrix}\boldsymbol{\Phi}_{\boldsymbol{X}}(1) & 0\\ 0 & \boldsymbol{\Phi}_{\boldsymbol{Y}}(1)\end{bmatrix} \tag{58}$$

where $\mathbf{\Phi}_{\boldsymbol{X}}(1) = 0.3\boldsymbol{I}_k$ and $\mathbf{\Phi}_{\boldsymbol{Y}}(1) = -0.8\boldsymbol{I}_l$, and $\mathbf{\Sigma} = \boldsymbol{I}_m$. Instead of increasing the filter order, however, we generate these state equations for an increasing dimension of k and l, with $k, l \in [1, 5]$. These signals are then filtered along the temporal dimension using 8th order FIR and IIR filters. By doing so, we ensure that there is no correlation between signals within the same subprocess, i.e., $\rho_{ij} = 0$ for all $i, j \in [1, m]$. An internal correlation between any of these subprocesses may further increase the FPR for the LR test; our tests of Granger causality in Sec. IV F-IV G naturally incorporate examples with correlated subprocesses in the underlying mutual information.

The results are shown in Fig. 5, where the increasing the dimension approximately linearly increases the FPR of the original LR test, yet our F-test remains constant at 0.05 FPR. This increase of the FPR is expected, given the increase in dimensionality of the conditional process was not taken into account in the LR test.

F. Univariate Granger causality

As discussed in the above, Granger causality is equivalent to conditional mutual information for Gaussian variables. As such, we can build on the previously presented hypothesis tests to obtain Bartlett-corrected distributions for Granger causality when applied to univariate or multivariate processes for an arbitrary AR order.

We will begin by discussing the simplest case of Granger causality where the model comprises a fixed predictor order of $q = 1$ and an arbitrary predictee order p. Here, the variable describing the history of $\boldsymbol{X}$ at time t is $\boldsymbol{X}^{(p)}(t)$ and the variable describing the history of $\boldsymbol{Y}$ is $Y^1(t)$. From (23), we can see that Granger causality can be computed as the partial correlation between the sample paths of x and y^1, given $\boldsymbol{x}^{(p)}$,

$$\hat{\mathcal{F}}_{y\to x}(p, 1) = -\log\Big(1 - r^2_{xy^1\cdot\boldsymbol{x}^{(p)}}\Big). \tag{59}$$

Univariate Granger causality is often assumed to be χ^2-distributed under the null hypothesis of no partial correlation [4, 40, 42, 49]

$$T\,\hat{\mathcal{F}}_{y\to x}(p, 1) \sim \chi^2(1). \tag{60}$$

However, we can obtain the finite-sample distribution for Granger causality between univariate processes X and Y in the same way as above for conditional mutual information IV D. That is, by using residuals obtained from a regression of x on $\boldsymbol{x}^{(p)}$ and of y^1 on $\boldsymbol{x}^{(p)}$

$$\begin{aligned} e_{x|\boldsymbol{x}^{(p)}} &= x - \bar{x}\big(\boldsymbol{x}^{(p)}\big), \\ e_{y|\boldsymbol{x}^{(p)}} &= y^1 - \bar{y}^1\big(\boldsymbol{x}^{(p)}\big), \end{aligned}$$

we get an effective sample size of

$$N_{y\to x} = \eta\big(e_{x|\boldsymbol{x}^{(p)}}, e_{y|\boldsymbol{x}^{(p)}}\big) - \dim\Big(\boldsymbol{x}^{(p)}(t)\Big),$$

and an F-distributed statistic under the null hypothesis can be constructed as:

$$(N_{y\to x-2})[\exp\left(2\,\hat{\mathcal{F}}_{y\to x}(p, 1)\right) - 1] \sim F(1, N_{y\to x-2} - 2). \tag{61}$$

The sampling distributed of the null can also be constructed empirically from:

$$\widetilde{\mathcal{F}}_{y\to x}(p, 1) \coloneqq \log\left(\frac{\widetilde{F}(1, N_{y\to x} - 2)}{N_{y\to x} - 2} + 1\right). \tag{62}$$

From Eq. (24), for higher order predictor processes—i.e., an arbitrarily large q—Granger causality is computed from a sum of partial correlation terms:

$$\hat{\mathcal{F}}_{y\to x}(p, q) = -\sum_j^q \log\Big(1 - r^2_{xy^j\cdot\boldsymbol{v}^{\{j\}}_{y\to x}}\Big). \tag{63}$$

Here, the jth term can be computed explicitly from residuals. Recall that the set of variables up to j samples back is denoted by a bracketed upper index $Y^{(j)}(t)$. Then, the design matrix becomes

$$\boldsymbol{v}^{\{j\}}_{y\to x} = \Big[\boldsymbol{x}^{(p)}\,;\, y^{(j-1)}\Big],$$

The null distribution of this measure from a log-likelihood perspective is [40]

$$T\,\hat{\mathcal{F}}_{y\to x}(p, q) \sim \chi^2(q). \tag{64}$$

For these models, however, the Granger causality decomposes as a sum of log terms (63), where each term involves a squared partial correlation $r^2_{xy^j\cdot\boldsymbol{v}^{\{j\}}_{y\to x}}$, between residuals computed as

$$\begin{aligned} e^{\{j\}}_{x|\boldsymbol{v}_{y\to x}} &= x - \bar{x}\Big(\boldsymbol{v}^{\{j\}}_{y\to x}\Big) \\ e^{\{j\}}_{y|\boldsymbol{v}_{y\to x}} &= y^j - \bar{y}^j\Big(\boldsymbol{v}^{\{j\}}_{y\to x}\Big), \end{aligned}$$

and thus the effective sample size is different for each term:

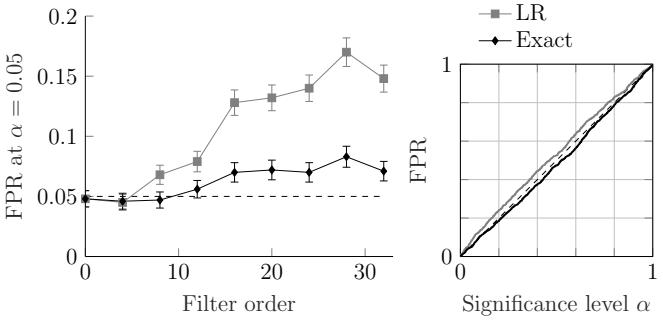
$$N^{\{j\}}_{y\to x} = \eta\Big(e^{\{j\}}_{x|\boldsymbol{v}_{y\to x}}, e^{\{j\}}_{y|\boldsymbol{v}_{y\to x}}\Big) - \dim\Big(\boldsymbol{v}^{\{j\}}_{y\to x}(t)\Big).$$

We can empirically generate the sampling distribution of the null, following Eq. (57) akin to conditional mutual information, via:

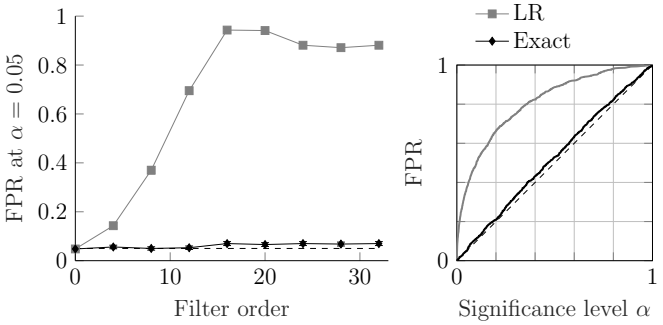
$$\widetilde{\mathcal{F}}_{y\to x}(p, q) \coloneqq \sum_{j=1}^q \log\left(\frac{\widetilde{F}(1, N^{\{j\}}_{y\to x} - 2)}{N^{\{j\}}_{y\to x} - 2} + 1\right). \tag{65}$$

1. Numerical evaluation

In order to evaluate the performance of both the LR and our new hypothesis test on estimates of Granger causality, we use the same procedure discussed in

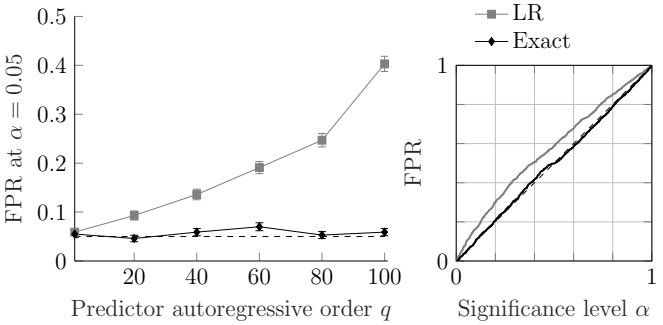


(a) FIR filter.

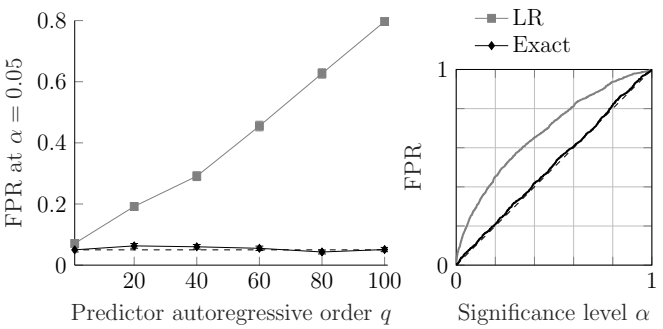


(b) IIR filter.

FIG. 6: Correction of otherwise inflated FPRs with our technique: for univariate Granger causality estimates under autocorrelation, as a function of filter type and order. Granger causality, with history lengths p and q chosen optimally, is estimated (using Eq. (63)) between univariate time series (generated by Eq. (46)) after filtering. Estimates are then tested using the LR test (Eq. (64)) and our exact test (Eq. (65)). The subplots on the right show the FPR for each significance level α with an 8th order filter.



(a) FIR filter.



(b) IIR filter.

FIG. 7: Correction of otherwise inflated FPRs with our technique: for univariate Granger causality estimates under autocorrelation, as a function of predictor history length q. Time series are generated and analysed as per Fig. 6, aside from specification of q and use of 8th order FIR and IIR filters. The subplots on the right show the FPR for each significance level α when $q = 20$; this was chosen because the average optimal order was inferred to be approximately 18.

Sec. IV C. That is, the univariate model (46) is used to generate 512 samples of the X and Y processes, which are then passed through a filter. 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 partial autocorrelation (see Appendix A). We then compute Granger causality via Eq. (63) and use this estimate to obtain a p-value from both the LR χ^2-distribution (64) and our Bartlett-corrected F-distribution (65). This is performed 1000 times in order to obtain a FPR of each approach. Crucially, 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 correlation amongst the variables within the source $\boldsymbol{Y}^{(q)}(t)$ and conditional $\boldsymbol{X}^{(p)}(t)$.

The results shown in Fig. 6 illustrate that increasing the autocorrelation length via filtering increases the FPR of Granger causality under the LR tests. In contrast, the corrected tests for mutual information perform better than the corrected tests for Granger causality, remaining consistent with the nominal significance of $\alpha = 0.05$ for the IIR filters as well as the lower FIR filter orders tested. However, it is clear that the FPR of Granger causality using our new approach is slightly above the nominal level for higher FIR filter orders (and higher than that of mutual information, c.f. Fig. 2). We hypothesise that this is because any small numerical errors in each of the q terms used to compute the distribution will compound. We discuss the potential source of these errors in Appendix B. Note however that the problematic high FIR filter orders here (above 15) are unlikely to be used in practice for such short time series, due to the dramatic impact they have on the signal here.

In addition to optimally setting the history lengths p and q, it common to manually set a large enough history to presumably capture all past dependence, e.g., see [4, 61]. In Fig. 7, we show that this approach actually inflates the FPR even further when using the LR hypothesis test. This is illustrated by increasing the predicter history length q from one to 200 (in increments of 20), whilst holding the autocorrelation length (filter or-

der) constant. This effectively introduces more terms in Eq. (63), causing the χ^2-distribution to diverge further from the adjusted distribution, however we may also expect small errors to accumulate in our proposed distribution. As expected, the FPR of the LR test linearly increases in this range, however our distribution remains relatively consistent with the 5% FPR. This linear increase of the FPR is somewhat counter-intuitive to the notion of Granger causality, where one may expect capturing more history would reduce spurious correlations. However, the opposite is true, simply due to a lack of the correct distributions (in the case of the LR test).

G. Multivariate Granger causality

Finally, we present the most general case: the sample distribution of Granger causality with an arbitrarily large dimension for both $\boldsymbol{X}$ and $\boldsymbol{Y}$, and arbitrary AR orders p and q for each of those dimensions. From Eq. (26) this can be computed as:

$$\hat{\mathcal{F}}_{\boldsymbol{x}\to\boldsymbol{y}}(p,q) = -\sum_{g=1}^{k}\sum_{h=1}^{l}\sum_{j=1}^{q}\log\Big(1 - r^2_{x_g y_h^j \cdot \boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}}^{\{ghj\}}}\Big). \quad (66)$$

That is, the (g,h,j) term is the conditional Granger causality for dimension g of $\boldsymbol{X}$, dimension h of $\boldsymbol{Y}$, and the predictor sample j steps back of the hth sub-process of $\boldsymbol{Y}$. This is conditioned on all dimensions and predictor samples below g, h, and j. We can express this through the design matrix for the (g,h,j) term:

$$\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}}^{\{ghj\}} = \Big[\boldsymbol{x}_{1:g}^{(p)}\,;\,\boldsymbol{y}_{1:h-1}^{(q)}\,;\,\boldsymbol{y}_h^{(j-1)}\Big].$$

By virtue of Wilk's theorem, the classical assumption is that—when $\boldsymbol{X}$ does not depend on $\boldsymbol{Y}$—estimates of multivariate Granger causality are distributed as [40]:

$$T\,\hat{\mathcal{F}}_{\boldsymbol{y}\to\boldsymbol{x}}(p,q) \sim \chi^2(klq). \quad (67)$$

From (66), however, the sample estimate involves sums of correlations between residuals. The (g,h,j) term is computed from the residuals

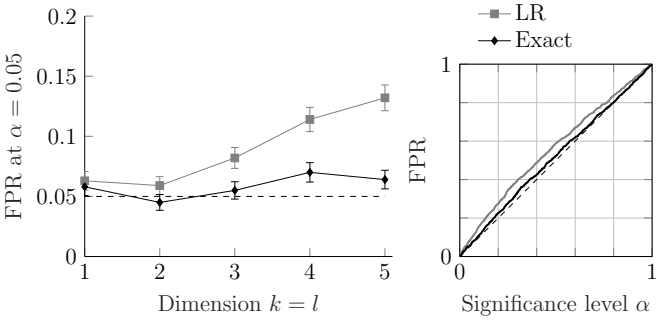
$$e_{\boldsymbol{x}|\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}}}^{\{ghj\}} = x_g - \bar{x}_g\Big(\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}}^{\{ghj\}}\Big)$$
$$e_{\boldsymbol{y}|\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}}}^{\{ghj\}} = y_h^j - \bar{y}_h^j\Big(\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}}^{\{ghj\}}\Big),$$

Again, the effective sample size is different for each regression and computed as:

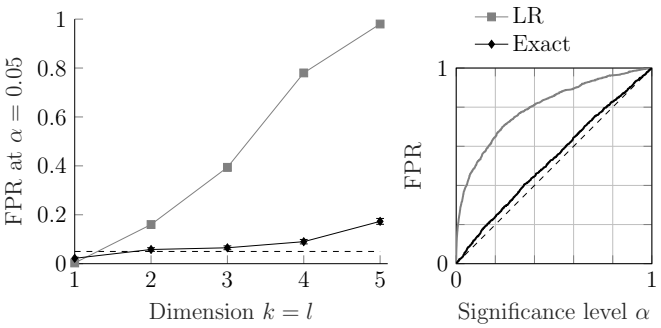
$$N_{\boldsymbol{y}\to\boldsymbol{x}}^{\{ghj\}} = \eta\Big(e_{\boldsymbol{x}|\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}}}^{\{ghj\}}, e_{\boldsymbol{y}|\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}}}^{\{ghj\}}\Big) - \dim\Big(\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}}^{\{ghj\}}(t)\Big)$$

As per conditional mutual information and univariate Granger causality, we can empirically sample from the null distribution as:

$$\widetilde{\mathcal{F}}_{\boldsymbol{y}\to\boldsymbol{x}}(p,q) \coloneqq \sum_{g=1}^{k}\sum_{h=1}^{l}\sum_{j=1}^{q}\log\left(\frac{\widetilde{F}(1, N_{\boldsymbol{y}\to\boldsymbol{x}}^{\{ghj\}} - 2)}{N_{\boldsymbol{y}\to\boldsymbol{x}}^{\{ghj\}} - 2} + 1\right). \quad (68)$$



(a) FIR filter.



(b) IIR filter.

FIG. 8: Correction of otherwise inflated FPRs with our technique: for multivariate Granger causality estimates under autocorrelation, as a function of dimensions k and l of predictor and predictee processes. Granger causality, with each subprocess' history length chosen optimally, is estimated (using Eq. (66)) between multivariate time series (generated by Eq. (58)) after filtering via 8th order FIR and IIR filters. Estimates are then tested using the LR test (Eq. (67)) and our exact test (Eq. (68)). The subplots on the right show the FPR for each significance level α when $k = l = 3$.

Here, we have used the same history lengths (p and q for each predictee and predictor subprocess), however this could be easily generalised.

Conditioning Granger causality on another variable $\boldsymbol{W}$ can be treated equivalently. The measure is computed as

$$\hat{\mathcal{F}}_{\boldsymbol{x}\to\boldsymbol{y}|\boldsymbol{w}}(p,q) = -\sum_{g=1}^{k}\sum_{h=1}^{l}\sum_{j=1}^{q}\log\Big(1 - r_{x_g y_h^j \cdot \boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}^{\{ghj\}}}\Big). \quad (69)$$

The only difference is that the effective sample size for each term is reduced further. By letting

$$\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}^{\{ghj\}} = \Big[\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}}^{\{ghj\}}\,;\,\boldsymbol{w}\Big],$$

we obtain residuals $e_{\boldsymbol{x}|\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}}^{\{ghj\}}$ and $e_{\boldsymbol{y}|\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}}^{\{ghj\}}$ in the same way as above. From this, the effective sample size is given as:

$$N_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}^{\{ghj\}} = \eta\Big(e_{\boldsymbol{x}|\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}}^{\{jhg\}}, e_{\boldsymbol{y}|\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}}^{\{ghj\}}\Big) - \dim\Big(\boldsymbol{v}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}^{\{ghj\}}(t)\Big),$$

and the null distribution for multivariate conditional Granger causality can be obtained through samples:

$$\widetilde{\mathcal{F}}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}(p,q) \coloneqq \sum_{g=1}^{k}\sum_{h=1}^{l}\sum_{j=1}^{q}\log\left(\frac{\widetilde{F}(1, N^{\{ghj\}}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}-2)}{N^{\{ghj\}}_{\boldsymbol{y}\to\boldsymbol{x}|\boldsymbol{w}}-2}+1\right). \tag{70}$$

1. Numerical evaluation

Using the multivariate state equations (58), we can evaluate the effect of increasing the dimensionality of both processes on Granger causality inference. Here, the dimension of the subprocesses $\boldsymbol{X}$ and $\boldsymbol{Y}$ range from one to five. Note that the number of terms involved in computing Granger causality (or the null distributions) is $k \times l \times q$ (59). 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 our tests. Thus, for these experiments we fix the history length of the predictor to $q = 1$.

The results are shown in Fig. 8, where the LR hypothesis tests inflate the FPR close to 100% for higher dimensional processes. Although our corrected tests perform well for moderate dimensionality, when $k, l > 3$ with the IIR filter, the FPR of our F-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 much lower FPR than the LR tests. Moreover, a poorly conditioned regression can be easily tested for in practice. So, we conclude that with minimally-sufficient samples, 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.

V. CASE STUDY: HUMAN CONNECTOME PROJECT DATASET

Functional magnetic resonance imaging (fMRI) research is particularly abundant with short, autocorrelated time series [1–5]. In these applications, the blood-oxygen level-dependent (BOLD) data is translated into a slowly-varying (and thus highly autocorrelated) multivariate time series that traces the haemodynamic response of different regions of the brain. Digital filtering is then commonly used as a preprocessing step to reduce line noise and other common 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. In this section, we validate our theory by showing that the numerical simulations presented earlier reflect realistic scenarios that are encountered in practice. To this end, we use the Human Connectome Project (HCP) resting state fMRI (rsfMRI) dataset [13] to characterise the FPR of linear-dependence measures between completely independent brain regions.

The HCP rsfMRI dataset comprises 500 subjects imaged at a 0.72 s sampling rate for 15 minutes in the (relatively quiescent) resting state. This results in 1200 samples of spatially dense time series data, which is 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 samples. The raw (BOLD) data of each region was then preprocessed by removing the DC component, detrending, applying a 3rd order zero-phase Butterworth bandpass filter (0.01-0.08 Hz). These are common techniques used to remove potential filtering artefacts. We also removed 200 samples from the start and the end of the time series, in order to minimise filter initialisation effects. This leaves $T = 800$ samples 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, multivariate mutual information (this time between pairs of two rather than single, unrelated time series) and Granger causality in the same way as discussed for the simulated time-series experiments in Sec. IV. We use the same hypothesis testing procedure of both the classical LR and our newly proposed Bartlett-corrected test for both the filtered and unfiltered data with a 5% significance level.

Referring to Fig. 1, our results clearly demonstrate the inflated FPR under the LR tests as well as the corrected FPRs due to our tests. Briefly, the FPR of the LR test for: mutual information was 44% (before filtering) and 55% (after); multivariate mutual information was 82% (before filtering) and 90% (after); Granger causality was 20% (before filtering) and 63% (after). This represents an inflated FPR of between nine and 18 times the nominal significance level for a real-world dataset. As expected, the inherent autocorrelation in the unfiltered time series led to inflated FPR, which was worsened by filtering. In contrast, by using our Bartlett-corrected hypothesis-testing procedure, we maintain between 5-7% FPRs for all measures of linear dependence when applied to both unfiltered and filtered time series. These results are in agreement with those of our numerical simulations, indicating that the classical LR hypothesis test is invalid for fMRI analysis and any other short, autocorrelated dataset, whereas our hypothesis tests do not suffer from the same biases.

VI. DISCUSSION

We have shown that the autocorrelation exhibited in typical stationary time-series data inflates the FPR of linear-dependence inference procedures when using the classical hypothesis tests. By framing different dependence measures in unified theoretical terms, we provide the first demonstration of how Bartlett's correction can be applied to derive exact hypothesis tests for Granger causality and mutual information, in their unconditional, conditional, and multivariate forms. These measures—particularly Granger causality—are used in a very wide range of disciplines, modelling myriad physical processes from anthropogenic climate change [21, 22] to the brain dynamics of dementia patients [62]. The continued use of flawed classical testing procedures in empirical sciences is problematic, making it imperative that the corrections reported here are incorporated into such analysis.

Indeed, it was known that digital filtering led to an increased FPR for Granger causality for shorter time-series [4], 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, climate analysis 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. The issue was magnified in fields with short, highly-autocorrelated time-series, as demonstrated in Fig. 1 and Sec. V for computational neuroscience using fMRI recordings. Using classical LR testing after standard pre-processing and filtering of typical fMRI data, we found arbitrarily large FPRs. Crucially, applying our new method of correcting the distribution under the null hypothesis of no relationship resulted in FPRs consistent with the expected levels across a wide combination of filtering types and linear-dependence measures. These results strongly suggest that our method should be incorporated into all null hypothesis testing for Granger causality and related dependence measures. Moreover, our work suggests that relationships established using classical testing should perhaps be revisted—particularly in fields that have high levels of autocorrelation and limited data.

In more detail, we first demonstrated how a large class of dependence measures (including mutual information and Granger causality) can be written as functions of independent partial correlations. We then derived an exact test for partial correlation between autocorrelated processes based on seminal work by Bartlett and others. Since conditional mutual information can be expressed in terms of partial correlation, we were able to derive its null distribution. Using the chain rule, we then showed how to derive the null distribution for any linear-Gaussian model information-theoretic measure, with particular examples of multivariate mutual information and Granger causality. Through comprehensive numerical simulations and widely-accessible brain-imaging data, we have shown that the classical hypothesis tests can have arbitrarily poor specificity and sensitivity when testing the linear dependence between autocorrelated processes. In contrast, our exact distributions maintain the expected FPR with minimally-sufficient samples.

A major challenge of handling autocorrelation for more involved dependence measures, such as mutual information and Granger causality, 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 multivariate mutual information in Sec. IV E used examples that involved dependence between multivariate processes, where the subprocesses were all independent. When instead extending this approach to Granger causality in Sec. IV G 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 robust extension of Bartlett's dependence studies 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 dependence between autoregressive models. The classical approach to inferring the significance of (potentially multivariate) Granger causality estimates has been via asymptotically valid hypothesis tests such as the LR and Wald tests [40]. 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, resulting in statistical errors under classical hypothesis testing. This means that, ironically, using these tests induces errors exactly when the assumptions underlying Granger causality are valid. One might logically surmise that this issue could be mediated by accounting for a longer history of the process, i.e., condition on additional autoregressive variables. However, in Fig. 7 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 [6, 7, 39]—the other is its mean, which has no effect on scale-invariant dependence measures such as Granger causality, mutual information, and Pearson correlation. We conclude that the classical hypothesis tests cannot be used to reliably infer the significance of Granger causality or mutual information estimates when applied to autoregressive signals and instead our exact tests should be used, particularly for limited time-series data.

Throughout this work, we have made the assumption that the time-series innovations are Gaussian and that all autoregressive 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 non-parametric density estimation techniques such as nearest neighbour or kernel

methods [31]. Spurious estimates of the nearest neighbour counts have previously been observed for autocorrelated signals by Theiler [29], who provided a solution by excluding samples that are close in time to a target. This is now a popular approach to effectively account for autocorrelation in density estimation for nonlinear time-series analysis [63]. In fact, in introducing transfer entropy—now understood as a model-free extension of Granger causality [42]—Schreiber explicitly recommended the use of a Theiler window (also known as serial- or dynamic-correlation exclusion) when kernel estimation methods are used [50]. The Theiler window approach has been demonstrated to control FPR for such estimators in practice, e.g., see the Supporting Information of Novelli et al. [48] for a similar experiment with transfer entropy on the same HCP rsfMRI dataset used in our Fig. 1 and Sec. V. Yet the Theiler window approach, however effective, remains a heuristic with no theoretical guarantees and, similar to Pearson correlation, autocorrelation is often poorly accounted for in practical estimation of transfer entropy (e.g. Weber et al. [64] report a contrasting finding of increased FPR in transfer entropy inference due to filtering, however the use of a Theiler window is not specified). We hypothesize that the methods outlined in our work could be extended to provide a more rigorous approach to handling autocorrelated nonlinear time series through, e.g., nonlinear versions of Bartlett's formula [65], facilitating a broader class of information-theoretic measures.

Finally, the dependence structure of the stochastic processes discussed in this work is assumed to be in the time domain and on the (observable) output of a process. However, other models and approaches consider different forms of dependence with correlation structures that could similarly increase FPRs. Future work will be required to consider handling such correlation structures in a similar fashion to that which we have presented. For instance, one could consider autoregression on: the innovation terms (the moving average (MA) or linear state-space model); both the output and the innovation terms (the ARMA model), and the output and innovations in the frequency domain (spectral models). Many extensions to Granger causality have been made to handle these scenarios, e.g., the ARMA model has been addressed by [38] and more recently through state-space Granger causality [66]. The latter was recently shown to still exhibit significant bias with a reasonable number of samples [67]. It thus remains for future investigations to determine whether corrections of the style presented here are able to be, and should be, incorporated into these extensions.

Appendix A: Partial autocorrelation and active information storage

The partial autocorrelation function conveys important information regarding the dependence structure of an autoregressive process. 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 $\boldsymbol{Z}^{(u-1)}(t) = [Z(t-1)\,;\,\ldots\,;\,Z(t-u+1)]$. Using notation from the main text, we denote Z^u as the process of Z lagged by u time steps and $Z^{(u)}$ as the history up until that lag (inclusive) $\boldsymbol{Z}^{(u)} = [Z^1\,;\,\ldots\,;\,Z^u]$. Then, for a stationary time series, the partial autocorrelation function is defined by [6]

$$\alpha_Z(1) = \rho_{ZZ^1}, \tag{A1}$$

and

$$\alpha_Z(u) = \rho_{ZZ^u.\boldsymbol{Z}^{(u-1)}}, \quad u > 1. \tag{A2}$$

In this work, we employed the common practice of using the partial autocorrelation function to identify the relevant history length p for an autoregressive model of Z, since $\alpha_Z(u) = 0$ for $u > p$ [6, 7, 39]. Again, this is a statistical estimate and thus we must infer the order p by testing each sample partial autocorrelation $\hat{\alpha}_Z(u)$ for significance against the null distribution. To simplify inference, we z-transformed the estimate $f = \tanh^{-1}(\hat{\alpha}_Z(u))$ and considered it statistically insignificant when $|f| < 1.96\,(T-u)^{-1/2}$.

Intriguingly, our work reveals a relationship between the partial autocorrelation function and active information storage [34]—a recently developed model-free measure for quantifying memory in a process—under the linear Gaussian assumption. Active information storage $\mathcal{A}_X$ quantifies the information storage component of a process that is directly in use in the computation of the next value of a sequence. More precisely, active information storage is the mutual information between the (semi-infinite) past state of the process $\boldsymbol{X}^{(p)}(t)$ and its next value $X(t)$, i.e.,

$$\mathcal{A}_X(p) = \mathcal{I}_{X;\boldsymbol{X}^{(p)}}. \tag{A3}$$

Since the active information storage can be computed from multivariate mutual information, we can use Eq. (19) to decompose it into a sum of squared partial autocorrelations:

$$\begin{aligned}\mathcal{I}_{X;\boldsymbol{X}^{(p)}} &= -1/2 \sum_{u=1}^{p} \log\left(1 - \rho^2_{XX^u.\boldsymbol{X}^{(u-1)}}\right), \\ &= -1/2 \sum_{u=1}^{p} \log\left(1 - [\alpha_X(u)]^2\right).\end{aligned} \tag{A4}$$

This same logic can be straightforwardly applied to other measures such as excess entropy [32] and predictive information [33].

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 [31]. This is typically achieved

by using the LR test procedure to infer the significance of increasing the embedding lengths p. For linear-Gaussian processes, 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 approximately equivalent to squared partial autocorrelation up to a factor of two. This can be seen from Eq. (A4):

$$\begin{aligned}\delta_X(u) &= \mathcal{A}_X(u+1) - \mathcal{A}_X(u) \\ &= -\frac{1}{2}\log\left(1 - [\alpha_X(u+1)]^2\right) \\ &\approx \frac{1}{2}[\alpha_X(u+1)]^2, \end{aligned} \tag{A5}$$

through a Taylor expansion. In contrast to measures of dependence between multiple processes, the LR hypothesis test is valid, even without Bartlett-correcting the distributions, for testing any $\delta_X(u)$ for $u > p$. This is because, after the full set of past variables is included in the regression, any higher order residuals $X^u - \bar{x}^u(\bar{x}^{(p)})$ with $u > p$ have zero expected autocorrelation for every lag. Equivalently, the residuals on the next value $X - \bar{x}(\bar{x}^{(p)})$ have zero expected autocorrelation. Thus, no Bartlett correction is required, however the conditionals should be included, such that the LR test should be done with a modified effective sample size of $T - p$.

Appendix B: Considerations and extensions of Bartlett's formula

In our derivations we use one of the earliest versions of Bartlett's formula, as described in [9]. However, since his seminal work, there have been a number of 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 [12], who provided the large-sample distribution between pairs of sample cross correlations at differing lags. Consider the four processes Z_a, Z_b, Z_d, Z_e. By letting

$$\Delta_v(a,b,d,e) = \sum_{u=-\infty}^{\infty} \rho_{ab}(u)\rho_{de}(u+v), \tag{B1}$$

and

$$s_{ab}(v) = T^{1/2}\left[r_{ab}(v) - \rho_{ab}(v)\right], \tag{B2}$$

the asymptotic distribution of the standard error $s_{ab}(v)$ is Gaussian with zero mean and covariance

$$\begin{aligned}&\lim_{T\to\infty} \mathrm{Cov}(s_{ab}(v), s_{de}(w)) \\ &\quad\approx \Delta_{w-v}(a,d,b,e) + \Delta_{w+v}(b,d,a,e) \\ &\qquad - \rho_{ab}(v)\big[\Delta_w(a,d,a,e) + \Delta_w(b,d,b,e)\big] \\ &\qquad - \rho_{de}(w)\big[\Delta_v(b,d,a,d) + \Delta_v(b,e,a,e)\big] \\ &\qquad + \tfrac{1}{2}\rho_{ab}(w)\rho_{de}(w)\big[\Delta_0(a,d,a,d) + \Delta_0(a,e,a,e) \\ &\qquad\qquad + \Delta_0(b,d,b,d) + \Delta_0(b,e,b,e)\big] \end{aligned} \tag{B3}$$

This derivation can be further generalised to the non-Gaussian case, for instance by allowing for skewed distributions [68]. As discussed below, the full covariance structure (B3) facilitates testing against an alternative hypothesis (of non-zero corrrelation), and could further help in situations when the partial correlation terms are themselves correlated.

First, however, 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 (B3) by setting $d = a$ and $e = b$, reducing to the results reported in [69] and [6] (Theorem 11.2.3). Using this special case, the sampling distribution of Pearson (zero-lag) correlation between two univariate processes $\mathrm{Var}\left(s_{ab}(0)\right)$ can be obtained by setting $v = w = 0$. Finally, under the null hypothesis of $\rho_{ab}(0) = 0$, most of these terms disappear and we are left with Bartlett's original formula:

$$\lim_{T\to\infty} \mathrm{Var}\left(s_{ab}(0)\right) = \lim_{T\to\infty} \mathrm{Var}\left(T^{1/2}\left[r_{ab}(0)\right]\right).$$

This can be alternatively represented as an approximation for a large-sample size:

$$\mathrm{Var}\left(s_{ab}(0)\right) \approx \sum_{u=-\infty}^{\infty} \rho_{aa}(u)\rho_{bb}(u), \tag{B4}$$

which was the form used in the main text of this paper. For the exact number of samples to constitute a large-sample dataset, we refer the reader to the discussions in [9–11]. Bartlett did present a formula irrespective of sample size in [9], which may yield an improvement for small sample distributions and give minor practical advantages (e.g., reducing the bias slightly in Fig. 2b), however, we did not find this necessary for practical purposes here and instead follow the approximations given in [10, 11].

Under the alternative hypothesis of $\rho_{ab}(0) \neq 0$, we obtain the finite-sample approximation discussed in [2]. That is, we can obtain an alternative hypothesis test for the dependence measures discussed in this paper. Specifically, by inferring the effective sample size (38) with the variance $\mathrm{Var}\left(r_{ij}\right)$ that is used in obtained from (B3) (instead of (34)), we can derive the exact sampling distribution for the dependence measures under the alternative.

Previous work has followed similar logic for obtaining the alternative hypothesis for correlation coefficients [2]. A challenge in this approach is that computing the exact distribution correlation coefficients is non-trivial, and so it is often more common to use Fisher's z-transformation to approximate the coefficient as a normal distribution. This approach may also be suitable for deriving alternative distributions for dependence measures, however the effect of a short sample size on Fisher's approximation may present issues in applications with short datasets and high autocorrelation such as neuroimaging.

1. Correlated residuals

In our derivation of the sampling distribution of multivariate mutual information (52) and much of the discussion above, we have assumed that each term is independent. This logic then follows to all other distributions by virtue of the chain rule. To expand on this point, let Eq. (52) be expressed as a sum of kl random variables L_i, i.e.,

$$\hat{\mathcal{I}}_{\boldsymbol{x};\boldsymbol{y}|\boldsymbol{w}} = -\frac{1}{2}\sum_{g=1}^{k}\sum_{h=1}^{l}\log\left(1 - r^2_{x_g y_h \cdot \boldsymbol{v}^{\{gh\}}_{\boldsymbol{xy}|\boldsymbol{w}}}\right) = \sum_{i=1}^{kl} L_i. \tag{B5}$$

In general, the variance of the sum of kl random variables $\{L_1, \ldots, L_{kl}\}$ is

$$\begin{aligned}\mathrm{Var}\left(\sum_{i=1}^{kl} L_i\right) &= \sum_{i,j=1}^{kl}\mathrm{Cov}(L_i, L_j) \\ &= \sum_{i=1}^{kl}\mathrm{Var}\left(L_i\right) + \sum_{i\neq j}\mathrm{Cov}(L_i, L_j).\end{aligned} \tag{B6}$$

If the random variables L_i in (B5) are uncorrelated, then the second term $\mathrm{Cov}(L_i, L_j) = 0, \forall (i \neq j)$. In this paper, we derived the variance for each L_i as a log F-distribution in (57). Since they are all independent, we can then obtain the sampling distribution of $\hat{\mathcal{I}}_{\boldsymbol{x};\boldsymbol{y}|\boldsymbol{w}}$ as a sum of all these log F-distributions. However, if the random variables were correlated, then we must use the multivariate form of Bartlett's formula (B3), which provides the covariance of each variable L_i term with all other variables L_j. In expectation, we expect this to be unnecessary, since there is theoretically no covariance between each term. However, in practice, the residuals are computed iteratively and thus each have the same standard error in them, which can induce spurious correlation between residual processes. The fact that this is not accounted for in our work might be the source of some small biases that we see in, e.g., Fig. 3 and 6. However, accounting for this covariance would require knowing the distribution of a sum of correlated log F-distributions that, to the best of our knowledge, is not an established result.

2. Regularising autocorrelation estimates

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 regularise the autocorrelation samples to better estimate their true value [2, 41]. 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

$$\mathrm{Var}\left(s_{ab}(0)\right) \approx 1 + 2\sum_{u=1}^{U}\lambda(u) r_{aa}(u) r_{bb}(u), \tag{B7}$$

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 [41] for details). Numerous truncation points U have also been proposed, e.g., $N/4$, $N/5$, $\sqrt{N}$, and $2\sqrt{N}$ [2]. We tested a number of these techniques to verify whether or not they had an effect on our results.

3. Comparing these distributions

In this above few sections we outlined a number of potential factors that could introduce small biases in our sampling distributions. To test our approach against these the more complex extensions, we compared the FPR of hypothesis tests that compute the effective sample size based on the full covariance matrix (B3), both with and without tapering. This was performed for each of the experiments performed in the paper. As mentioned in Sec. B 1, the sum of correlated log F-distributions is unknown, so we instead use sums of partial correlations, rather than sums of squared partial correlations. Obtaining this distribution is tractable since the sum of Gaussian distributed random variables is another Gaussian variable with variance from (B6). In performing these tests, we found no notable difference in the FPR for any of the validation experiments, and thus conclude that Eq. (34) (the formula used in the main text) is suitable in practice.

ACKNOWLEDGMENTS

JTL was supported through the Australian Research Council DECRA grant DE160100630. JMS was supported through a University of Sydney Robinson Fellowship and NHMRC Project Grant 1156536. JMS and JL were supported through The University of Sydney Research Accelerator (SOAR) Fellowship program. JL and JMS were supported through The Centre for Translational Data Science at The University of Sydney's Research Incubator Funding Scheme. High performance

computing facilities provided by The University of Sydney (Artemis) have contributed to the research results reported within this paper.

[1] C. E. Davey, D. B. Grayden, G. F. Egan, and L. A. Johnston, NeuroImage **64**, 728 (2013).
[2] S. Afyouni, S. M. Smith, and T. E. Nichols, NeuroImage (2019).
[3] E. Florin, J. Gross, J. Pfeifer, G. R. Fink, and L. Timmermann, NeuroImage **50**, 577 (2010).
[4] L. Barnett and A. K. Seth, Journal of Neuroscience Methods **201**, 404 (2011).
[5] A. K. Seth, Journal of Neuroscience Methods **186**, 262 (2010).
[6] P. J. Brockwell, R. A. Davis, and S. E. Fienberg, *Time Series: Theory and Methods: Theory and Methods* (Springer Science & Business Media, 1991).
[7] G. C. Reinsel, *Elements of Multivariate Time Series analysis* (Springer Science & Business Media, 2003).
[8] M. Bartlett, Journal of the Royal Statistical Society **98**, 536 (1935).
[9] M. S. Bartlett, Supplement to the Journal of the Royal Statistical Society **8**, 27 (1946).
[10] M. H. Quenouille, Biometrika **34**, 365 (1947).
[11] G. Bayley and J. Hammersley, Supplement to the Journal of the Royal Statistical Society **8**, 184 (1946).
[12] R. Roy, Biometrika **76**, 824 (1989).
[13] 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.*, NeuroImage **62**, 2222 (2012).
[14] J. R. Bence, Ecology **76**, 628 (1995).
[15] M. Macias-Fauria, A. Grinsted, S. Helama, and J. Holopainen, Dendrochronologia **30**, 179 (2012).
[16] K. J. Friston, Brain Connectivity **1**, 13 (2011).
[17] C. W. Granger, Econometrica: Journal of the Econometric Society , 424 (1969).
[18] A. Roebroeck, E. Formisano, and R. Goebel, NeuroImage **25**, 230 (2005).
[19] W. Liao, J. Ding, D. Marinazzo, Q. Xu, Z. Wang, C. Yuan, Z. Zhang, G. Lu, and H. Chen, NeuroImage **54**, 2683 (2011).
[20] K. Friston, R. Moran, and A. K. Seth, Current Opinion in Neurobiology **23**, 172 (2013).
[21] R. K. Kaufmann and D. I. Stern, Nature **388**, 39 (1997).
[22] D. D. Zhang, H. F. Lee, C. Wang, B. Li, Q. Pei, J. Zhang, and Y. An, Proceedings of the National Academy of Sciences **108**, 17296 (2011).
[23] J. R. Freeman, American Journal of Political Science **27**, 327 (1983).
[24] R. Reuveny and H. Kang, American Journal of Political Science **40**, 943 (1996).
[25] H. Hotelling, Biometrika **28**, 321 (1936).
[26] P. M. Robinson, Journal of Multivariate Analysis **3**, 141 (1973).
[27] D. J. MacKay, *Information theory, inference and learning algorithms* (Cambridge University Press, 2003).
[28] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, 2012).
[29] J. Theiler, Physical Review A **34**, 2427 (1986).
[30] J. Kay, in *Proc. of IEEE IJCNN*, Vol. 4 (1992) pp. 79–84.
[31] J. T. Lizier, Frontiers in Robotics and AI **1**, 11 (2014).
[32] W. Bialek, I. Nemenman, and N. Tishby, Physica A: Statistical Mechanics and its Applications **302**, 89 (2001).
[33] J. P. Crutchfield and D. P. Feldman, Chaos: An Interdisciplinary Journal of Nonlinear Science **13**, 25 (2003).
[34] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, Information Sciences **208**, 39 (2012).
[35] We mainly refer to the LR hypothesis test in this paper. However, the issue of autocorrelation also applies to another common approach of using a standard F-test; see [4, 5] for details.
[36] The F-test used in [4] does account for this particular effect; however, it is not extended to testing a multivariate predictee nor does it account for autocorrelation.
[37] G. C. Tiao and G. E. Box, Journal of the American Statistical Association **76**, 802 (1981).
[38] H. Boudjellaba, J.-M. Dufour, and R. Roy, Journal of the American Statistical Association **87**, 1082 (1992).
[39] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control* (John Wiley & Sons, 2015).
[40] J. Geweke, Journal of the American Statistical Association **77**, 304 (1982).
[41] C. Chatfield, *The Analysis of Time Series: An Introduction* (Chapman and Hall/CRC, 2003).
[42] L. Barnett, A. B. Barrett, and A. K. Seth, Physical Review Letters **103**, 238701 (2009).
[43] J. T. Lizier and M. Prokopenko, The European Physical Journal B **73**, 605 (2010).
[44] F. Takens, in *Dynamical Systems and Turbulence*, Lecture Notes in Mathematics, Vol. 898, edited by D. Rand and L.-S. Young (Springer, Berlin / Heidelberg, 1981) pp. 366–381.
[45] O. M. Cliff, M. Prokopenko, and R. Fitch, Frontiers in Robotics and AI **3**, 71 (2016).
[46] O. M. Cliff, M. Prokopenko, and R. Fitch, Entropy **20**, 51 (2018).
[47] L. Faes, G. Nollo, and A. Porta, Physical Review E **83**, 051112 (2011).
[48] L. Novelli, P. Wollstadt, P. Mediano, M. Wibral, and J. T. Lizier, Network Neuroscience **3**, 827 (2019).
[49] L. Barnett and A. K. Seth, Journal of Neuroscience Methods **223**, 50 (2014).
[50] T. Schreiber, Physical Review Letters **85**, 461 (2000).
[51] T. Bossomaier, L. Barnett, M. Harré, and J. T. Lizier, *An Introduction to Transfer Entropy: Information Flow in Complex Systems* (Springer, 2016).
[52] O. M. Cliff, J. T. Lizier, X. R. Wang, P. Wang, O. Obst, and M. Prokopenko, Artificial Life **23**, 34 (2017).
[53] J. T. Lizier, J. Heinzle, A. Horstmann, J.-D. Haynes, and M. Prokopenko, Journal of Computational Neuroscience **30**, 85 (2011).
[54] N. R. Draper and H. Smith, *Applied Regression Analysis*, Vol. 326 (John Wiley & Sons, 1998).
[55] T. L. Kelley, *Crossroads in the Mind of Man: A Study of Differentiable Mental Abilities* (Stanford University Press, 1928).

[56] C. E. Davey, D. B. Grayden, M. Gavrilescu, G. F. Egan, and L. A. Johnston, Human Brain Mapping **34**, 1999 (2013).
[57] See Appendix B for a differentiation of the large-sample from asymptotic results on Bartlett's formula.
[58] J. Wishart, Biometrika **20A**, 32 (1928).
[59] D. R. Brillinger, Brazilian Journal of Probability and Statistics **18**, 163 (2004).
[60] M. J. Anderson and J. Robinson, Australian & New Zealand Journal of Statistics **43**, 75 (2001).
[61] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich, NeuroImage **54**, 875 (2011).
[62] R. Franciotti, N. Falasca, L. Bonanni, F. Anzellotti, V. Maruotti, S. Comani, A. Thomas, A. Tartaro, J. Taylor, and O. M, Neurobiology of Aging **34**, 1148 (2013).
[63] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis*, Vol. 7 (Cambridge University Press, 2004).
[64] I. Weber, E. Florin, M. Von Papen, and L. Timmermann, PloS One **12** (2017).
[65] C. Francq and J.-M. Zakoïan, Journal of Time Series Analysis **30**, 449 (2009).
[66] L. Barnett and A. K. Seth, Physical Review E **91**, 040101 (2015).
[67] A. Gutknecht and L. Barnett, arXiv preprint arXiv:1911.09625 (2019).
[68] N. Su and R. Lund, Journal of Multivariate Analysis **105**, 18 (2012).
[69] M. Bartlett, *An Introduction to Stochastic Processes* (Cambridge University Press, 1935).