Rigorous and flexible statistical tests for correlations between stationary or nonstationary time series

Alexander Yuan

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Wenying Shou, Chair

Sean Gibbons

Benjamin Kerr

Program Authorized to Offer Degree:

Molecular and Cellular Biology

University of Washington

Abstract

Rigorous and flexible statistical tests for correlations between stationary or nonstationary time series

Alexander Yuan

Chair of Supervisory Committee:

Wenying Shou

Department of Biology

A growing body of life-sciences research seeks to infer correlational and causal relationships from time series. Yet, these analyses can encounter challenges in practice, and this dissertation focuses on three such challenges: First, many popular statistical approaches for correlational and causal analysis of time series come with assumptions and caveats that are easy to overlook. Second, time series typically exhibit autocorrelation, which violates the fundamental assumptions of many standard statistical tests. Third, researchers are increasingly using correlation statistics that lack a known analytical null distribution and which therefore lack a standard parametric test, thus requiring the development of appropriate nonparametric methods. To address the first challenge (overlooked assumptions), Part I uses a multimedia strategy (including video) to illustrate key concepts and caveats of three popular statistical approaches that are used to make causal claims. Although primarily an interdisciplinary synthesis, Part I also describes some novel pathologies of existing methods. The later parts report methodological advances targeted at the second (autocorrelation) and third (nonparametric correlation) challenges outlined above. Both parts describe new tests that are statistically valid (meaning that they can guarantee a false positive rate that does not surpass a user-defined "significance level"). Part II reports a significance test that is applicable to any pairwise correlation statistic and which is valid as long as one of the two time series is stationary (i.e. behavior does not change systematically over time). As a demonstration, the test is used to detect known statistical dependence relationships in disciplines from microbiome science to climatology. Part III tackles the difficult setting of nonstationary

time series, for which multiple biological replicates are often needed. In this context, I describe a valid significance test for correlation between time series that enables detections with higher confidence and fewer replicates than similar approaches. The test is used to verify the previously observed relationship between swimming speed and directional alignment in zebrafish, using a publicly available data set with only 3 replicates of likely-nonstationary time series. These efforts seek to empower scientists to use current and future data-analytic approaches without sacrificing the benefits of statistical rigor.

Acknowledgement

I am indebted to Dr. Wenying Shou for support, training, and endless encouragement. This work would not be possible without her. During my time with the Shou Lab, nearly all of the members of the group provided some form of feedback on the presentation of the ideas. Without their suggestions, this text would be worse. In particular, Dr. Li Xie checked a number of the formal mathematical arguments, often pointing out where writing was unclear. I am grateful to my committee (which, apart from Dr. Shou, includes Drs. Sean Gibbons, Sarah Holte, Ben Kerr, and Nathan Kutz) for their valuable guidance and technical suggestions. Moreover, this process would be a far lonelier task without my fiancée Nam Phuong, who as a current PhD student herself, has held my hand during the moments of frustration and joy that one encounters in graduate school.

Contents

Part I

An overview of data-driven causal analysis for observational biological time series

This part was published in 2022 in eLife with the title "Data-driven causal analysis of observational biological time series. Simulation results can be reproduced using the scripts included in the online article."

Abstract

Complex systems are challenging to understand, especially when they defy manipulative experiments for practical or ethical reasons. Several fields have developed parallel approaches to infer causal relations from observational time series. Yet these methods are easy to misunderstand and often controversial. Here, we provide an accessible and critical review of three statistical causal discovery approaches (pairwise correlation, Granger causality, and state space reconstruction), using examples inspired by ecological processes. For each approach, we ask what it tests for, what causal statement it might imply, and when it could lead us astray. We devise new ways of visualizing key concepts, describe some novel pathologies of existing methods, and point out how so-called "model-free" causality tests are not assumption-free. We hope that our synthesis will facilitate thoughtful application of causal methods, encourage communication across different fields, and encourage explicit statements of assumptions. A video walkthrough is available at https://youtu.be/AlV0ttQrjK8.

1 Introduction

Ecological communities perform important activities, from facilitating digestion in the human gut to driving the biogeochemical cycles of elements on the earth. Communities are often highly complex, with many species engaging in diverse interactions. To control communities, it helps to know causal relationships between variables (e.g. whether perturbing the abundance of one species might alter the abundance of another species). We can express these relationships either explicitly by proposing causal networks [1, 2, 3, 4, 5, 6, 7], or implicitly by simply predicting the effects of new perturbations [8, 9].

Ideally, biologists discover such causal relations from manipulative experiments. However, manipulative experiments can be infeasible or inappropriate: Natural ecosystems may not offer enough replicates for

comprehensive manipulative experiments, and perturbations can be impractical at large scales and may have unanticipated negative consequences. On the other hand, there exists an ever-growing abundance of observational time series (i.e. without intentional perturbations). The goal of obtaining accurate causal predictions from these or similar data sets has motivated several complementary lines of investigation.

Determining causal relationships can become more straightforward if one already knows, or is willing to assume, a model that captures key aspects of the underlying process. For example, the Lotka-Volterra model popular in mathematical ecology assumes that species interact in a pairwise fashion, that the fitness effects from different interactions are additive, and that all pairwise interactions can be represented by a single equation form where parameters can vary to reflect signs and strengths of fitness effects. By fitting such a model to time series of species abundances and environmental factors, one can predict, for instance, which species interact or how a community might respond to certain perturbations [10, 11, 12]. However, the Lotka-Volterra equations often fail to describe complex ecosystems and chemically-mediated interactions [13, 14, 15].

When our understanding is insufficient to support knowledge-based modeling, how might we formulate causal hypotheses? A large and rapidly growing literature attempts to infer causal relations from time series data without using a mechanistic model. Such methods are sometimes called "model-free" [16], although they typically rely on *statistical* models. Some of these methods avoid any equation-based description of the dynamics and instead examine some kind of "information flow" between time series [17, 18]. Others deploy highly flexible equations that are not necessarily mechanistic [19, 20].

Here we focus on three model-free approaches that have been commonly used to make causal claims in ecology research: pairwise correlation, Granger causality, and state space reconstruction. For each, we ask (1) what information does the method give us, (2) what causal statement might that information imply, and (3) when might the method lead us astray?

We found that answering these seemingly basic questions was at first surprisingly challenging for several reasons. First, modern causal discovery approaches have intellectual roots in several communities including philosophy, statistics, econometrics, and chaos theory, which sometimes use different words for the same idea, and the same word for different ideas. The word causality itself is a prime example: Many philosophers (and scientists) would say that X causes Y if an intervention upon X would result in a change in Y [21, 22]. Granger's original works instead defined causality to be about how important the history of X is in predicting Y [19, 17], and in the nonlinear dynamics field, causality is sometimes used to mean that the trajectories of X and Y have certain shared geometric or topological properties [23]. Such language, while unproblematic when confined to a single community, can nevertheless obscure important differences between methods from different communities. A second challenge is that in methodological articles, key assumptions are sometimes

hidden in algorithmic details, or simply not mentioned. Finally, some methods deal with nuanced or advanced mathematical ideas that can be difficult even for those with quantitative training. Given these challenges, it is no surprise that efforts to infer causal relationships from observational time series have sometimes been highly controversial, with an abundance of "letters to the editor", sometimes followed by impassioned dialogue [24, 25, 26, 27, 28].

We have tried to balance precision and readability in this review. To accomplish this, we devised new ways to visualize key concepts. We also provide refreshers and discussions of mathematical notions in the Appendices. Lastly, we compare all methods to a common definition of causality that is useful to experimental scientists. Our goals are to inform practitioners who wish to statistically test causal hypotheses using temporal data, to facilitate communication across different fields, and to encourage explicit statements of methodological assumptions and caveats. For a broad overview of time series causal methods in Earth sciences or more technical reviews, see [5] and [29, 30] respectively.

2 Dependence, correlation, and causality

Causality

We use the definition of "causality" that is common in statistics and intuitive to scientists: X has a causal effect on Y ("X causes Y" or "X is a causer; Y is a causee" or "X is a cause; Y is an effect") if some externally applied perturbation of X can result in a perturbation in Y (Figure 1A). We say that X and Y are *causally related* if X causes Y, Y causes X, or some other variable ("common cause", "confounder") causes both. Otherwise, X and Y are *causally unrelated*. Additionally, one can talk about direct versus indirect causality (Figure 1B; see legend for definitions). A surprising result from past several decades of causality research is that there are in fact some conditions under which directional causal structures can be correctly inferred ("identified") from purely observational data [1, 29, 31] (e.g. Figure A13, last row). However, empirical time series often do not contain enough information for easy causal identifiability [1, 3].

Correlation versus dependence

The adage "correlation is not causality" is well-known to the point of being cliché [18, 33, 34, 35]. Yet, to dismiss correlative evidence altogether seems too extreme. To make use of correlative evidence without being reckless, it helps to distinguish between the terms "correlation" and "dependence". When applied to ecological time series, the term "correlation" is often used to describe some statistic that quantifies the similarity between two observed time series [36, 33]. Examples include Pearson's correlation coefficient and

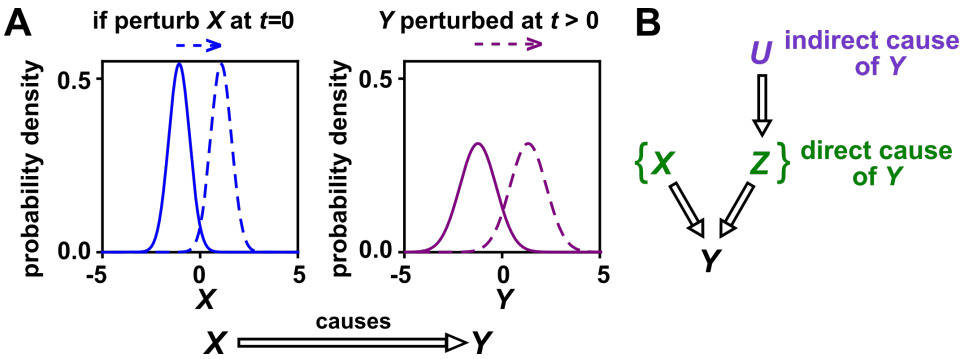


Figure 1: Causality. (**A**) Definition. If a perturbation in X can result in a change in future values of Y, then X causes Y. This definition does not require that *any* perturbation in X will perturb Y. For example, if the effect of X on Y has saturated, then a further increase in X will not affect Y. In this article, causality is represented by a hollow arrow. To embody probabilistic thinking (e.g. drunk driving increases the chance of accidents) [22], X and Y are depicted as histograms. Sometimes, perturbations in one variable can change the current value of another variable if, for example, the two variables are linked by a conservation law (e.g. conservation of energy). Some have argued that these are also causal relationships [21]. (**B**) Direct versus indirect causality. The direct causers of Y are given by the minimal set of variables such that once the entire set is fixed, no other variables can cause Y. Here, three variables X, Z, and U activate Y. The set $\{X, Z\}$ constitutes the direct causers of Y (or Y's parents [32, 22]), since if we fix both X and Z, then Y becomes independent of U. If a causer is not direct, we say that it is indirect. Whether a causer is direct or indirect can depend on the scope of included variables. For example, suppose that yeast releases acetate, and acetate inhibits the growth of bacteria. If acetate is not in our scope, then yeast density has a direct causal effect on bacterial density. Conversely, if acetate is included in our scope, then acetate (but not yeast) is the direct causer of bacterial density (since fixing acetate concentration would fix bacterial growth regardless of yeast density). When we draw interaction networks with more than two variables, hollow arrows between variables denote direct causation.

local similarity [37]. In contrast, statistical dependence is a hypothesis about the probability distributions that produced those time series, and has close connections to causality.

Dependence has a precise definition in statistics, and is most easily described for two binary events. For instance, if the incidence of vision loss is higher among diabetics than among the general population, then vision loss and diabetes are statistically dependent. In general, events A and B are dependent if across many independent trials (e.g. patients), the probability that A occurs given that B has occurred (e.g. incidence of vision loss among diabetics only) is different from the background probability that A occurs (e.g. background incidence of vision loss). If A and B are not dependent, then they are called independent. The concept of dependence is readily generalized from binary events to numerical variables, and also to vectors such as time series (Appendix A1).

Dependence is connected to causation by the widely accepted "Common Cause Principle": *if two variables are dependent, then they are causally related (one causes the other, or both share a common cause)* [29, 5, 31, 38]. Note however that if one mistakenly introduces selection bias, then two independent variables can appear to be dependent (Figure A14). The closely related property of conditional dependence (see Appendix A1), i.e. whether two variables are dependent after statistically controlling for ("conditioning on") certain other variables, can be even more causally informative. In fact, when conditional dependence (and conditional independence) relationships are known, it is sometimes possible to infer most or all of the direct causal relationships at play, even without manipulative experiments or temporal information. Many of the algorithms that accomplish this rely on two technical but often reasonable assumptions: the "causal Markov condition", which allows one to infer causal information from conditional *dependence*, and the "causal faithfulness condition", which allows one to infer causal information from conditional *independence* ([29, 3, 31]; Appendix A2).

In sum, whereas a correlation is a statistical description of data, statistical dependence is a hypothesis about the relationship between the underlying probability distributions. Dependence is in turn linked to causality. Below, we discuss how to use correlation to detect dependence in time series.

Testing for dependence between time series using surrogate data

Despite its scientific usefulness, dependence between time series can be treacherous to test for. This is because time series are often autocorrelated (e.g. what occurs today influences what occurs tomorrow), so that a single pair of time series contains information from only a single trial. If one has many trials that are independent and free of systematic differences (e.g. ≥ 20 as in some laboratory microcosm experiments), the task is relatively easy: One can test whether species X and Y are statistically dependent by comparing the

correlation between X and Y abundance series from the same trial with those between X and Y abundance series from different trials (Figure A11; see also [39]). However, a large trial number is generally a luxury and often only one trial is available. In such cases, attempting to discern whether two time series are statistically dependent is like attempting to divine whether diabetes and vision loss are dependent with only a single patient (i.e. we have an "n-of-one problem"). As one possible remedy, there are parametric tests using the Pearson correlation coefficient that account for autocorrelation. In these tests, one estimates the correlation coefficient between time series, and evaluates its statistical significance using the variance of the null distribution [40]. However, the calculation of this variance relies on estimates of the autocorrelation at each lag for both time series, which can be highly uncertain [41, 42]. Furthermore, even if the variance can be accurately estimated, the shape of the null distribution still needs to be assumed before one can assign a p-value to the correlation.

Alternatively, the n-of-one problem is often addressed by a technique called surrogate data testing. Specifically, one computes some measure of correlation between two time series X and Y. Next, one uses a computer to simulate replicates of Y that might have been obtained if X and Y were independent (see below). Each simulated replicate is called a "surrogate" Y. Finally, one computes the correlation between X and each surrogate Y. A p-value (the probability of observing the original correlation or a more impressive one under the null hypothesis that X and Y are independent) is then determined by counting how many of the surrogate Ys produce a stronger correlation than the real Y. For example, if we produced 19 surrogates and found the real correlation to be stronger than all 19 surrogate correlations, then we would write down a p-value of $1/(1 + 19) = 0.05$. Ideally, if two time series are independent, then we should register a a p-value of 0.05 (or less) in only 5% of cases.

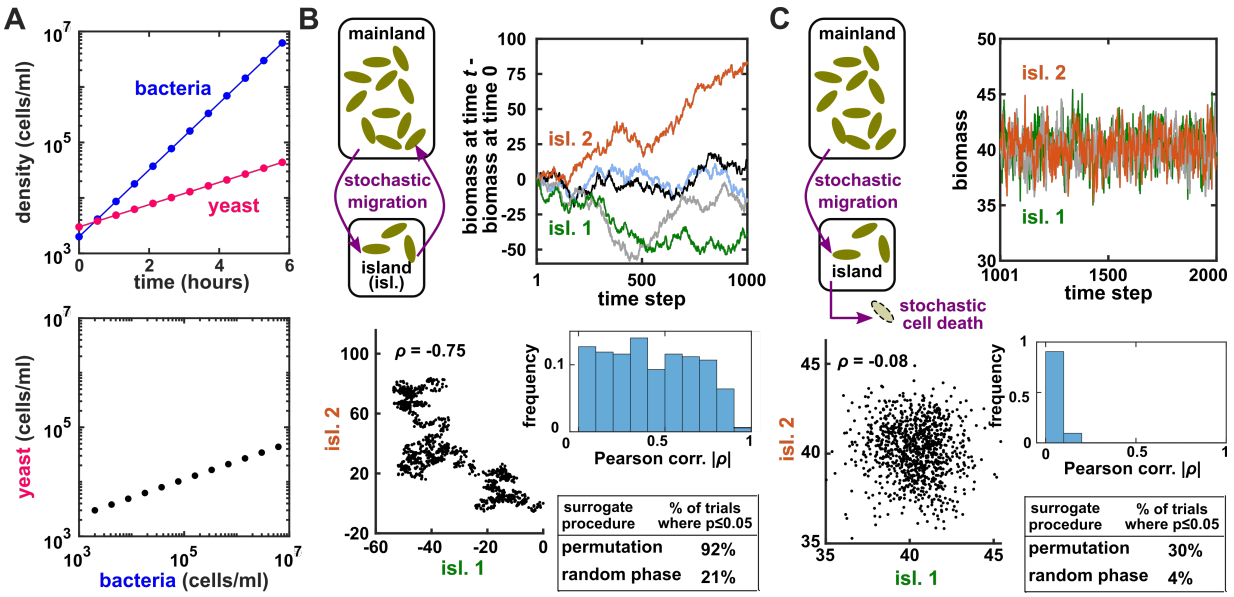


Figure 2: Two independent temporal processes can appear correlated, especially when compared to an inappropriate null model. (**A**) Densities of independent yeast and bacteria cultures growing exponentially are correlated. (**B**, **C**) Correlation between time series of two independent island populations can appear significant if inappropriate tests are used. (**B**) In an island ("isl"), individuals stochastically migrate to and from the mainland (so that total island biomass follows a random walk). At each time step, the net change in island biomass is drawn from a standard normal distribution (mean = 0; standard deviation = 1 biomass unit). (**C**) An island population receives cells through migration and loses cells to death. Observations are made after 1000 steps, so that the population size has reached an equilibrium. For both (B) and (C), we performed 1000 simulations in which we calculated the Pearson correlation coefficient of a pair of independent islands populations. Both panels contain: example time series (upper right), a scatterplot comparing two independent islands (lower left), the distribution of Pearson correlation coefficient strength (blue shading), and the proportion of simulations in which the correlation was deemed significant ($p \leq 0.05$) by surrogate data tests using either permutation or phase randomization (see main text). Ideally, the proportion of correlations that are significant (false positives) should not exceed 5%. The strength of correlation is weaker in (C) compared to (B), yet still often significant according to the permutation test. See Appendix A5 for more details.

Several procedures can used to produce surrogate time series, each corresponding to an assumption about how the original time series was generated [43]. One popular procedure is to simply shuffle the values of a time series [37, 44, 45, 46]. This procedure, often called permutation, assumes that all possible orderings of the time points in the series are equally likely. This assumption is commonly violated in time series due to autocorrelation, and thus the test is often invalid. For example, for independent time series in Figure 2B-C, this test returns $p < 0.05$ at rates of $30 \sim 92\%$, much higher than 5%. Nevertheless, permutation testing has appeared in many applied works, perhaps because it has been the default option in some popular software packages. Another procedure for generating surrogates is called phase randomization. It first uses the Fourier transform to represent a time series as a sum of sine waves, then randomly shifts each of the component sine waves in time, and finally sums the phase-shifted components [42, 47, 48] (Figure A16).

This procedure assumes that the original time series was obtained from a Gaussian, linear, and stationary process [48, 43], where "Gaussian" means that any subsequence follows a multivariate Gaussian distribution, "stationary" means that this distribution does not change over time, and "linear" means that future values depend linearly on past values and past random events ("process noise", Figure 7A). See [49] for a discussion of exact requirements. Indeed, this test performed well (with a false positive rate of 4%) when time series satisfied the three assumptions (Figure 2C), and poorly when the stationarity assumption was violated (with a false positive rate of 21%; Figure 2B). Other surrogate data procedures include time shifting [48], block bootstrap [50], and the twin method [51]. Some surrogate data tests have been shown to perform reasonably well even when exact theoretical requirements are unmet or unknown [51, 50], but a more comprehensive benchmarking effort is needed to map out each method's valid domain in practice.

In sum, surrogate data allow a researcher to use an observed correlation statistic to test for dependence under some assumption about the data-generating process. Dependence indicates the presence of a causal relationship, and conditional dependence can sometimes even indicate the direction [31, 3, 52] (Figure A13). Below we consider Granger causality and state space reconstruction, two approaches which can be used to directly infer the direction of causality from time series.

3 Granger causality: intuition, pitfalls, and implementations

Intuition and formal definitions

In simple language, X is said to Granger-cause Y if a collection of time series containing all historical measurements predicts Y's future behavior better than a similar collection that excludes the history of X. An important consequence of this definition is that Granger causality excludes indirect causes, as illustrated in Figure 3A. In practice, whether a causal relationship is direct or indirect depends on which variables are observed. For instance, in Figure 3A, if Y were not observed, then X would "directly" cause (and Granger-cause) Z.

Granger causality has many related but nonequivalent quantitative incarnations in the literature, including several that were proposed by Granger himself [19, 17]. Box 1 presents two definitions: one based on a linear regression which we call "linear Granger causality" [53, 54, 55, 35] and a second, more general, definition which we call "general Granger causality" (and which is also sometimes called nonlinear Granger causality) [17, 56, 57, 58, 59, 50]. See theorem 10.3 of [29] for a discussion of the theoretical relationship between general Granger causality and (true) causality.

Box 1: Granger causality

1. Linear Granger causality:

Under linear Granger causality, X Granger-causes Y if including the history of X in a linear autoregressive model (Eq 1) allows for a better prediction of future Y than not including the history of X (i.e. setting all α_k coefficients to zero). By "linear autoregressive model", we mean that the future value of variable Y is modeled as a linear combination of historical values of X and Y and all other observed variables that might help predict Y "$\cdots$":

$$Y_{t+1} = c + \sum_{k=0}^{n} (\alpha_k X_{t-k} + \beta_k Y_{t-k} + \cdots) + \varepsilon_t \tag{1}$$

Here, t is the time index, $k = 0, 1, ..., n$ is a time lag index, c is a constant, coefficients such as α_k and β_k represent the strength of contributions from the respective terms, and ε_t represents independent and identically-distributed (IID, A1) process noise (Figure 7A).

2. General Granger causality [17, 29]:

Let X_t, Y_t, and Z_t be series of random variables indexed by time t. X Granger-causes Y with respect to the information set $\{X_t, Y_t, Z_t\}$ if:

$$P(Y_t|\{X_k, Y_k, Z_k \text{ for all } k < t\}) \neq P(Y_t|\{Y_k, Z_k \text{ for all } k < t\}) \tag{2}$$

at one or more times t. Here, $P(Y_t|\mathcal{S})$ is the probability distribution of Y_t conditional on the variable set $\mathcal{S}$. Note that Z_k in Eq. 2 may include multiple variables and thus plays the same role as "$\cdots$" in Eq. 1.

Granger causality failure modes

We discuss four important instances where Granger causality can fail as an indicator of direct causality (Figure 3B). These pathologies can be understood intuitively and can apply to both linear and general Granger causality. First, if a system has deterministic dynamics (see Appendix A3), then Granger causality may fail to detect causal relations (Figure 3Bi). More generally, if dynamics have a low degree of randomness, Granger causality signals can be very weak (e.g. knowing X's past improves predictions of Y's future only slightly) [60, 29]. Moreover, as we will see, this limitation has motivated other methods that take a primarily deterministic view [18]. Second, Granger causality may erroneously assign a direct causal relation between a pair of variables that have an unobserved common cause (Figure 3Bii). Third, recording data at a frequency below that of the original process by "subsampling" (e.g. taking weekly measurements of a daily process) or by "temporal aggregation" (e.g. taking weekly averages of a daily process) can alter the inferred causal

structure (Figure 3Biii), although recent techniques can help with these issues [61, 62, 63]. Lastly, when measurements are noisy (Figure 3Biv), Granger causality can assign false interactions and also fail to detect true causality [64], although some progress has been made on this front [65].

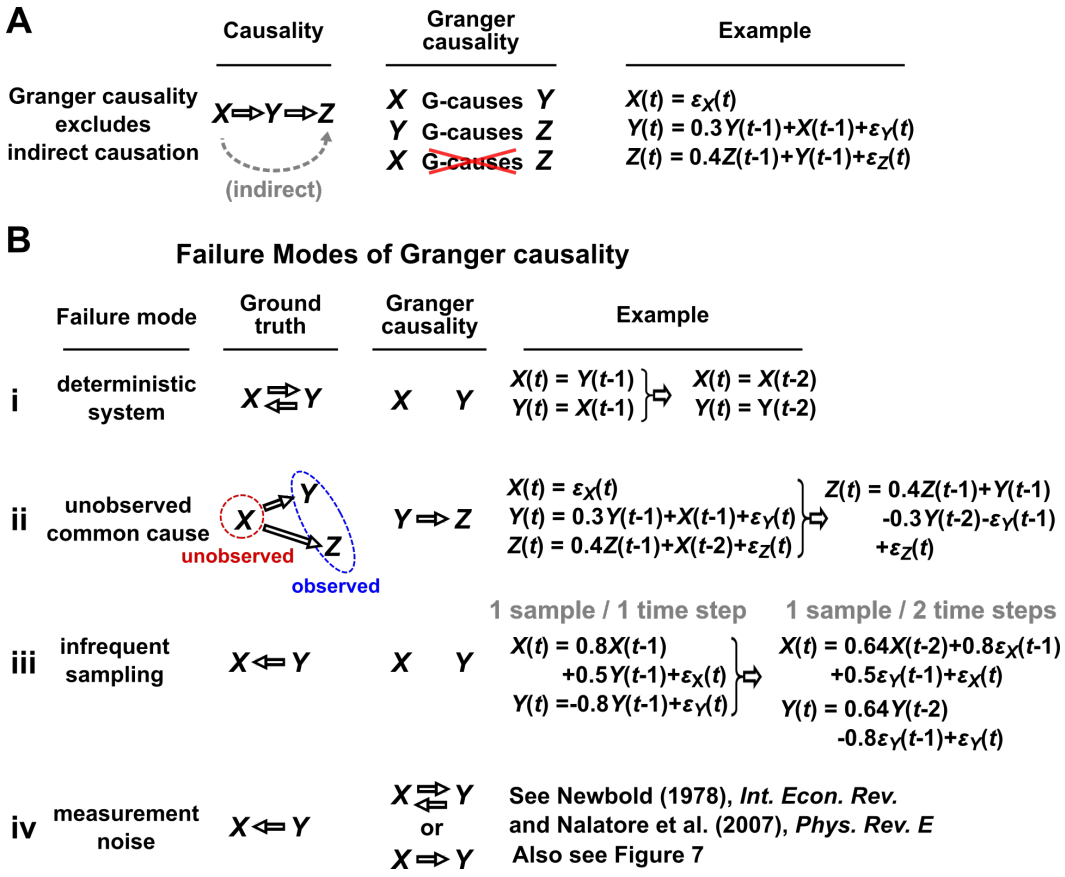


Figure 3: Causality versus Granger causality. (**A**) Granger causality is designed to reveal direct causes, not indirect causes. Although X causes Z, X does not Granger-cause Z because with the history of Y available, the history of X no longer adds value for predicting Z. This also shows that Granger causality is not transitive: X Granger-causes Y and Y Granger-causes Z, but X does not Granger-cause Z. (**B**) Failure modes of Granger causality when inferring direct causality. (**i**) False negative due to lack of stochasticity. X and Y mutually and deterministically cause one another through a copy operation [66, 29]: $X(t)$ copies $Y(t-1)$ and vice versa. Since $X(t-2)$ already contains sufficient information to know $X(t)$ exactly, the history of Y cannot improve prediction of X, and so Y does not Granger-cause X. By symmetry, X does not Granger-cause Y. (**ii**) False positive due to unobserved common cause. X causes Y with a delay of 1, and causes Z with a delay of 2. We only observe Y and Z. Since Y receives the same "information" before Z, the history of Y helps to predict Z, and thus Y Granger-causes Z, resulting in a false positive. (**iii**) Infrequent sampling can induce false negatives. Although there is a Granger causality signal when we sample once per time step, the signal is lost when we sample only once per 2 steps [61]. (**iv**) Measurement noise can lead Granger causality to suffer both false positives and false negatives. ϵ_X, ϵ_Y, and ϵ_Z represent process noise and are normal random variables with mean of 0 and variance of 1. All process noise terms are independent of one another.

Practical testing for linear and general Granger causality

One might still attempt to infer Granger causality despite the above caveats, especially in situations where they can be largely avoided. Linear Granger causality, popular in microbiome studies [53, 54, 55, 35], uses standard parametric tests: if any of the α_k terms in Eq. 1 is nonzero, then X linear Granger-causes Y. Parametric tests are computationally inexpensive and available in multiple free and well-documented software packages [67, 20]. These tests assume that time series are "covariance-stationary", which means that certain statistical properties of the series are time-independent [20] (see Appendix A3), and can fail when this assumption is violated [68, 69, 70]. Additionally, applying linear Granger causality to nonlinear systems can lead to incorrect causal conclusions [71]. One can assess whether the linear model (Eq. 1) is a reasonable approximation, for instance by checking whether the model residuals ε_t are uncorrelated across time [72] (as is assumed by Eq. 1).

Tests for general Granger causality often use a statistic known as transfer entropy [73]. Roughly, the transfer entropy from X to Y is the extent to which the entropy (a measurement of uncertainty) of Y's future is reduced when we account for (specifically, condition on) the past of X [74, 75, 76, 50]. A significant transfer entropy thus indicates the presence of general Granger causality. Surrogate data are typically used to evaluate significance [76, 50, 77]. However, the previously discussed surrogate data procedures are designed to test the null hypothesis of independence, which is different from the null hypothesis of general Granger non-causality (i.e. Eq. 2, but replace "$\neq$" with "$=$"). More recent surrogate procedures have been proposed to resolve this issue [78, 77]. Several software implementations of Granger causality tests based on transfer entropy statistics are available (e.g. [76, 79, 80]).

Granger causality methods face challenges when datasets have a large number of variables (e.g. in microbial ecology). In this case, the summation in Eq. 1 will contain a large number of terms, and so a regression procedure may fail to detect many true interactions [5, 4]. To handle systems with many variables, one can impose the assumption that only a small number of causal links exist [53, 35]. This is sometimes called sparse regression or regularization. Additionally, under certain technical assumptions, it is possible to use a series of logical rules to remove unnecessary terms in a purely data-driven way [4, 5]. As an example, suppose that we wish to test whether pH is a Granger-cause of chlorophyll concentration in some aquatic environment and we infer based on a prior analysis that chlorophyll concentration is always independent of fluctuations in salinity. Then, most likely, salinity is irrelevant to the pH-chlorophyll relationship and can be safely omitted from our Granger causality analysis. (However, this reasoning could theoretically fail if pH, salinity, and chlorophyll took on the respective roles of X, Y, and W in the 'cancellation' diagram of Figure A12.) These rules and their associated assumptions are formalized in "constraint-based causal

discovery" algorithms [29, 3] (Appendix A2). The development of new causal discovery algorithms, and their application to time series, is a very active area of research [4, 5, 6, 81].

4 State space reconstruction (SSR): intuition, pitfalls, and implementations

The term "state space reconstruction" (SSR) refers to a broad swath of techniques for prediction, inference, and estimation in time series analysis [82, 83, 84, 18, 85]. In this article, when we use the term SSR, we refer only to SSR methods for causality detection. The SSR approach is especially popular in empirical ecology [86, 87, 88, 89, 90]. SSR methods are intended to complement Granger causality: Whereas Granger causality has trouble with deterministic dynamics (Figure 3B), the SSR approach is explicitly designed for systems that are primarily deterministic [18]. Since SSR is less intuitive than correlation or Granger causality, we introduce it with an example rather than a definition.

Visualizing SSR causal discovery

Consider the deterministic dynamical system in Figure 4. Here, Z is causally driven by Y (and X), but not by W or V. We can make a vector out of the current value $Z(t)$ and two past values $Z(t-\tau)$ and $Z(t-2\tau)$ (Figure 4B, red dots), where τ is the time delay and $[Z(t), Z(t-\tau), Z(t-2\tau)]$ is called a "delay vector". The delay vector can be represented as a single point in the 3-dimensional Z "delay space" (Figure 4C, red dot). We then shade each point of the trajectory in Z delay space according to the contemporaneous value of its causer $Y(t)$. Since in this example each point of the trajectory in Z delay space corresponds to one and only one $Y(t)$ value, we call this a "delay map" from Z to Y. Notice that the $Y(t)$ gradient in this plot looks gradual in the sense that if two points are nearby in the delay space of Z, then their corresponding $Y(t)$ shades are also similar. This property is called "continuity" (Figure A20). We provide additional details on continuity in Appendix A4. Overall, there is a continuous map from the Z delay space to Y, or more concisely, a "continuous delay map" from Z to Y. A similar continuous delay map also exists from Z to its other causer X. On the other hand, if we shade the delay space of Z by W or V (neither of which causes Z), we do not get a continuous delay map (Figure 4D-E).

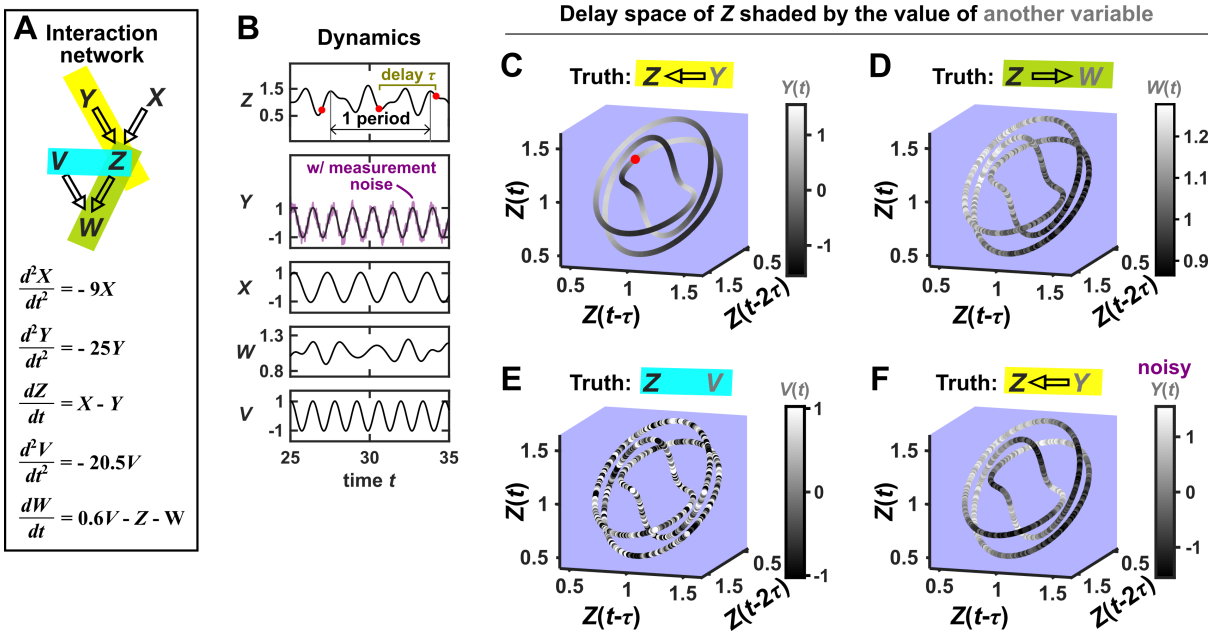


Figure 4: SSR methods look for a continuous map from the delay space of a causee to the causer, and becomes more difficult in the presence of noise. (**A**) A 5-variable toy (linear) system. Filled arrows and blunt head arrows represent activation and inhibition, respectively. (**B**) Time series. The delay vector $[Z(t), Z(t-\tau), Z(t-2\tau)]$ (shown as three red dots) can be represented as a single point in the 3-dimensional Z delay space (**C**, red dot). (**C**) We then shade each point of the Z delay space trajectory by its corresponding contemporaneous value of $Y(t)$ (without measurement noise). The shading is continuous (i.e. gradual transitions in shade), and note that Y causes Z (correctly, in this case). (**D**) When we repeat this procedure, but now shade the Z delay space trajectory by $W(t)$, the shading is bumpy, and note that W does not cause Z (even though Z causes W). (**E**) Shading the delay space trajectory of Z by the causally unrelated V also gives a bumpy result. (**F**) Dynamics as in (**C**), but now with noisy measurements of Y (purple in **B**). The shading is no longer gradual. Thus with noisy data, inferring causal relationships becomes more difficult. See Appendix A5 for more details.

In this example, there is a continuous delay map from a causee to a causer, but not the other way around, and also no continuous delay map between causally unrelated variables. If this behavior reflects a broader principle, then perhaps continuous delay maps can be used to infer the presence and direction of causation. Is there in fact a broader principle?

In fact, there is a sort of broader principle, but it may not be fully satisfying for causality testing. The principle stems from a classic theorem due to Floris Takens [91]. A rough translation of Takens' theorem is the following: If a particle follows a deterministic trajectory which forms a surface (e.g. an ant crawling all over a donut), and if we take one-dimensional measurements of that particle's position over time (e.g. the distance from the ant's starting position), then we are almost guaranteed to find a continuous delay map from our measurements (of current distance) to the original surface (the donut), as long as we use enough delays. (We walk through visual examples of these ideas in detail in Appendix A4.) A key result that follows from this theorem is that we can typically ("generically") expect to find continuous delay maps from "dynamically

driven" variables to "dynamically driving" variables in a coupled deterministic dynamical system, as long as certain technical requirements are met [85]. Although the notion of "dynamic driving" differs from our definition of causation [85], the two are related and we will still use the standard notion of causation when evaluating the performance of SSR methods. In theory, Takens' theorem says that almost any choice of delay vector should work as long as it contains enough delays. However in practice, with finite noisy data, the behavior of SSR methods can depend on the delay vector selection procedure ([92]; see also Appendix A4). Overall, Takens' theorem and later results [93, 85] form the theoretical basis of SSR techniques.

SSR techniques attempt to detect a continuous delay map (or a related feature) between two variables and use this to infer the presence and direction of causation [18, 94, 23]: A continuous delay map from Y to X is taken as an indication that X causes Y. The fact that the map points in the opposite direction as the expected causation is potentially counterintuitive. One informal explanation is that the delay vectors of the causee can contain a record of past influence from the causer [18]. As a word of warning, while causation is one possible explanation for a continuous delay map, it is not the only possible explanation. Indeed, we now illustrate scenarios where a causal relationship and a continuous delay map do not coincide.

SSR failure modes

Figure 5 illustrates four failure modes of SSR. In the first failure mode, which we refer to as "nonreverting continuous dynamics" (top row of Figure 5; Appendix A4), a continuous map arises from the delay space of X to Z because a continuous map can be found from the delay space of X to time ("nonreverting X") and from time to Z ("continuous Z"). This pathology leads to false causal conclusions and may explain apparently causal results in some early works where SSR methods were applied to time series with a clear trend. We are not aware of statistical tests for this problem, but Clark et al. [95] recommend shading points in the delay space with their corresponding time to visually check for a time trend. In the second failure mode [18] (Figure 5, second row), one variable drives another variable in such a way that the dynamics of the two variables are synchronized. Consequently, although the true causal relationship is unidirectional, bidirectional causality is inferred. Although the "prediction lag test" (Figure 6B right panel) can sometimes alleviate this problem [96, 92], it is not foolproof as we demonstrate in Appendix A4. In the third failure mode (Figure 5 third row; [92]), X and Z both oscillate and X's period is an integer multiple of Z's period. In this case, Z is inferred to cause X even though they are causally unrelated. In the fourth failure mode (Figure 5, bottom row), SSR gives a false negative error due to "pathological symmetry", although this may be rare in practice.

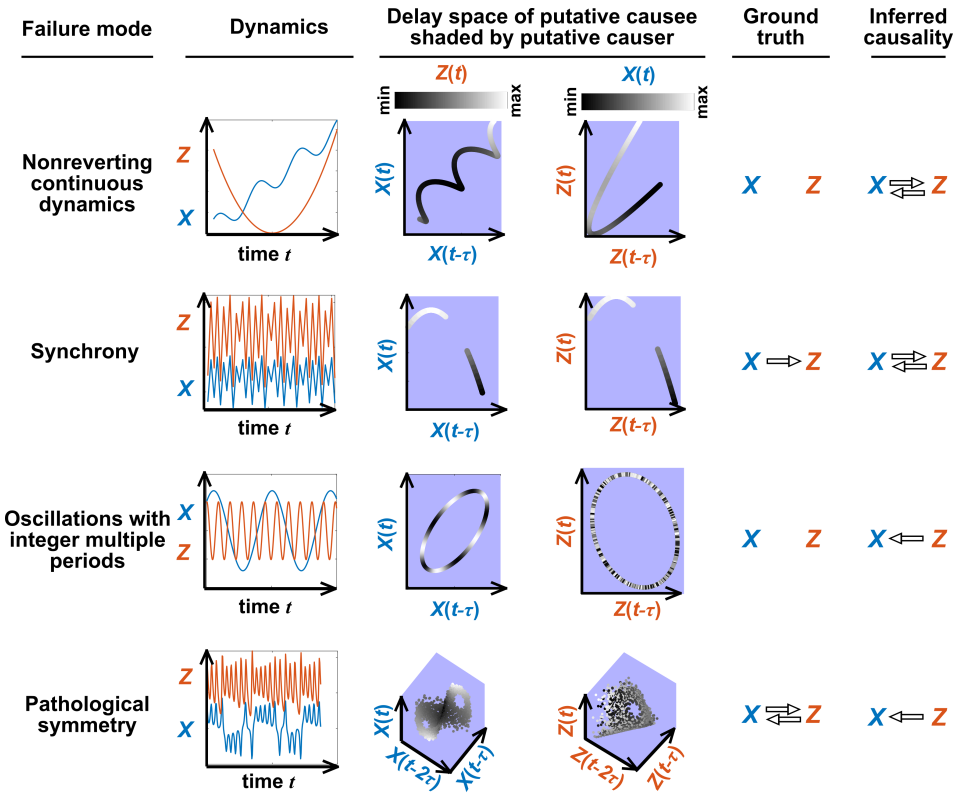


Figure 5: Failure modes associated with state space reconstruction. **Top row**: Nonreverting continuous dynamics may lead SSR to infer causality where there is none. This example consists of two time series: a wavy linear increase and a parabolic trajectory. Although they are causally unrelated, we can find continuous delay maps between them. This is because there is (i) a continuous map from the delay vector $[X(t), X(t-\tau)]$ to t (X is "nonreverting"), and (ii) a continuous map from t to Z (Z is "continuous"), and thus there is a continuous delay map from X to Z ("nonreverting continuous dynamics"). Thus, one falsely infers that Z causes X, and with similar reasoning that X causes Z. **Second row**: X drives Z such that their dynamics are "synchronized", and consequently, we find a continuous delay map also from X to Z even though Z does not drive X. Note that the extent of synchronization is not always apparent from inspecting equations (e.g. Figure 12 of [97]) or dynamics (Figure A24). **Third row**: X oscillates at a period that is 5 times the oscillatory period of Z. There is a continuous delay map from X to Z even through X and Z are causally unrelated. Note that true causality sometimes also induces oscillations where the period of one variable is an integer multiple of the period of another (e.g. in Figure 4, the period of Z is 3 times the period of X). **Bottom row**: In the classic chaotic Lorenz attractor, the X and Z cause one another, but we do not see a continuous map from the delay space of Z to X. This is because, as mentioned earlier, satisfying the conditions in Takens' theorem makes a continuous mapping likely but not guaranteed (Appendix A4). Here, Z is an example of this lack of guarantee [98] due to a symmetry in the system (see "Background definitions for causation in dynamic systems" in the supplementary information of [18]).

Convergent cross mapping: Detecting SSR causal signals from real data

SSR causal discovery methods require testing for the existence of continuous delay maps between variables. However, testing for continuity in real data is complicated by noise and discrete sampling (Figure 4, compare panels D and F; see also Figure A20).

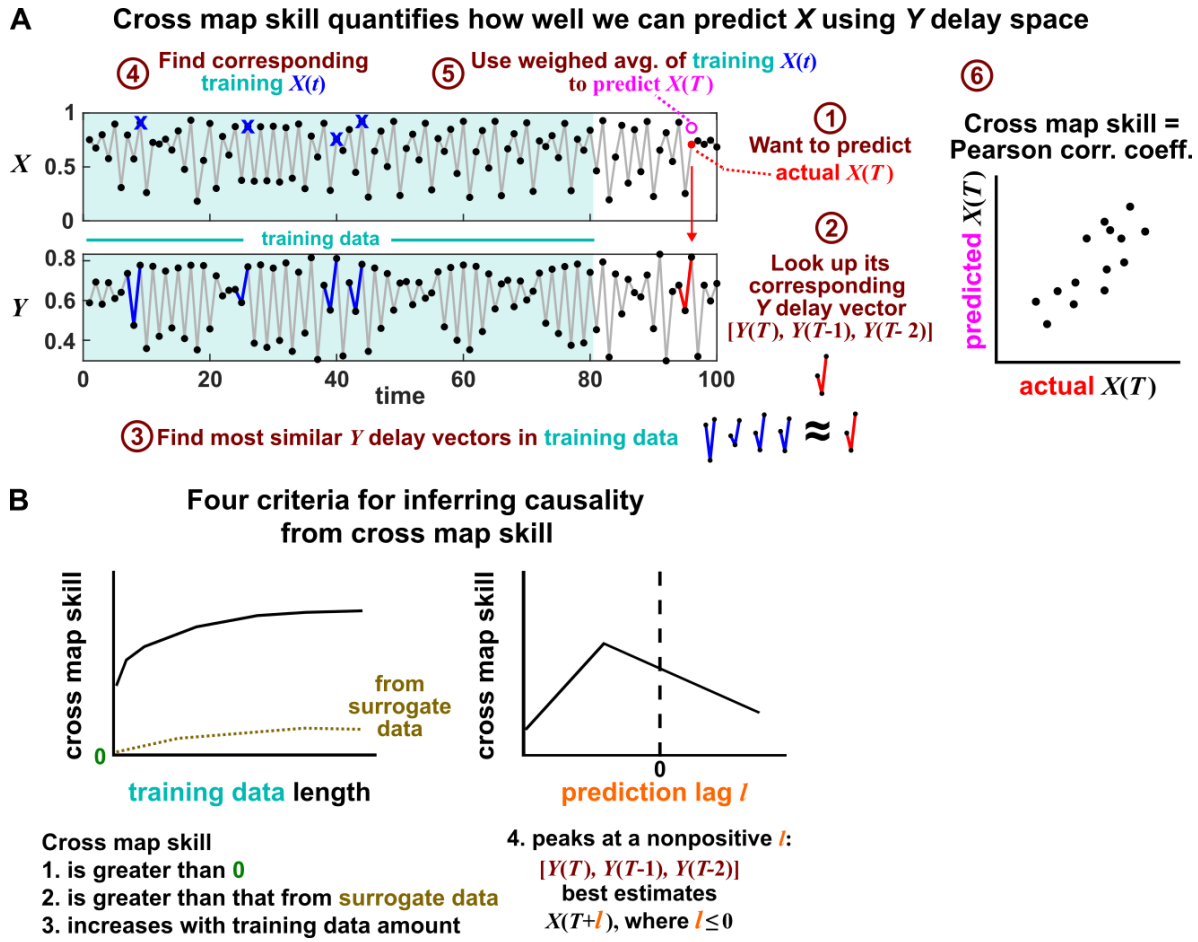


Figure 6: Illustration of the convergent cross mapping (CCM) procedure for testing whether X causes Y. (**A**) Computing cross map skill. Consider the point $X(T)$ denoted by the red dot ("actual $X(T)$" in ①), which we want to predict from Y delay vectors. We first look up the contemporaneous Y delay vector $[Y(T), Y(T-1), Y(T-2)]$ (②, red dynamics), and identify times within our training data when delay vectors of Y were the most similar (i.e. least Euclidean distance) to our red delay vector (③, blue segments). We then look up their contemporaneous values of X (④, blue crosses), and use their weighted average to predict $X(T)$ (⑤, open magenta circle; weights are given as equations S2 and S3 in the supplement of [18]). We repeat this procedure for many choices of T and calculate the Pearson correlation coefficient between the actual $X(T)$ and predicted $X(T)$ (⑥). This correlation is called the "cross map skill". While other measures of cross map skill, such as mean squared error, may also be used [18], our choice follows the convention of [18]. (**B**) Four criteria for inferring causality from the cross map skill. Data points in **(A)** are marked by dots and connecting lines are visual aids.

Several methods have been used to detect SSR causal signals by detecting approximate continuity [85] or related properties [18, 94, 23]. The most popular is convergent cross mapping (CCM), which has been applied to nonlinear [18] or linear deterministic systems [55]. CCM is based on a statistic called "cross map skill" that quantifies how well a causer can be predicted from delay vectors of its causee (Figure 6A), conceptually similar to checking for gradual transitions when shading the causee delay space by causer values (Figure 4). Four criteria have been proposed to infer causality [18, 96, 92] (Figure 6B): First, the cross map skill must

be positive. Second, the cross map skill must be significant according to some surrogate data test. Third, the cross map skill must increase with an increasing amount of training data. Lastly, cross map skill must be greater when predicting past values of the causer than when predicting future values of the causer (the prediction lag test [96, 92] in the right panel of Figure 6B, but see Appendix A4 for caveats of this test). In practice, many if not most CCM analyses use only a subset of these four criteria [18, 86, 87, 99]. Other approaches to detect various aspects of continuous delay maps have also been proposed [94, 85, 23, 7]. We do not know of a systematic comparison of these alternatives.

5 Simulation examples: External drivers and noise jointly influence causal discovery performance

In this section we examine how environmental drivers, process noise, and measurement noise can influence the performance of Granger causality and CCM, using computer simulations. We constructed a toy ecological system with a known causal structure, obtained its dynamics (with noise) through simulations, and applied a linear Granger causality test (using the MVGC package [20]) and CCM (using the R language package rEDM) to test how well we could infer causal relationships.

We simulated a two-species community in which one species (S_1) causally influences the other species (S_2) but S_2 has no influence on S_1 (Figure 7B). Additionally, S_1 is causally influenced by an unobserved periodic external driver and S_2 either is (Figure 7D) or is not (Figure 7E) causally influenced by its own (also unobserved) periodic external driver. In an ecosystem, external drivers might appear as changes in temperature, light, or water levels, for example. We also added process noise to model the stochastic nature of natural ecosystems and added measurement noise to model measurement uncertainty. Process noise propagates to future time steps and can result from, for instance, stochastic migration and death (Figure 7A). In contrast, measurement noise does not propagate over time, and includes instrument noise as well as ecological processes that occur during sampling. Unlike in linear Granger causality, there is no default test procedure for CCM causality criteria [92, 55]. We therefore tested for CCM criteria using two different procedures (Figure 7 legend and Appendix A5).

Granger causality and CCM can perform well when their respective requirements are met, but both are fairly sensitive to the levels of process and measurement noise (Figure 7D and E, correct inferences colored as green in pie charts) and to details of the ecosystem (whether or not S_2 has its own external driver; compare Figure 7D with E). In both methods, detection of the true causal link is disrupted by either the strongest measurement noise (standard deviation of 1) or the strongest process noise (standard deviation of 8) used

here.

For Granger causality (Figure 7D and E, left panels), the MVGC package correctly rejects the data as inappropriate in the deterministic setting (lower left corner). When process and/or measurement noises are present, their relative amount is important: As measurement noise increases (from bottom to top), process noise needs to increase (from left to right) for Granger causality to perform well. Indeed, prior analytical results [64, 65] show that measurement noise can induce false positives (e.g. red slices in row 2, column 2) and hide true positives (e.g. grey slices in row 1). Surprisingly, increasing measurement noise can sometimes improve performance (in column 3 of both panels, row 2 has a larger green slice than row 3).

To understand the CCM results (Figure 7D and E, right panels), recall that CCM is designed for deterministic systems, and fails when dynamics of variables are synchronized. When S_2 has its own external driver (Figure 7D), there is no synchrony, and CCM performs admirably in the deterministic setting (lower left corner). CCM performs less well when measurement or process noise is introduced. Strikingly, when we remove the external driver of S_1 (Figure 7E), CCM performs poorly. This is likely because the two species are now synchronized in the absence of noise (violating the "no synchrony" requirement of CCM). However, adding noise, which removes the synchrony problem, violates the determinism requirement. So CCM is hapless either way. Note that unlike CCM, Granger causality is less sensitive to the presence of underlying synchrony as long as this synchrony is disrupted by process noise. Additionally, the performance of CCM (Figure 7D and E, right panels) is sensitive to the test procedure (olive brackets).

In reality, where a system lies in the spectrum of process versus measurement noise is often unknown, and we are not aware of any method that reliably distinguishes between process noise and measurement noise without knowing the functional form of the system. Furthermore, how might one tell if a time series is stochastic or deterministic so that one can choose between Granger causality versus CCM? One idea is that deterministic processes tend to be more predictable than stochastic processes, at least in the short term [100]. Indeed, the inventors of CCM have recommended checking whether historical values of a time series can be used to accurately predict future values [101] before applying CCM (i.e. [95]). However, practical time series found in nature are most likely somewhere between the extremes of "fully deterministic" (i.e. no measurement or process noise) and "fully stochastic" (i.e. IID). Time series are often partly deterministic due to autocorrelation and partly stochastic due to random fluctuations. Indeed, simulations have found that SSR-based and Granger causality-based methods can both potentially succeed for such systems [55]. Future work is needed to flesh out the nuances of when and why methods from these two classes provide similar or different performance [55].

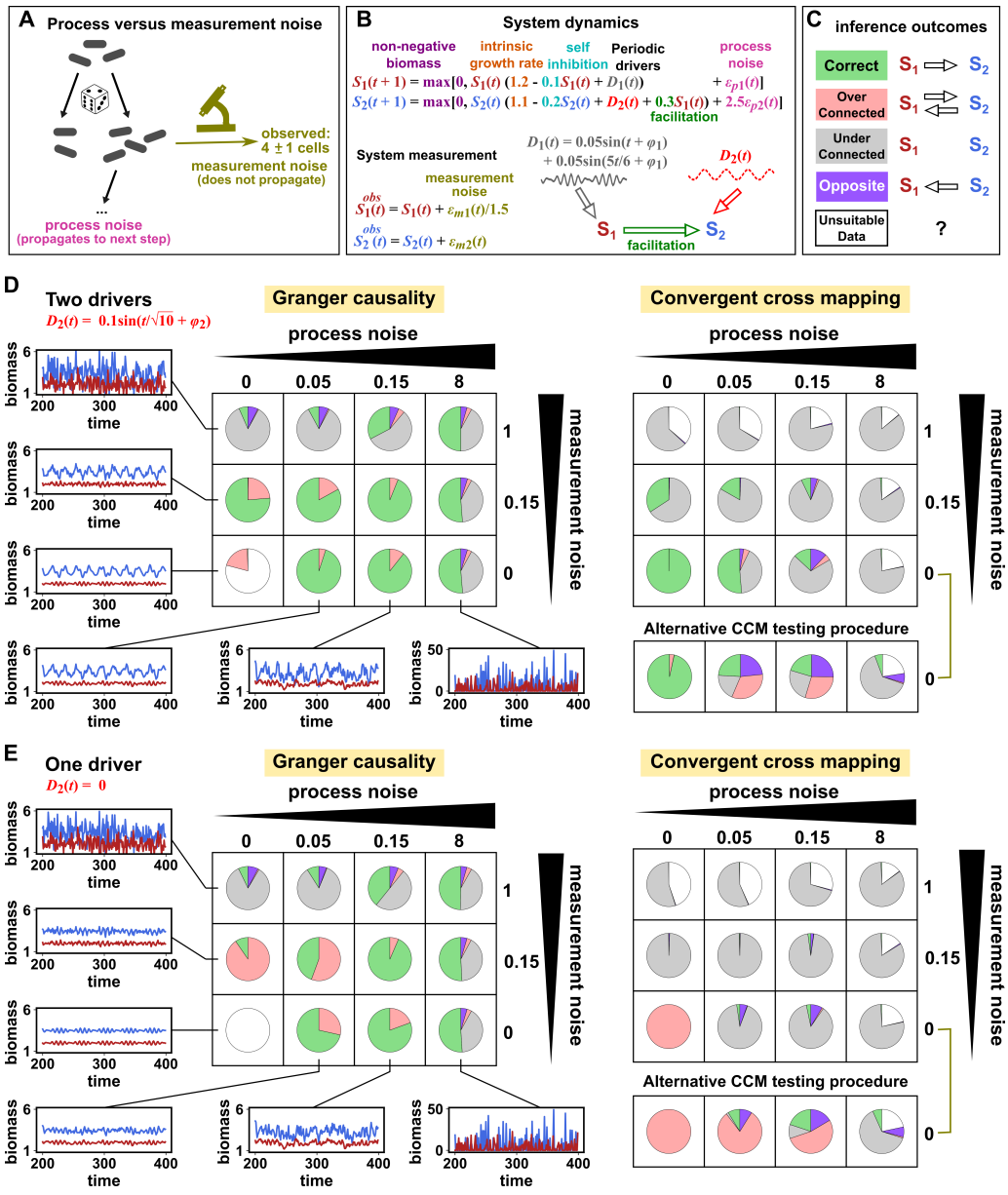


Figure 7: Performance of Granger causality and convergent cross mapping in a toy model with noise. (**A**) The effect of process noise but not measurement noise propagates to samples taken at subsequent time points. (**B**) We simulated a two-species community. The process noise terms $\epsilon_{p1}(t)$ and $\epsilon_{p2}(t)$, as well as the measurement noise terms $\epsilon_{m1}(t)$ and $\epsilon_{m2}(t)$, are IID normal random variables with a mean of zero and a standard deviation whose value we vary. (**C**) Five possible outcomes of the causal analysis. (**D**, **E**) Community dynamics and causal analysis outcomes. We varied the level (i.e. standard deviation) of process noise and measurement noise. For Granger causality, we used the MVGC package (Appendix A5). For convergent cross mapping, we used the rEDM package to calculate cross map skill and to construct surrogate data, and custom codes for other tasks (Appendix A5). Each pie chart shows the distribution of inference outcomes from 1000 independent replicates. Note that MVGC package for inferring Granger causality does not necessarily flag data corrupted by a problematic level of measurement noise [102]. . In both the main and alternative CCM procedures, criterion 1 (positive ρ) was checked directly and random phase surrogate data were used to test criterion 2 (significance of ρ). Criterion 4 (prediction lag test) was not used, because the test is difficult to interpret for periodic dynamics where cross map skill can oscillate as a function of prediction lag length (Figure A24). The two procedures differ only in how they test criterion 3 (ρ increases with more training data): the main procedure uses bootstrap testing following [92] while the alternative procedure uses a Kendall's τ as suggested by [103].

6 Summary: Model-free causality tests are not assumption-free

We have described three causal discovery approaches for observational time series (Table 1). Although the techniques explored in this article have been called model-free and do not depend on prior mechanistic knowledge, they are by no means free from assumptions [16]. The danger that arises when we replace knowledge-based modeling with model-free inference is that we can replace explicitly stated assumptions with unstated and unscrutinized assumptions. Too frequently, both methodological and applied works fall into this trap. Nevertheless, when assumptions are clearly articulated and shown to be reasonable, model-free causal discovery techniques have the potential to jump-start the discovery process where little mechanistic information is known. Still, experimental follow-up (when possible) remains valuable since any technique that seeks to infer causality from observational measurements will typically require at least some assumptions that are difficult to fully verify.

We have discussed several failure modes of various causal discovery approaches (Table 1). Among these failure modes, measurement noise and nonstationarity have been repeatedly singled out as crucial considerations for real data [104, 28, 105]. While the deleterious effect of excessive measurement noise is intuitive, the pernicious effect of nonstationarity is not always appreciated. This is perhaps because the stationarity requirement, although ubiquitous, is sometimes hidden in the analysis pipeline. For example, when testing whether cross map skill (Fig 6B, condition 2) or correlation is significant, one must choose from among a handful of surrogate data tests (e.g. [43]), nearly all of which require stationary data. Granger causality tests also typically require the data to be stationary.

What comes next? We cannot cover all open fronts in data-driven causal discovery from time series, but do note a few important directions here. First, given that practical ecological time series can rarely be shown to satisfy the assumptions of tests with mathematical exactness, we need a more complete understanding of how well tests for dependence and/or causality tolerate moderate deviations from assumptions. In a different direction, one may sometimes possess not a complete mathematical model, but instead some pieces of a model, such as the knowledge that nutrients influence the growth of organisms according to largely monotonic saturable functions. Techniques that attempt to make use of such partial models have recently obtained intriguing results [8, 106, 9], and more would be welcome. Moreover, natural experiments often involve known external perturbations that are random or whose effects are poorly understood. An important question is how inference techniques might best take advantage of such perturbations [107, 108].

Perhaps most importantly, how can method developers best communicate their assumptions and caveats to method users who are potentially unfamiliar with technical terms or concepts? One effective strategy is to provide simulation examples of how applying techniques to pathological data may give incorrect results

[95, 106]. Video walkthroughs (e.g. [109, 110]) may be another useful way to communicate how a method works as well as method assumptions. Finally, we recommend that editors and reviewers work with authors to ensure that failure modes and caveats are clearly articulated in the main text, along with with accessible explanations of any necessary technical terms or concepts.

	What is the method detecting?	Implied causal statement	What are some possible failure modes?
Correlation	Whether X and Y are statistically dependent	X causes Y, Y causes X, or Z causes both.	Surrogate null model may make incorrect assumptions about the data-generating process.
Granger causality	Whether the history of X contains unique information that is useful for predicting the future of Y.	X directly causes Y.	Hidden common cause; infrequent sampling; deterministic system (no process noise); excessive process noise; measurement noise
State space reconstruction	Whether the delay space of X can be used to estimate Y.	Y causes X.	Nonreverting continuous dynamics; synchrony; integer multiple periods; pathological symmetry; measurement or process noise

Table 1: A comparison of three statistical causal discovery approaches

Appendix to Part I

A1 Random variables and their relationships

Dependence between random variables and between vectors of random variables

The concepts of dependence and independence between random variables are central to many statistical methods, including those that concern causality. A random variable is a variable whose values or experimental measurements depend on outcomes of a random phenomenon and follow a particular probability distribution. Reichenbach's common cause principle states that if X and Y are random variables with statistical dependence (such as a nonzero covariance), then one or more of three statements is true: X causes Y, Y causes X, or a third variable Z causes both X and Y. The common cause principle cannot be proven from the axioms of probability; rather, the principle is itself a fundamental assumption that supports much of the modern statistical theory of causality ([22] Section 1.4.2).

As an example, consider the size and length of a bacterial cell. If a larger cell tends to be longer, then cell volume and cell length covary and are thus dependent. A mathematical definition of dependence (and its opposite, independence) is presented in Figure A8B.

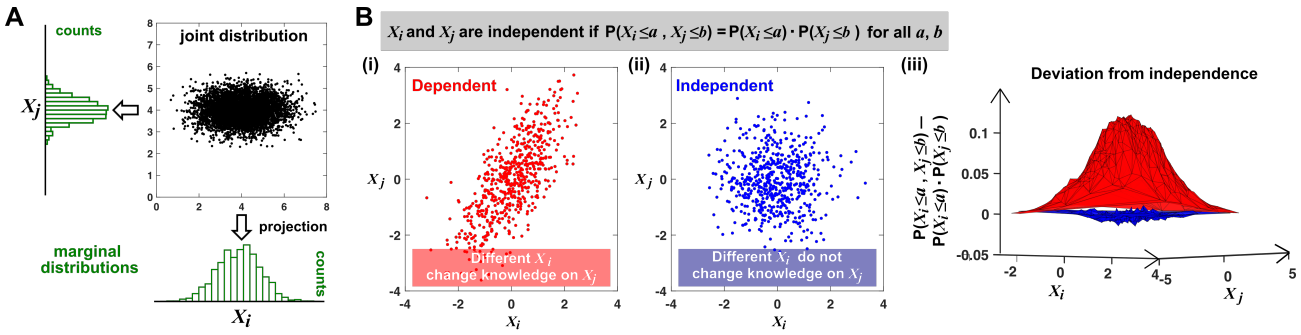


Figure A8: Joint distribution, marginal distributions, and dependence of two random variables. (**A**) A scatterplot of data associated with random variables X_i and X_j represents a "joint distribution" (black). Histograms for data associated with X_i and for data associated with X_j represent "marginal distributions" (green). Strictly speaking, joint and marginal distributions must be normalized so that probabilities (here represented as "counts") sum to 1. Graphically, marginal distributions are projections of the joint distribution on the axes. Two random variables are identically distributed if their marginal distributions are identical. (**B**) Independence between two random variables. Gray box: the mathematical definition of independence, where "P" means probability. Two random variables are dependent if and only if they are not independent. Visually, if two random variables are independent, then different values of one random variable will not change our knowledge about another random variable. In (**i**), X_j increased as X_i increased (different X_i led to different knowledge on X_j), and thus, X_i and X_j are not independent (i.e. they are dependent). In (**ii**), X_i and X_j were repeatedly drawn from two normal distributions. Thus, the two random variables are independent. One might argue that when X_i values become extreme, X_j values tend to land in the middle. However, this is a visual artifact caused by fewer data points at the more extreme X_i values. If we had plotted histograms of X_j at various X_i values, we would see that X_j is always normally distributed with the same mean and variance. (**iii**) Indeed, when we plotted the difference between the observed probability $P(X_i \le a, X_j \le b)$ and the probability expected from X_i and X_j being independent $P(X_i \le a) \cdot P(X_j \le b)$, (ii) showed a near-zero difference (blue), while (i) showed deviation from zero (red). This is consistent with X_i and X_j being independent in (ii) but not in (i).

Random sampling from a population with replacement is one way to produce "IID data" (which we use as a shorthand for "data which can be modeled as IID random variables"). For example, repeatedly rolling a standard die can be thought of as randomly sampling from the set $\{1, 2, 3, 4, 5, 6\}$ without replacement: if the first trial registers 1, then the second trial can register 1 as well. Otherwise, if sampling was done *without* replacement, then the second trial must not register 1, which means that the outcome of the second trial would depend on the outcome of the first trial.

Dependence can be readily generalized from the definition in Figure A8 to become a property between two vectors of random variables. (Note that a time series can be viewed as a vector of random variables.) For example, suppose that we measure two variables X and Y over two days. Our (very short) time series are then $[X_1, X_2]$ and $[Y_1, Y_2]$ where the subscript index denotes the day of measurement. Similar to Figure A8B, we would say that our two time series are independent if

$$P(X_1 \le x_1, X_2 \le x_2, Y_1 \le y_1, Y_2 \le y_2) = P(X_1 \le x_1, X_2 \le x_2) P(Y_1 \le y_1, Y_2 \le y_2)$$

for all choices of x_1, x_2, y_1, y_2.

When are two random variables independent and identically distributed (IID)?

Many statistical techniques require repeated measurements that can be modeled as independent and identically distributed (IID) random variables, and passing non-IID data (such as time series) into such techniques can lead to spurious results (Fig 2; [111]). Random variables are IID if they have the same probability distribution and are independent (Figure A8). In Figure A9 we give examples of pairs of random variables that are (or are not) identically distributed, and that are (or are not) independent. Note that two dependent random variables can be linearly correlated (Figure A9 3rd column), or not (Figure A9 4th column).

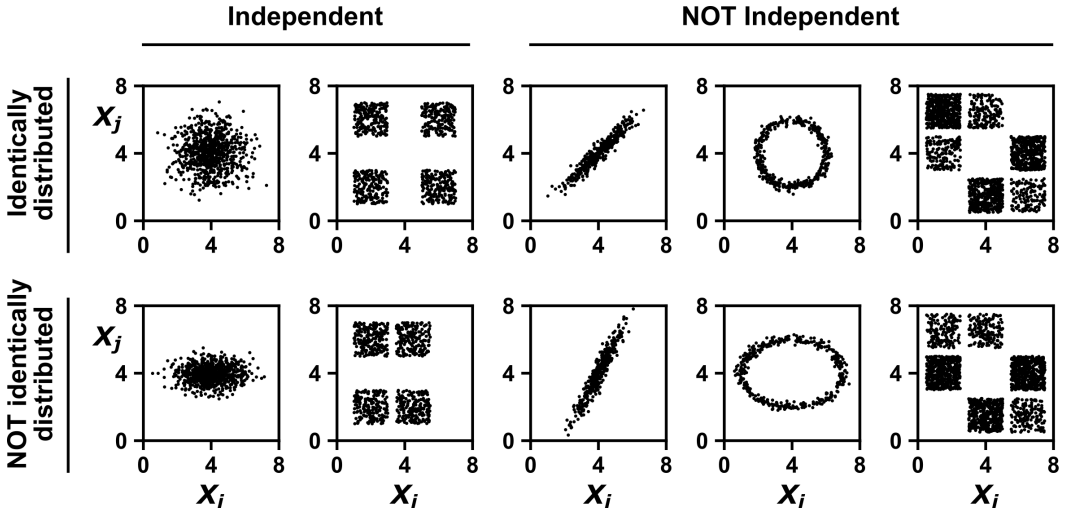


Figure A9: **Examples of random variables that are identically distributed or not identically distributed, and independent or not independent.** In the top row, X_i and X_j are identically distributed (projections of the scatter plot on both axes would have the same shape, as in Figure A8A). Note that in the top row of the rightmost column, the scatter plot is not symmetric along the diagonal line, yet projections on both axes yield identical marginal distributions: three segments of equal densities. Thus, the two random variables are identically distributed. In the bottom row, X_i and X_j are not identically distributed. In the leftmost two columns, the two random variables are independent (for more details about independence, see Figure A8B). In the last three columns, the two random variables are dependent: different X_i values alter our knowledge of X_j.

A sample drawn from a mixed population can still be IID, as long as sample members are chosen randomly and independently

Since the IID concept is so central to statistical analysis, we wish to further clarify one conceptual difficulty that may arise. To set the stage, suppose that a scientist measures the levels of voluntary physical activity in a collection of mice that includes both males and females. Also suppose that female mice tend to be more physically active than male mice [112]. Since this dataset now contains measurements from both the less

active males and the more active females, we might naively think that these data cannot be IID.

In fact, such a dataset still might be IID, but this depends on how the scientist chooses which mice to measure. To illustrate this fact, consider the highly simplified scenario in which only two mice are assayed for physical activity. Let X_1 and X_2 be random variables that describe the activity levels of these two mice. We consider three different ways that the scientist might select which mice to assay. Only one of these ways will result in an IID dataset.

First, suppose that the scientist chooses to measure X_1 from a male mouse and X_2 from a female mouse. In this case, to see whether X_1 and X_2 are IID, we can use the same visualization strategy as in Figure A8. That is, we imagine many possible "parallel universes", each with a different possible two-mouse dataset (left panel of Figure A10). This allows us to visualize the joint distribution of X_1 and X_2. We can then see that X_1 and X_2 are independent, but not identically distributed.

Second, suppose that the scientist again selects exactly one mouse of each sex, but randomizes the order so that both X_1 and X_2 have an equal chance of being measured from a male or female mouse (middle panel of Figure A10). We can now see that X_1 and X_2 are identically distributed, but not independent.

Lastly, suppose that the scientist selects mice randomly, and without any information about whether a mouse is male or female. In this case, the two-mouse sample might be all male, all female, or have one of each. Once again we plot the joint distribution of X_1 and X_2 by imagining their values across many different parallel universes (right panel of Figure A10). We then see that that X_1 and X_2 are finally independent identically distributed. Overall, a set of measurements can be IID even if they are taken from a mixed population, as long as they are sampled randomly from among different subpopulations.

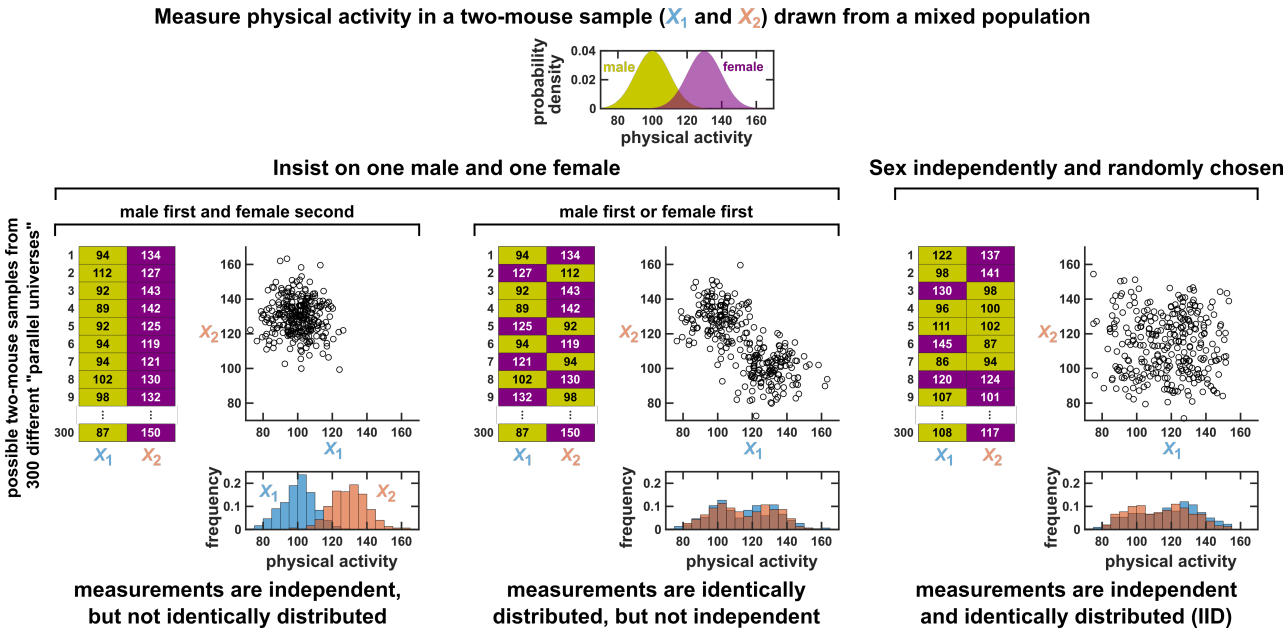


Figure A10: **Measurements taken from a mixed population may still be IID, as long as sampling is independent and random.** Consider a study in which physical activity is measured from a mixed population of low-activity male mice and high-activity female mice. For simplicity, suppose that the study uses only two mice. To see whether this could be an IID dataset, we imagine drawing many possible versions of that sample, and ask whether our first measurement X_1 and second measurement X_2 are identically distributed and independent. We could draw this sample in 3 different ways (3 sets of charts). On the left, we take our first measurement X_1 from a male and second measurement X_2 from a female. In this case, our two measurements are independent, but not identically distributed, and thus not IID. In the middle, we choose one male and one female per sample, but choose the first measurement randomly from a male or female. Now, our measurements are identically distributed (save for sampling error) but not independent (so also not IID). On the right, the sex of each measurement is randomly and independently chosen so that, for example, a sample might have two measurements from the same sex. In this case our sample is an IID dataset.

Independence and statistical conditioning

Here, we will restate the concept of independence in equivalent forms that will allow us to more easily transition to the concept of conditional independence.

It is intuitive that two variables are independent if knowledge of one variable tells us nothing about the other. The statistical notion of independence captures this intuition: Random variables X and Y are independent if the conditional distribution of X given Y is always equal to the marginal distribution of X. For discrete random variables, this condition can be written

$$P(X = x|Y = y) = P(X = x) \tag{3}$$

, or equivalently written $P(X = x, Y = y) = P(X = x)P(Y = y)$, for all x and y. For continuous random

variables, independence can be written in terms of probability density functions as $f_X(x|y) = f_X(x)$ or equivalently, $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ where $f_X(x|y)$ is the conditional density of X given Y, $f_{X,Y}(x,y)$ is the joint density of X and Y, and $f_X(x)$ and $f_Y(y)$ are the marginal densities of X and Y, respectively.

The statement "X and Y are conditionally independent given Z" intuitively means that X and Y are independent when we only analyze outcomes where Z has a certain value. For discrete random variables, this condition is written $P(X = x|Y = y, Z = z) = P(X = x|Z = z)$, or equivalently, $P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z)$, for all x, y, and z. For continuous random variables, we have a similar formulation except that probability P is replaced by probability density f (i.e. $f_{X,Y}(x,y|z) = f_X(x|z)f_Y(y|z)$ for all x, y, z). If X and Y are not conditionally independent given Z, then X and Y are conditionally dependent given Z.

One could be forgiven for worrying about the feasibility of testing for dependence between long time series. This is because as a time series grows longer, the amount of data needed to get a sense of its probability distribution would seem to grow extremely rapidly. Thus, when X and Y are vectors that represent long time series, estimating the distributions in Eq. 3 seems unrealistic. However, establishing that two time series are dependent only requires that we show that the distributions on the left and right sides of Eq. 3 differ. Showing that two distributions differ can be much easier than actually estimating those distributions. For instance, if we know that the averages of two univariate distributions are different, then we immediately know that the two distributions are not the same, even if we know nothing of their shapes. Indeed, Fig A11 demonstrates a way to test for dependence between time series with only a moderate number of replicates, and without any assumptions about the shapes of the distributions. Additionally, if certain assumptions can be made, then surrogate data can be used to test for dependence with only one replicate of each time series, as discussed in the main text.

When multiple trials exist, the significance of a correlation between time series can be assessed by swapping time series among trials

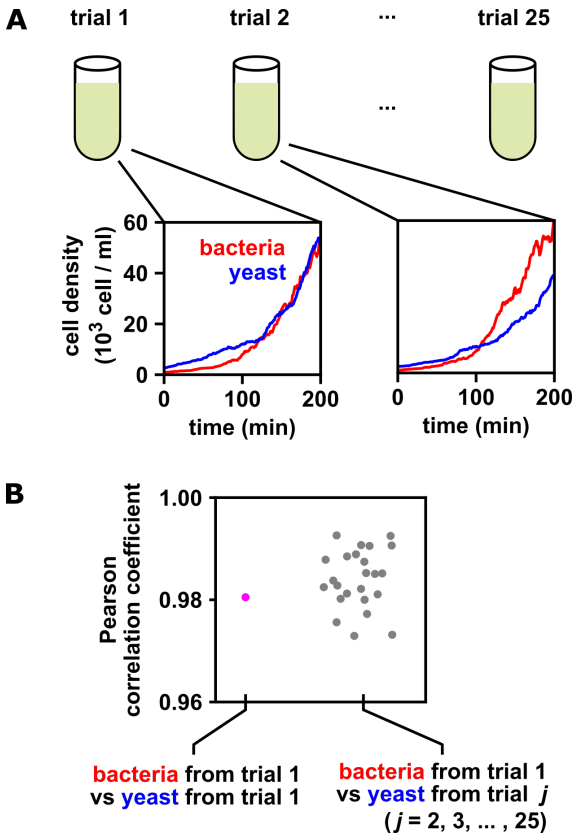


Figure A11: When multiple identical and independent trials are available, the significance of a correlation between time series within a trial can be assessed by comparing it to correlations between trials. (A) A thought experiment in which yeast and bacteria are grown in the same test tube, but follow independent dynamics. We imagine collecting growth curves from 25 independent replicate trials. (B) Correlations within and between trials. The Pearson correlation coefficient between yeast and bacteria growth curves from trial 1 is a seemingly impressive ~ 0.98 (pink dot). But does this result really indicate that the two growth curves are dependent? To answer this question, notice that the yeast curves from other trials are similarly highly correlated to the bacteria curve from trial 1, even though they all come from independent trials (grey dots). Therefore, the pink dot cannot be used as evidence that the yeast and bacteria growth are dependent. If the within-trial correlation (pink dot) were stronger than, for instance, 95% of the between-trial correlations (grey dots), we would have evidence of dependence.

A2 Causal discovery with directed acyclic graphs

Discovering causal relationships and their associated directed acyclic graphs (DAGs)

Many theoretical results and data-driven methods for causal analysis begin by representing causal relationships as an acyclic directed graph (DAG). That is, one makes a graph by representing random variables as nodes and by drawing a directed edge from each direct cause (or parent) to its causee (or child), as in Fig 1B; additionally, the graph is acyclic, meaning that that it does not contain any directed paths from any variable back to itself. The acyclicity condition is often required for nice theoretical properties and ease of analysis [1]. Additionally, when data are temporal, a particular node in the graph commonly refers to a particular variable measured at a particular time (e.g. chapter 10 of [29]). If we follow this convention and note that causation cannot flow backward in time, and if we additionally exclude instantaneous causation, then our causal graph will be acyclic, even for systems with feedback (Figure A15).

DAGs are useful visual tools in their own right, but for many purposes we need to be more mathematically precise about what we mean when we draw an edge from one variable to another. Thus, often one interprets a causal DAG as corresponding to a set of equations with the following two conditions: First, each variable can be written as a function of (only) the variable's direct causers and a random process noise term unique to the variable. Models that satisfy this condition are called structural equation models (SEMs) [31]. Second, all process noise terms are (jointly) independent of one another. SEMs that satisfy this second condition are called Markovian and have a useful property called the "causal Markov condition" [22]. (Some authors [29], but not all [31], require that all SEMs be Markovian by definition.) The causal Markov condition, along with the related "causal faithfulness condition" are key assumptions that allow one to connect statistical structure to causal structure and infer aspects of causal structure from data, even in observational settings.

The causal Markov condition states that if there is no path from X to Y in a DAG (i.e. we cannot go from X to Y by following a sequence of edges in the forward direction), then X and Y are conditionally independent given X's parents [113, 22]. In this context Y can be either a variable or a set of variables. As an example, consider the boxed DAG in Figure A13Bii. Here, X and Y share the common cause Z. Each variable depends on its parents, and on its own process noise term. Although X and Y are dependent, the causal Markov condition expresses the intuitive idea that if we were to control for Z, X and Y would become independent. Note that if X does not have any parents, then the statement "X and Y are conditionally independent given X's parents" reduces to "X and Y are independent".

The causal faithfulness condition is, like the Markov condition, very useful in causal discovery and

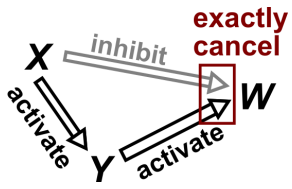


Figure A12: Violation of faithfulness due to cancellation of causal effects. Although X has a direct causal effect on W, we assume here that this is exactly cancelled out by an opposing influence via the indirect route of $X \rightarrow Y \rightarrow W$. Thus, although the Markov condition does not require that X and W be independent, X and W are actually independent. We thus say that the joint probability distribution of the variables $\{X, Y, W\}$ is not faithful to the graph.

often quite reasonable. However, faithfulness is more difficult to state precisely and concisely without first introducing technical notation such as "d-separation" (as in definition 6.33 of [29]). We attempt to give the gist of the idea here and direct readers to other sources [113, 29] for more precise definitions. The causal faithfulness condition is a kind of converse to the causal Markov condition. Recall that the causal Markov condition requires certain conditional (or unconditional) independence relationships based on the causal graph structure. Let us call any other independence relationships (i.e. those not required directly or indirectly by the causal Markov condition) "extra" independence relationships. The (joint) probability distribution of random variables is causally faithful to the DAG if no "extra" independence relationships exist [31]. An imprecise shorthand for the faithfulness condition is "independence relationships indicate the absence of certain causal relationships". The faithfulness condition can be violated when two effects precisely cancel each other (Figure A12).

Existing causal discovery methods for observational nontemporal data are diverse. Such methods can differ greatly in the assumptions they make (e.g. whether there are hidden variables or "unknown shift interventions" [52]), the lines of reasoning they employ, and the resolution of causal detail they provide (e.g. a unique causal graph versus a set of several plausible graphs) [52]. We will briefly introduce two classes of causal methods: (1) constraint-based search and (2) structural equation models (SEMs) with assumptions about the functional forms of equations [1]. However, these two classes, while illustrative of different modes of causal discovery, are far from an exhaustive list [52].

Constraint-based search uses independence and dependence relationships (and their conditional counterparts) to narrow down the scope of causal graphs without exhaustively checking all possibilities (which can be enormous in number even for a handful of variables). The PC algorithm (named after its inventors Peter Spirtes and Clark Glymour) and the fast causal inference algorithm are examples of constraint-based search methods [3]. However, constraint-based methods often find multiple graphs that are consistent with the same set of data (e.g. the second to last row of Figure A13 ii, see legend; see also [1]).

Functional form-based (or SEM-based) approaches to causal discovery begin by assuming a particular

functional form for causal relationships, and then assess a given causal hypothesis by inspecting the joint distribution between a potential causer and its potential causee [1]. These methods rely on the fact that in a Markovian SEM, each variable has a noise term that is independent of the noise terms of all other variables [29]. Given two dependent variables with no hidden common causes, one can use an appropriate regression to estimate values of a proposed causee based on the proposed causer [1]. If the residuals of this regression are independent from the proposed causer, then the proposed causal direction is consistent with the data [114]. Crucially, theoretical results indicate that for a fairly wide variety of scenarios (e.g. linear non-Gaussian and post-nonlinear models), we can expect the data to be consistent with only one causal direction, thus enabling unambiguous identification of the causal direction [1]. An illustrative graphic example is given in Fig 3 of [1] and also in Fig 3 of [3]. Similar ideas can be applied to multivariate systems [115, 114].

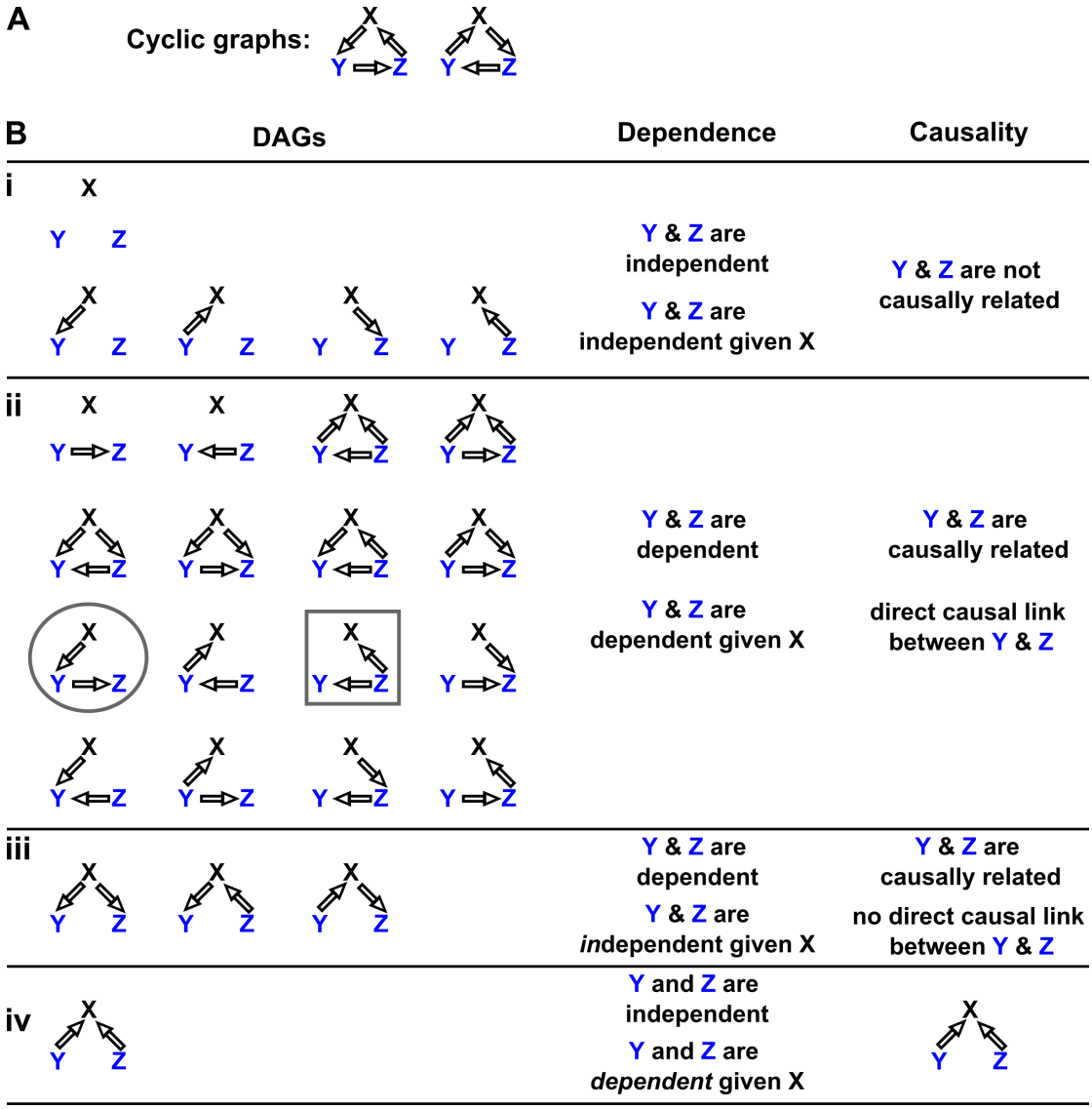


Figure A13: Probability distributions alone can specify causal structure to varying degrees of resolution. Consider a system of 3 and only 3 random variables X, Y, and Z. Between each pair of variables, there are in three possible unidirectional relationships: causation in one direction, causation in the opposite direction, and no causation. With three pairs of variables and three types of relationships, there are $3^3 = 27$ possible graphs. (A) Two of these graphs are cyclic, while the rest are DAGs. (B) If our system is described by a Markovian and causally faithful SEM, we can infer some aspects of causal structure from probability distributions alone. We demonstrate this by using dependence relationships between Y and Z (blue) to infer causal relationships. (**i**): Y and Z are always independent. Y and Z are not causally related. (**ii**): Y and Z are dependent, implying that they are causally related. (Recall that in this article, two variables are "causally related" if one causes the other, or they share a common cause.) Furthermore, Y and Z are conditionally dependent given X. For example, in the circled graph, variation in Y will affect Z, leading to dependence between Y and Z, even if we control for X. (**iii**): Y and Z are dependent, but are conditionally independent given X. There is no direct link between Y and Z, but they are causally related. Note that all three graphs are consistent with the following observations: Y and Z are conditionally independent given X; X and Z are conditionally dependent given Y; X and Y are conditionally dependent given Z. Thus, we cannot uniquely identify the causal structure from probability distributions alone. (**iv**): Y and Z are independent, but are conditionally dependent given X; see Figure A14 for an example of this scenario . This scenario corresponds to one and only one possible DAG.

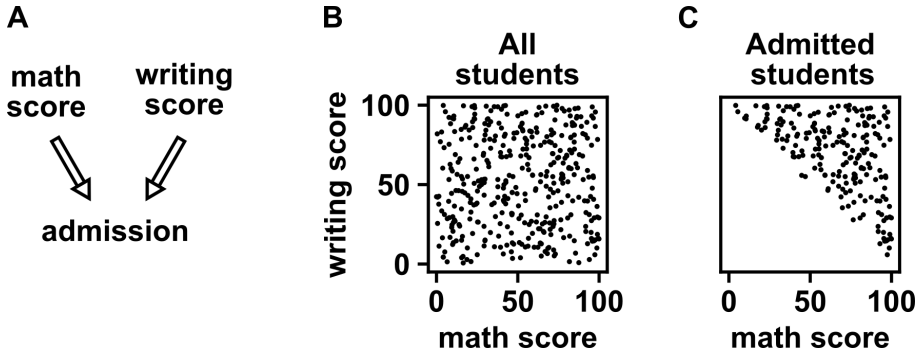


Figure A14: Math exam scores and writing exam scores are independent, but become dependent when we condition on college admission, which both scores jointly influence. (A) DAG depicting the assumed causal relationship between math scores, writing scores, and admission to a certain college. (B) Math and writing scores in a fictitious student population are independent of each other, and take on random values distributed uniformly between 0 and 100. (C) A college admits a student if and only if their combined score exceeds 100. It is apparent that when we condition on college admission (by plotting only the scores of admitted students), math and writing scores show a negative association, indicating that they are dependent.

Time series of systems with feedback loops can be analyzed by methods designed for directed acyclic graphs

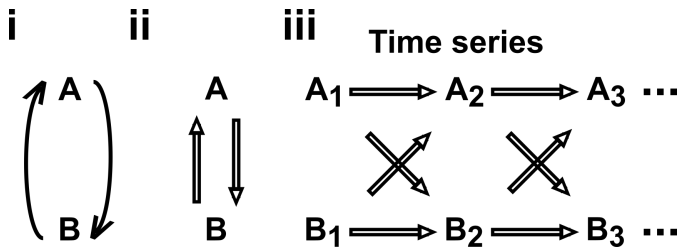


Figure A15: Causal discovery approaches designed for directed acyclic graphs (DAGs) can be applied to time series from systems with feedback. (**i**) Consider a mutualistic system where A and B represent the population sizes of two species that mutually facilitate each other's growth. (**ii**) When the role of time is ignored, the causal graph is cyclic and thus not a DAG. (**iii**) For time series data where A_1, A_2, ... represent the population size of A at times $1, 2, ...$, the causal graph is no longer cyclic since A_1 causes B_2 and B_1 causes A_2 etc. Note that A_1 causes A_2 (and similarly B_1 causes B_2). This framework [29] has helped one of the authors classify mutations in communities with feedback [116, 117].

A3 Mathematical concepts for stochastic time series

Intuition for random phase surrogate data

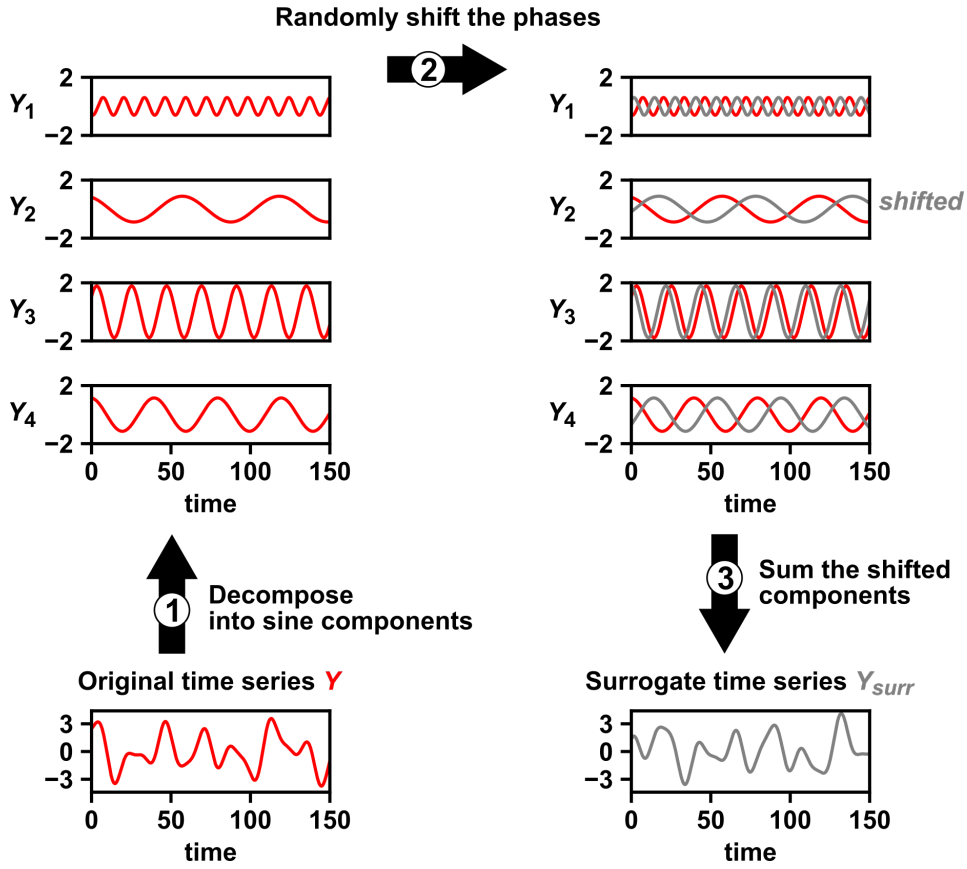


Figure A16: Intuition for random phase surrogate data methods. Random phase surrogate data methods generate Y_{surr} by representing Y as a sum of sine waves (upper left), randomly shifting the phases of the component sine waves (upper right grey), and summing up the shifted sine waves (lower right green).

Covariance-stationarity

A stochastic process X_t is covariance-stationary (or wide-sense stationary) if:

1. $\mathbb{E}[X_t]$ (the ensemble mean) does not depend on t
2. $\text{Var}[X_t]$ is finite and does not depend on t
3. For all choices of h, $\text{Cov}(X_t, X_{t+h})$ does not depend on t

As an example similar to Figure 2C, consider a population whose dynamics are governed by death and stochastic migration:

$$X_t = (1-a)X_{t-1} + c + \epsilon_t \tag{4}$$

Here, X_t is the population size at time t, a is the probability of death during the time interval of 1, c is the average number of individuals migrating into the population during the time interval of 1, and ϵ_t is a random variable with a mean of zero which represents temporal fluctuations in the number of migrants. Suppose that we observed the dynamics of 10 populations governed by Eq. 4 such that the populations all have the same parameters, but are independent (Figure A17A). Then, at each time point t, we will have some distribution of values of X_t. In fact, if we have not just 10, but 1,200 replicates, we can see that the distribution of values of X_t does not appear to depend on time (Figure A17B, top). Furthermore, the covariance between X_t and X_{t+1} does not appear to depend on time either (Figure A17B, bottom).

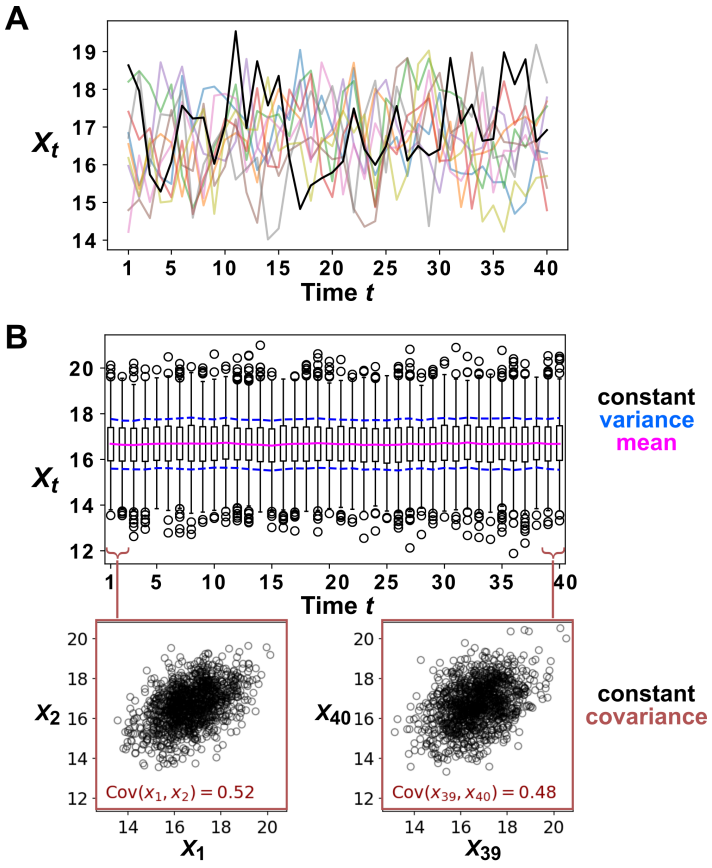


Figure A17: Example of a stationary process. (**A**) Ten replicate runs of the stochastic process described in Eq. 4 with parameter choices $a = 0.6$, $c = 10$, and ϵ is a normal random variable with mean of zero and standard deviation of 1. To illustrate the behavior of a single replicate, we highlight one representative trajectory in black. (**B**) The distribution X_t values shown for $1,200$ replicates runs of the same stochastic process as in (A). The mean of X_t is given as a solid red line and the mean $\pm$ the standard deviation of X_t is given by dashed blue lines. Bottom: X_t is plotted against X_{t+1} for two values of t.

Although it is common to talk about a time series being stationary or nonstationary, this is technically a slight abuse of language. Just as the mean and variance are properties of a random variable (and not of any single data point obtained from that random variable), stationarity is a property of a stochastic process (and

not of any single time series produced by that process). This fact is illustrated by comparing the middle and bottom rows of Figure A18. If we examine any one time series from the middle or bottom rows (e.g. the black curves in each), we see that they have essentially the same dynamics (i.e. they are sine waves with the same frequency). However, the process shown in the middle row is covariance-stationary (as shown below), whereas the process shown in the bottom row is not since its mean changes over time.

To see that the middle row of Figure A18 shows a covariance-stationary process, we can show that the mean, variance, and covariance of the process are independent of time:

$$\mathbb{E}[X_2(t)] = \frac{1}{2\pi}\int_0^{2\pi} \cos\left(\frac{t}{3} + \theta_2\right) d\theta_2 = 0$$

$$\text{Var}[X_2(t)] = \frac{1}{2\pi}\int_0^{2\pi} \left(\cos\left(\frac{t}{3} + \theta_2\right)\right)^2 d\theta_2 = \frac{1}{2}$$

$$\text{Cov}(X_2(t), X_2(t+h)) = \frac{1}{2\pi}\int_0^{2\pi} \cos\left(\frac{t}{3} + \theta_2\right)\cos\left(\frac{t+h}{3} + \theta_2\right) d\theta_2 = \frac{1}{2}\cos\left(\frac{h}{3}\right)$$

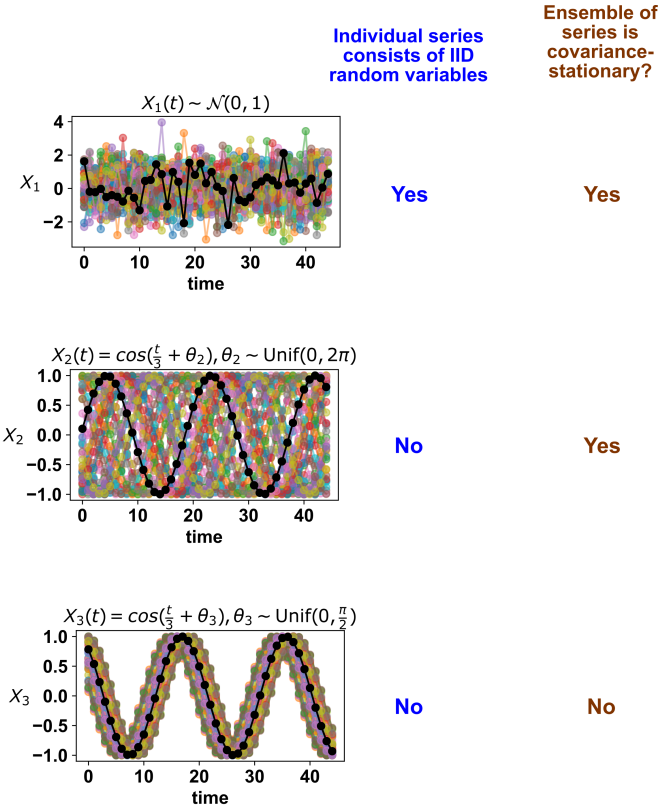


Figure A18: Whether a stochastic process is stationary depends on its entire ensemble of time series. The top panel shows IID standard normal noise. The middle and bottom panel both show sinusoidal curves. Although an individual time series from the middle panel looks similar to that from the bottom panel, only the middle panel shows a covariance-stationary process.

Deterministic processes with many variables may appear stochastic

A deterministic time series from a system with many variables can be approximated as stochastic. This is illustrated below in Figure A19. When we track the trajectory of a particle in a box with 99 other particles (Figure A19 bottom row), the observed trajectory appears random, even though the governing equations of motion are deterministic. In particular, the motion of our particle over each time step can be approximated as having a random component. Note that this flavor of randomness is in general different from the phenomenon called chaos. In chaotic dynamics, each time step needs not be random, but small changes in initial conditions lead to large changes at later times.

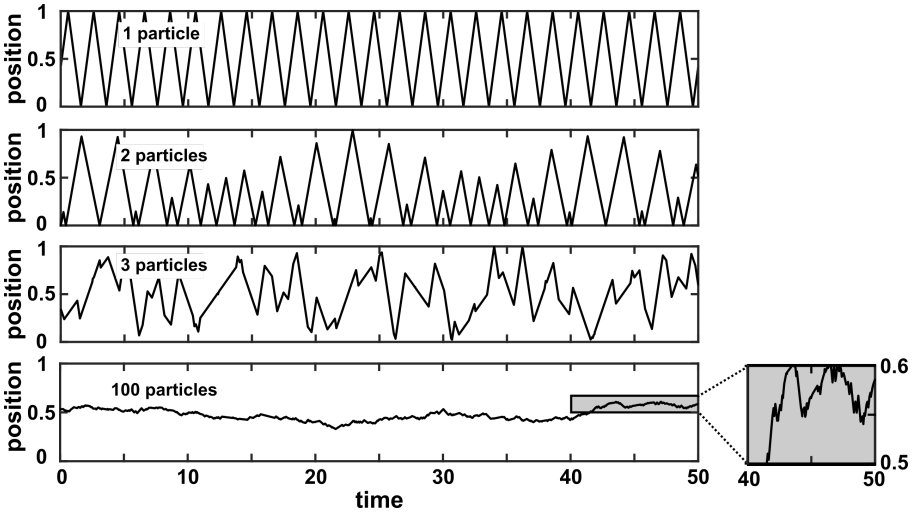


Figure A19: A many-variable deterministic system can be approximated as a stochastic system. The position of a particle in a system of particles bouncing in a 1-dimensional box is plotted over time. In each simulation, particles with radius 0 bounce around in a box with walls of infinite mass placed at positions 0 and 1. Each particle has a mass of 1 and is initialized at a random position between 0 and 1 according to a uniform distribution. Initial velocities are chosen in the following way: The initial velocity of each particle in a box is first randomly chosen from between -1 and 1 according to a uniform distribution. Then, all initial velocities in a given box are multiplied by the same constant to ensure that the total kinetic energy of each box is 0.5. Kinetic energy is conserved throughout the simulation. The simulation then follows the particles as they experience momentum-conserving collisions with one another and with the walls.

A4 State space reconstruction

Difficulty of evaluating the continuity or smoothness of a function with finite or noisy data

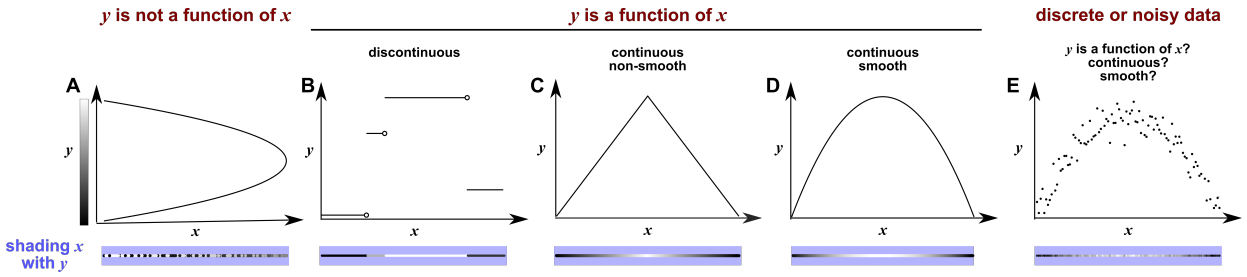


Figure A20: Continuity, smoothness, and the difficulty of evaluating the continuity or smoothness of a function with finite or noisy data. (**A**) y is not a function of x because a single x value can correspond to more than one y value. Here, when we shade x with y value, we randomly choose the upper or the lower y value, leading to bumpy shading, similar to what we might expect to occur in the real world. (**B**) y is a discontinuous function of x. This is because at any "breakpoint" (circle) between two adjacent segments, the limit taken from the left side is different from the limit taken from the right side. Shading x with y generates a "bumpy" pattern. (**C**) y is a continuous function of x, and shading x with y generates a gradual pattern. (**D**) y is a continuous and smooth function of x. A function is smooth if it has a derivative and this derivative is continuous. More generally, the term "smooth" may be used to refer to a function whose kth derivative is continuous for some specified choice of k (as in Takens' theorem, discussed below). Although the function in (**C**) is continuous, it is not smooth since one cannot take a derivative at the sharp point. A smooth function is always continuous. (**E**) With finite and noisy data, shading x with y often generates a bumpy pattern. It is unclear whether y is a function of x, and if yes, whether the function is continuous and/or smooth.

Considerations for selecting delay vector parameters for SSR

To construct delay vectors for SSR, one must choose the delay vector length E and the time delay τ. How does one choose E and τ? In general, detecting a continuous delay map requires that the delay vector length E be high enough so that no two parts of the delay space cross. For example, using $E = 2$ (instead of $E = 3$) to make Figure 4C would have projected the delay space onto 2 dimensions. This would introduce line crossings, which would in turn produce artifactual discontinuities in the shading. On the other hand, the amount of data required to perform SSR inference is said to grow with the delay vector length [18]. SSR is less sensitive to τ, although it is possible to mask a continuous delay map by choosing a "bad" τ. For example, consider what would happen to Figure 4C if we set τ to the period of Z. Since the delay vector is $[Z(t), Z(t-\tau), Z(t-2\tau)]$, setting set τ to the period of Z would force all 3 elements of the delay vector to always be equal. In geometric terms, this would compress the delay space onto a line, destroying the continuous delay map. However, bad choices of τ such as this are rare. Various practical methods are available for systematically choosing E and τ , and delay vectors with variable delays (e.g.

$[Z(t), Z(t-2), Z(t-7)]$) have also been used [92, 23, 118].

Historical notes on the basis of SSR

Takens' celebrated 1981 paper [91] was a major theoretical advance that has inspired a variety of data-driven methods for both causality detection and forecasting (e.g. [119]). Theorem 1 of [91] is reproduced below. Here, the term "map" can be used interchangeably with "function": a map from X to Y sends each point in X to exactly one point in Y.

Takens' theorem (theorem 1 of [91]): Let M be a compact manifold of dimension m. For pairs (ϕ, f), $\phi : M \to M$ a smooth diffeomorphism and $f : M \to \mathbb{R}$ a smooth function, it is a generic property that the map $\Phi_{(\phi,f)} : M \to \mathbb{R}^{2m+1}$, defined by

$$\Phi_{(\phi,f)}(p) = (f(p), f(\phi(p)), ..., f(\phi^{2m}(p))$$

is an embedding; by "smooth" we mean at least C^2.
Author's note: A function is in the class "C^k" if its kth derivative is continuous.

We will attempt to illustrate Takens' theorem using the example in Figure A21. In this system (Figure A21A), once the five initial conditions of $[X, Y, Z, dX/dt, dY/dt]$ are specified, the state space can be visualized in three dimensions (X, Y, Z) (Figure A21C). We can color the trajectory with time (a colored clock-like ring in Figure A21D to highlight the periodic nature of system dynamics). This trajectory is the manifold M in Takens' theorem and is 1-dimensional ($m = 1$) since it is a loop. Takens' theorem then asks us to choose a function ϕ, which we will define as a function that "points into the past". Specifically, ϕ is a function that maps a point p on the manifold M at the current time t to the point q at a previous time $t - \tau$. Similarly, $\phi^2(p)$ would apply ϕ twice and map p at the current time t to the point r at time $t - 2\tau$ (olive in Figure A21B), and $\phi^{-1}(q)$ would map point q at the past time $t - \tau$ to the point p at the current time t (Figure A21C). Note that ϕ and ϕ^{-1}, which are "discrete-time" mappings, are distinct from the differential equations that generated the system dynamics (which are continuous in time; Figure A21A). The term "diffeomorphism" in the theorem means that both this function ϕ and its inverse function (the map from past to present) are smooth (Figure A20).

The next symbol in the theorem is f, which can be viewed as an "observation" function that maps each point on the manifold to a single real number (e.g. in Figure A21E, $f(p) = p_Z$ so that f simply returns the Z coordinate of point p). Takens' theorem then asks us to consider a function Φ that maps a point p at time t on our state space manifold (Figure A21E) to a point in the "delay space". The coordinates of the delay space are given by applying the observation function to point p (which occurs at at time t), point q (at time $t - \tau$), and point r (at time $t - 2\tau$), so that a single point in the delay space is $[p_Z, q_Z, r_Z]$ with respect to a particular time t (Figure A21F). This choice of delay space comes from three earlier choices: First, we consider delayed values of Z since Z is what our observation function f returns; second, since $m = 1$, the

delay space should be of dimension 3 ($= 2m + 1$) per Takens' theorem; third, the delay length of τ comes from our diffeomorphism ϕ. Then, Takens' theorem states that for "most" (technically, "generic") choices of f and ϕ, Φ is an embedding. This means that Φ is diffeomorphic to its image, i.e. the curve in delay space will map smoothly to the manifold M and vice versa [120].

Indeed from Figure A21 (C-F), we can see that for our choice of observation function (i.e. $f = Z(t)$), there is a map from the state space manifold M to the delay space manifold. This is because each dot in the state space manifold corresponds to a single time color (i.e. a point within a period), and each time color corresponds to a single dot in the delay space manifold, and thus, each point in the state space manifold corresponds to a single point in the delay space manifold. Moreover, Φ is continuous because the maps from state space to the time ring and from the time ring to delay space are both continuous. Similarly, we can see that the inverse of Φ, which points from the delay space manifold to the state space manifold is also a continuous map, as guaranteed (generically) by Takens' theorem.

Strikingly, if the observation function is Y, we will no longer have a continuous map from the delay space trajectory (now of Y) to the state space trajectory. This is visualized as "bumpy coloring" in Figure A21H. In fact, we cannot even map the delay space to the time ring or the state space: (p, q, r) and (p', q', r') occupy the same point in the Y delay space, yet correspond to different times within a period (Figure A21B) and thus they correspond to different locations in the state space. Takens' theorem took care of this pathology using the word "generic". That is, Y is not considered a generic observation function here. On the other hand, if we use an observation function based on 95% Y mixed with 5% Z, we get an embedding from the state space to the delay space (Figure A21I-J). This is essentially what the term "generic" means in the context of topology: Although some observation functions do not give you an embedding, these "bad" observation functions can be tweaked just a tiny bit to become "good" ones. Similarly, some choices of ϕ do not work (i.e. $\tau = T$ for this system), but these are exceptions (see Theorem 2 of [120] for what makes a ϕ "generic").

At a conceptual level, SSR causality inference can be performed by shading the delay space of one variable (potential causee) with the contemporaneous value of another variable (potential causer), and inferring a causal link if this shading is continuous. In the example of Figure 4 in the main text, shading delay space of Z with Y generates a continuous pattern, consistent with Y causing Z. On the other hand, shading delay space of Z with W shows a bumpy pattern, consistent with W not causing Z.

Sauer and colleagues [93] later extended Takens' theorem by proving a similar result that is in some ways more general. Theorem 2.5 in [93] is distinct but related to Takens' theorem, and applies to cases that Takens' theorem does not cover, such as fractal spaces. Additionally, [93] replaces the concept of "generic" with a different notion ("prevalence"), which is closer to a probabilistic statement. Cummins et al. [85] then formally connected these results to a notion of potentially causal coupling between dynamic variables.

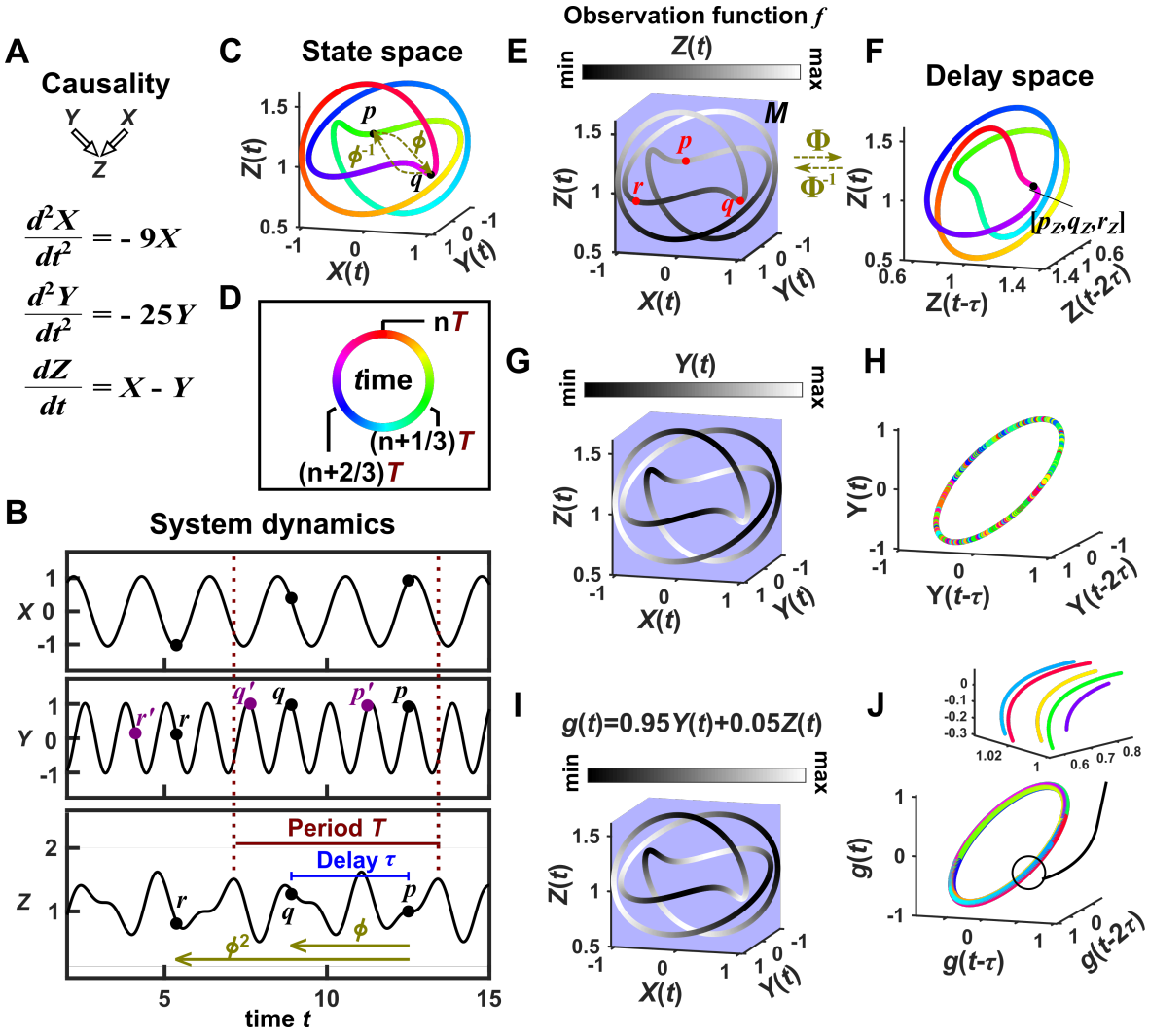


Figure A21: Illustration of Takens' theorem. (**A**) We consider a 3-variable toy system in which X and Y causally influence Z, but Z does not influence X or Y. (**B**) Time series of the three variables. (**C**) We can plot time series data in the state space manifold M. Takens' theorem requires that ϕ, a function that maps a point p at current time t to the point q at a previous time $t-\tau$, and its inverse ϕ^{-1} (from past to current time) are both smooth (C^2: the first and second derivatives of the function exist and are continuous at all times). To mark time progression, we color each point along the trajectory with its corresponding time value where time is represented as a color ring similar to a clock to reflect the periodic nature of system dynamics (**D**). (**E, G, I**) Shading the state space manifold with the observation function (f in Takens' theorem) marked above. (**F, H, J**) Delay space based on the observation function, colored with time. The map Φ in Takens' theorem maps, for example, point p in panel E to point $[p_Z, q_Z, r_Z]$ in panel F. The theorem states that for "generic" observation functions, this map Φ and its inverse Φ^{-1} are both smooth (differentiable). In this example $\tau = 3.6$. In panel J, multiple colors in a region are due to one period wrapping around the delay space multiple times (inset), but the color shading transition is continuous (similar to panel F).

Nonreverting continuous dynamics: Criteria and effects on convergent cross mapping

We first illustrate "nonreverting continuous dynamics", which reflects a nonstationarity pathology for SSR techniques. We then discuss how nonreverting continuous dynamics affects CCM.

We use the phrase "nonreverting continuous dynamics" to describe the following idea: If the X delay space maps continuously to t ("nonreverting" X), and t maps continuously to $Y(t)$ ("continuous" Y), then X delay space will map continuously to $Y(t)$, even if X and Y are causally unrelated (Figure A22A). Figure A22B illustrates this with three causally *independent* time series X and Y. In the top row, the X delay space maps continuously to t and t maps continuously to $Y(t)$, so we get nonreverting continuous dynamics and a continuous delay map from X to Y even though X and Y are independent. In the middle row, the X delay space maps continuously to t , but t does not map continuously to $Y(t)$, so we do not have nonreverting continuous dynamics (i.e. no continuous map from the X delay space to Y). In the bottom row, the X delay space does not map continuously to t. This is because a single delay vector (shown as a cyan dot in the delay space) occurs at multiple times (shown as a repeated cyan line segments whose starting and ending points denote the two values of the delay vector), generating a bumpy pattern similar to Figure A20A. In this case, even though t maps continuously to $Y(t)$, we do not have nonreverting continuous dynamics and we do not get a spurious continuous map from the X delay space to Y.

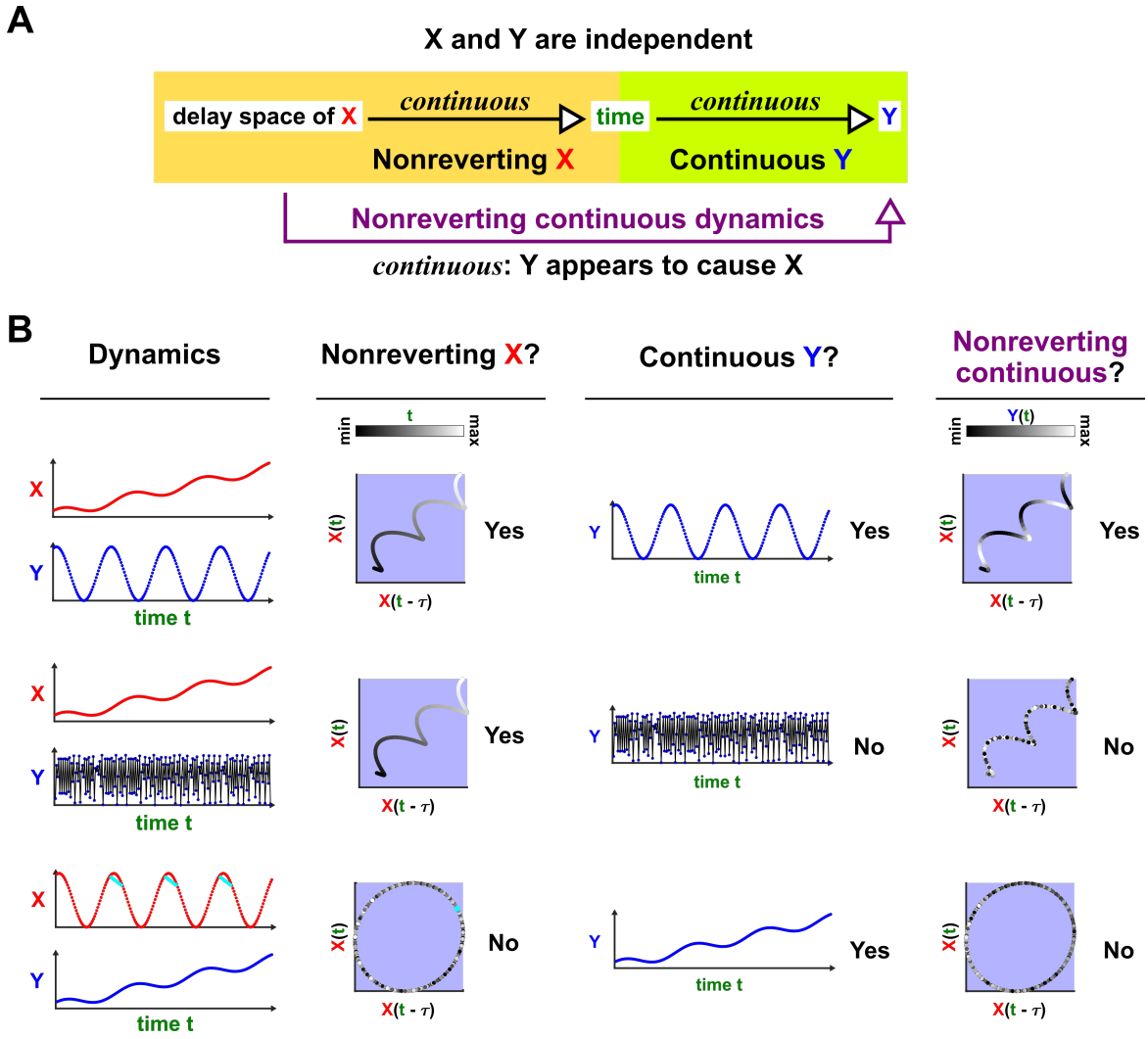


Figure A22: Nonreverting continuous dynamics. (**A**) Nonreverting continuous dynamics. We call X reverting if the delay space of X maps continuously to t (time). We call Y "continuous" if $Y(t)$ is a continuous function of t. If X is reverting and Y is continuous then we say that the pair of time series X, Y has nonreverting continuous dynamics. (**B**) Examples. In each row, X and Y are causally independent. Leftmost column: Dynamics. Each red or blue dot (visible upon zooming in on some of the charts) represents a single time point. Second column: Looking for a continuous map from the delay vectors of X (X delay space) to t, i.e. nonreverting X dynamics. Third column: Looking for a continuous map from t to Y by assessing whether Y at nearby times share similar values. Since the data occur at discrete times, the standard definition of continuity does not naturally apply, so "continuous Y" really means "highly autocorrelated". Fourth and final column: the presence or absence of "nonreverting continuous dynamics". With nonreverting continuous dynamics, there is a continuous map from the X delay space to Y, and thus Y appears to cause X even though X and Y are causally independent.

Nonreverting continuous dynamics interferes with CCM causal discovery. Although one could attempt to mitigate the nonstationarity problem by interspersing training and testing data before quantifying cross map skill [24] (Figure A23, Column 4), we find that this approach leads to false positive errors (Figure A23 bottom row). In contrast, the alternative (not interspersing training and testing data) can lead to false

negative errors (Figure A23, third row). Thus, the ability to correctly infer causality with CCM is vastly reduced when data exhibit nonreverting continuous dynamics.

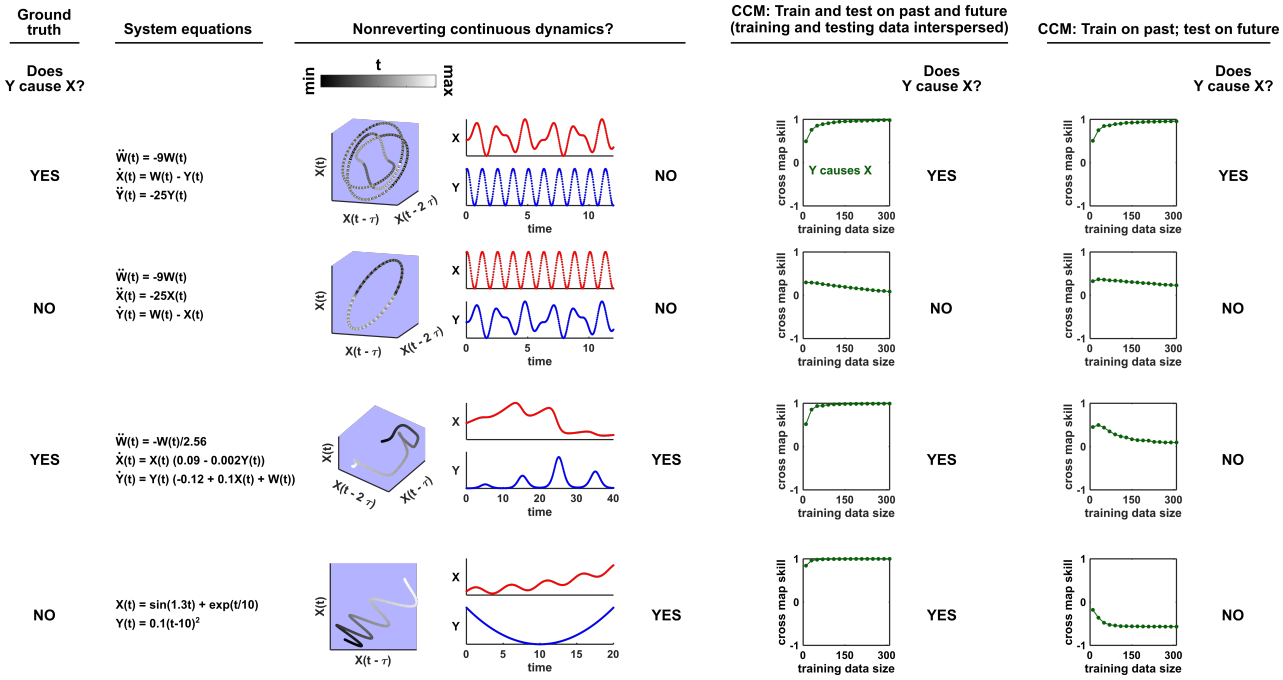


Figure A23: Nonreverting continuous dynamics reduce the ability of CCM to correctly infer causality. Each row represents a system where Y may or may not causally influence X (Column 1). Column 2: Governing equations. Column 3: Checking for nonreverting continuous dynamics as in Figure A22. The top two rows do not have nonreverting continuous dynamics since there is no continuous map from the delay space of X to time. The bottom two rows have nonreverting continuous dynamics. Columns 4 and 5: Results of CCM where training and testing data are interspersed or when we train on the past and test on future. In the bottom two rows, CCM suffers false negative or false positive errors depending on the analysis details (e.g. whether training and testing data are interspersed).

The prediction lag test: Intuition and some failure modes

State space reconstruction methods suffer false positive errors in the presence of synchrony [18]. This occurs when "the dependence of the dynamics of the forced variable on its own state is no longer significant" [18]. Ye et al. proposed a test in an effort to solve this problem [96]. Their procedure relies on finding mappings from the delay vector $[X(t), X(t-\tau), X(t-2\tau)...X(t-(E-1)\tau)]$ to $Y(t+l)$, where E is the delay vector length, τ is the time lag, and l is a key variable known as the "prediction lag". They then examine how the cross map skill (Figure 6B) varies with the prediction lag. According to this technique, if the cross map skill is maximized at a positive prediction lag ($l > 0$), then the putative causality is spurious and arose from, for example, strong unidirectional forcing. The reasoning is that if the causee were to predict the future of the causer, then causation would appear to flow backward in time, which is nonsensical. On the other hand, if the highest quality mapping occurs at a non-positive prediction lag ($l \leq 0$), then we have further evidence

that the detected causality is real and not spurious [96].

We find that while this test correctly distinguishes between real and spurious causal signals at some times, at other times it does not. Within each row of Figure A24, we examine a different system and ask whether Y causes X according to: (1) the ground truth model, (2) our visual continuity test, (3) a CCM cross map skill test (without the prediction lag test), and (4) the prediction lag test.

In rows 1 and 2 of Figure A24, the prediction lag test performs well, overturning the results of the visual continuity and CCM tests when apparent causality is spurious (row 1), and agreeing with the continuity and CCM tests (row 2) when apparent causality is real (modified from [96] Eq. 2). However in row 3, the prediction lag test dismisses a true causal link as spurious. Moreover, when we apply the prediction lag test to a system with a periodic putative driver (Figure A24 row 4), we find that cross map skill is a periodic function of the prediction lag. While this result is what we would expect mathematically, its causal interpretation is unclear. The fifth row of Figure A24 is an extreme case of strong forcing, where $Y(t+1)$ is a function of $X(t)$, but not $Y(t)$. Here the prediction lag test gives a false positive error. In the bottom row, X and Y do not interact, but are both driven by a common cause W with different lags. Specifically, $W(t)$ exerts a direct effect on $Y(t+1)$ and on $X(t+3)$. Thus, Y receives the same information as X, but at an earlier time, analogous to Figure 3iii. Consistent with this, delay vectors of X predict past values of Y better than future values of Y. Thus, the prediction lag test produces a false positive error.

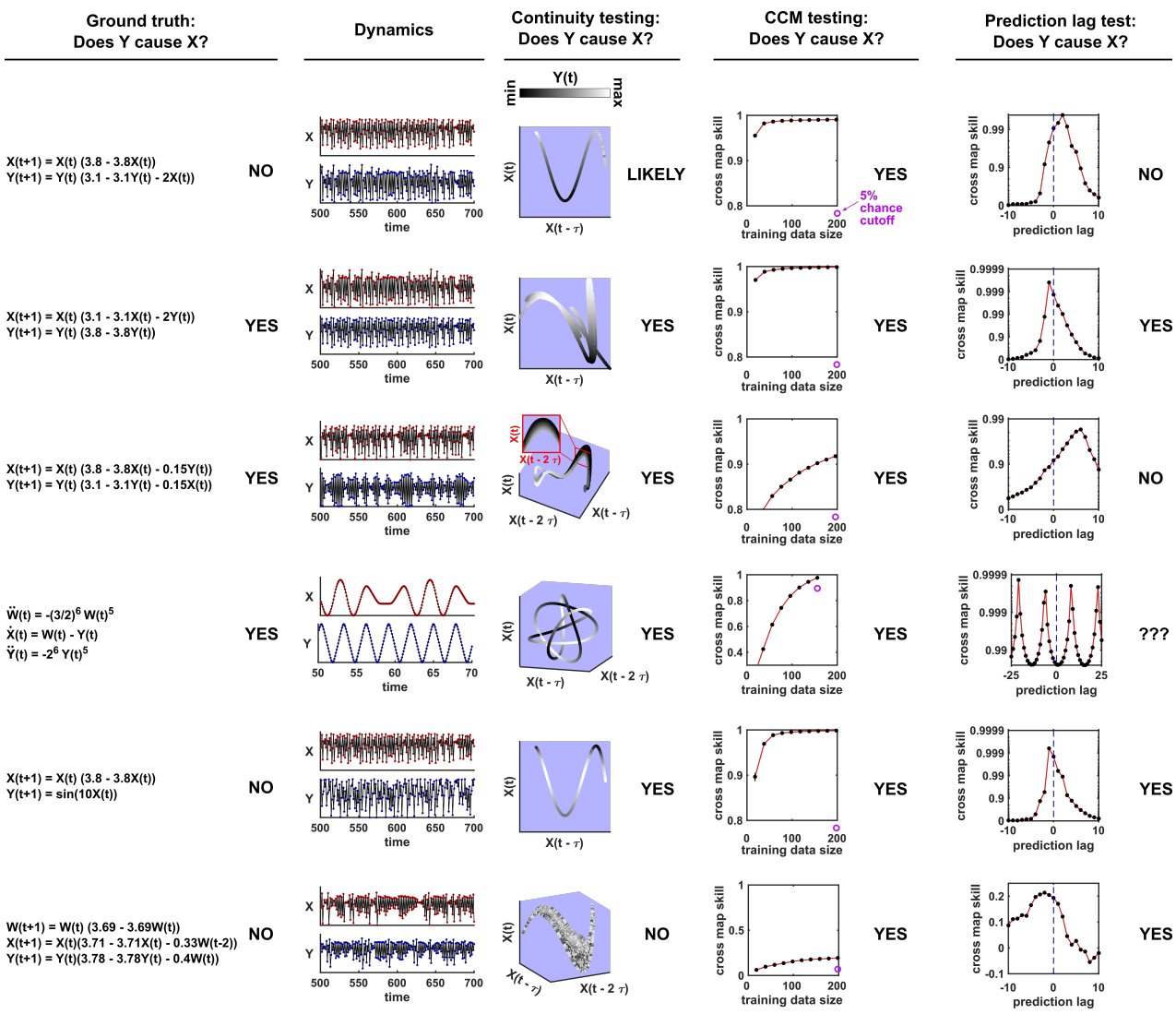


Figure A24: Comparison of visual continuity testing, cross map skill testing, and prediction lag testing in causal discovery. Each row represents a two-variable or three-variable system where Y may or may not causally influence X. The leftmost column shows the equations and ground truth causality. The second column shows a sample of X and Y dynamics. Red and blue dots represent X and Y values, respectively; black lines connecting the dots serve as a visual aid. The third column shows visual continuity testing and causal interpretation. We write "likely" in the top row because the map from X delay space to Y appears to have some small bumps on the right side of the plot. The fourth column shows cross map skill testing (without the prediction lag test) and causal interpretation. Black dots show cross map skill. Open purple dots show the 5% chance cutoff at the maximum library size according to random phase surrogate data testing (see Appendix A5), or are placed below the horizontal axis if the 5% chance cutoff is below the plot. In all systems Y appears to cause X according to cross map skill testing since cross map skill is positive, increases with training data size, and is significant according to the surrogate data test. The rightmost column shows the prediction lag test and causal interpretation.

Ye et al. [96] applied the prediction lag test to 500 systems with the same form as in the third row of Figure A24 but with randomly chosen parameters. They found that within the parameter range they sampled, false negative errors as in Figure A24 do occur, but such errors are rare. We repeated the randomized numerical

experiment from [96] for both the original parameter range of [96] (Figure A25B, "friendly" parameter regime) and a second parameter range of the same volume in parameter space (Figure A25B, "pathological" parameter regime). In this pathological parameter regime, false negative errors occur in the overwhelming majority of cases.

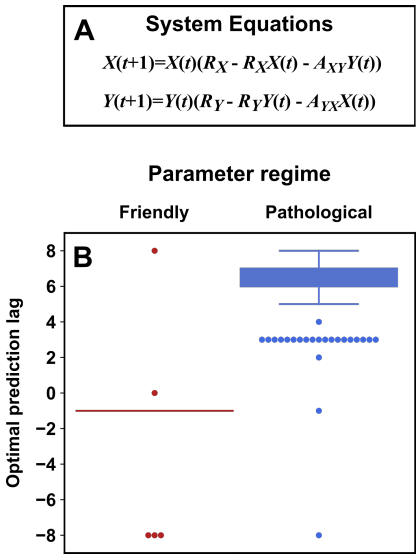


Figure A25: Parameters within a "pathological" regime almost always cause the prediction lag test to erroneously reject a true causal link. (**A**) System equations. For both "friendly" and "pathological" regimes, initial conditions $X(1)$ and $Y(1)$ were independently and randomly drawn from the uniform distribution between 0.01 and 0.99 ("$Unif(0.01, 0.99)$"), and R_X was drawn from $Unif(3.7, 3.9)$. R_Y was drawn from $Unif(3.7, 3.9)$ ("friendly") or $Unif(3.1, 3.3)$ ("pathological"). A_{XY} and A_{YX} were independently drawn from $Unif(0.05, 0.1)$ ("friendly") or $Unif(0.15, 0.2)$ ("pathological"). (**B**) Boxplots show the optimal prediction lag when using delay vectors made from X to predict values of Y in 250 systems with parameters selected randomly as described just now. In the ground truth model for this system, Y exerts a causal influence on X. In the "friendly" parameter regime, the optimal prediction horizon is negative, correctly indicating that Y does indeed cause X. In the "pathological" regime, the optimal prediction horizon is positive, and so the prediction lag test would wrongly conclude that Y does not cause X. In the friendly regime the "box" is shown as a line because the vast majority of trials had the same optimal prediction lag of -1.

A5 Detailed methods

Methodological details for Figure 2

For panel B, we simulated the random walk system

$$X(t+1) = X(t) + \epsilon(t)$$

where $\epsilon(t)$ terms were drawn independently from a normal distribution with mean of 0 and standard deviation of 1. We simulated this system from the initial condition of $X(1) = 0$ through 999 subsequent

steps. For panel C, we simulated the autoregressive system

$$X(t+1) = 0.75X(t) + 10 + \epsilon(t)$$

where $\epsilon(t)$ terms were again drawn independently from a normal distribution with mean of 0 and standard deviation of 1. We simulated this system from the initial condition of $X(1) = 40$ for 1999 subsequent steps. We only used the final 1000 steps for computing the correlation between two time series.

To compute the significance of the Pearson correlation between two time series, we used surrogate data generated by either permutation or the random phase procedure. Permutation surrogate time series were generated by randomly shuffling data. Random phase surrogate time series were generated by Ebisuzaki's random phase method [42] as implemented in the rEDM (version 1.5) function make_surrogate_data. For a pair of time series $[X_1(1), X_1(2), ..., X_1(1000)], [X_2(1), X_2(2), ..., X_2(1000)]$, we first computed the Pearson correlation $\hat{\rho}$ between the two time series. We then replaced the X_2 values with surrogate time series and recomputed the Pearson correlation as $\tilde{\rho}$. We computed this shuffled correlation 9999 times (permutation) or 499 times (random phase) to get a null distribution $[\tilde{\rho}_1, \tilde{\rho}_2, ..., \tilde{\rho}_n]$. Following [47], we computed the p value as

$$p = (N_{stronger} + 1)/(N_{surr} + 1) \tag{5}$$

where N_{surr} is the number of surrogates, $N_{stronger}$ is the number of surrogate correlations $\tilde{\rho}$ whose magnitude was greater than or equal to the magnitude of the original correlation $\hat{\rho}$, and the "+1" terms account for $\hat{\rho}$.

Methodological details for Figure 4

The system of equations was numerically integrated using the ode45 method in Matlab from $t = 0$ to $t = 200$ in time steps of 0.03, and plotted in the delay space Z with $\tau = 3.6$. The initial condition for all state variables (V, W, X, Y, Z, $\frac{dX}{dt}$, $\frac{dY}{dt}$, and $\frac{dV}{dt}$) was 1. For panel F, measurement noise was added to $Y(t)$. Specifically, noisy data were generated as:

$$Y^{obs}(t) \sim \text{Unif}\left(Y(t) - 3^{1/2}\left(0.15\Delta_Y\right), Y(t) + 3^{1/2}\left(0.15\Delta_Y\right)\right)$$

where $\text{Unif}(a, b)$ is a uniform random variable bounded by a and b, and Δ_Y is the difference between the maximum and minimum values of $Y(t)$ between $t = 0$ and $t = 200$. These noise parameters are chosen so that $Y^{obs}(t)$ is centered at $Y(t)$ and has a standard deviation of $0.15\Delta_Y$.

Methodological details for Figure 5

The dynamics in the top row of Figure 5 were generated from the equations:

$$X(t) = \sin(t) + 0.5t$$
$$Z(t) = 0.1(t - 10)^2$$

This continuous-time system was discretized from $t = 1$ to $t = 20$ on an evenly spaced grid of 400 data points for visualizing delay spaces where the time delay is 50 time points (i.e. $\tau = 50(20 - 1)/(400 - 1)$).

The dynamics in the second row of Figure 5 were generated from the equations:

$$X(t + 1) = X(t)(3.61 - 3.61X(t))$$
$$Z(t + 1) = Z(t)(3.61 - 3.61X(t))$$

with initial conditions of $X(1) = 0.4$ and $Z(1) = 0.7$. For this system, $\tau = 1$ and $t = 1, 2, ..., 2000$ were used to make the plots of delay spaces.

The dynamics in the third row of Figure 5 were generated from the equations:

$$\frac{dX^2}{dt} = -X(t)$$
$$\frac{dZ^2}{dt} = -25Z(t)$$

with initial conditions of $X(1) = X'(1) = Z(1) = Z'(1) = 1$. For this system, $\tau = 0.9$ was used for delay spaces. This continuous-time system was numerically integrated using the ode45 method in Matlab from $t = 0$ to $t = 13.998$ on a grid of 4667 evenly-spaced time points for plotting dynamics, and time points $t = 0.003$ through $t = 7.698$ were used for visualizing delay spaces.

The dynamics in the bottom row of Figure 5 were generated from the classic Lorenz attractor equations:

$$\frac{dX}{dt} = -10X(t) + 10Y(t)$$
$$\frac{dY}{dt} = 28X(t) - Y(t) - X(t)Z(t)$$
$$\frac{dZ}{dt} = -\frac{8}{3}Z(t) + X(t)Y(t)$$

with initial conditions of $X(0) = Y(0) = Z(0) = 1$. A delay of $\tau = 0.14$ was used to make delay spaces. This continuous-time system was numerically integrated using the ode45 method in Matlab from $t = 0$ to $t = 399.98$ on an evenly spaced grid of 5715 data points for visualizing delay spaces.

Methodological details for Figure 7

Ground truth model and data generation

We used the ground truth model:

$$S_1(t+1) = \max\left(0, S_1(t)\left(1.2 - 0.1S_1(t) + D_1(t)\right) + \epsilon_{p1}(t)\right)$$
$$S_2(t+1) = \max\left(0, S_2(t)\left(1.1 - 0.2S_2(t) + D_2(t) + 0.3S_1(t)\right) + 2.5\epsilon_{p2}(t)\right)$$

$S_1(t)$ and $S_2(t)$ represent the population sizes of species 1 and 2 at time t. $D_1(t)$ and $D_2(t)$ are the values of periodic drivers at time t. Specifically, in both the two-driver and one-driver cases:

$$D_1(t) = 0.05\sin\left(t + \phi_1\right) + 0.05\sin\left(\frac{5t}{6} + \phi_1\right)$$

In the two-driver case:

$$D_2(t) = 0.1\sin\left(\frac{t}{\sqrt{10}} + \phi_2\right)$$

Conversely, in the one-driver case $D_2(t) = 0$. The process noise terms $\epsilon_{p1}(t)$ and $\epsilon_{p2}(t)$ are both IID normal random variables with mean of 0 and with shared standard deviation σ_p. Specifically, for any pair of times $t_1 \neq t_2$, $\epsilon_{p1}(t_1)$ and $\epsilon_{p1}(t_2)$ are independent, and similarly for ϵ_{p2}. Also, all values $\epsilon_{p1}(1), \epsilon_{p1}(2), \cdots$ are independent of all values $\epsilon_{p2}(1), \epsilon_{p2}(2), \cdots$. At the beginning of each replicate simulation, the phases ϕ_1 and ϕ_2 are independently assigned a random number from a uniform distribution between 0 and 2π, and do not change with time.

To generate data without measurement noise, we simulated this system for $t = 1, 2, ..., 400$ with the initial conditions $S_1(1) = 2$; $S_2(1) = 4.5$. We used the final 200 time points for inference to help ensure that the system had reached equilibrium behavior.

We also introduced additive measurement noise to simulate instrument uncertainty:

$$S_1^{obs}(t) = S_1(t) + \epsilon_{m1}(t)/1.5$$
$$S_2^{obs}(t) = S_2(t) + \epsilon_{m2}(t)$$

where S_1^{obs} and S_2^{obs} represent the observed values (i.e. noisy measurements) of S_1 and S_2. $\epsilon_{m1}(t)$ and $\epsilon_{m2}(t)$ are also IID normal random variables with mean of 0 and standard deviation σ_m. The tables in Figure 7D are generated by varying σ_p from 0 to 8 and varying σ_m from 0 to 1.

Causal analysis using Granger causality and CCM

For each combination of σ_m and σ_p (the standard deviation of measurement noise and process noise, respectively), we generated 1000 time series for S_1 and S_2 as described above. For each replicate pair of time series, we used Granger causality and CCM to infer whether S_1 causes S_2 (it does) and whether S_2 causes S_1 (it does not).

Granger causality inference

We used the multivariate Granger causality Matlab package (MVGC, [20]). We used the following settings:

- regmode = 'OLS' (We fit the autoregressive model by the ordinary least squares method).
- icregmode = 'LWR' (We determined the information criterion using the LWR algorithm. This is the default setting).
- morder = 'AIC' (We used Akaike information criterion to determine the number of lags in the autoregressive model).
- momax = 50 (We used a maximum of 50 lags in the autoregressive model).
- tstat = '' (We used Granger's F-test for statistical significance. This is the default setting).

We inferred the presence of a causal link if the p-value was less than or equal to 0.05. We inferred no causal link otherwise. When σ_m and σ_p were both 0, the MVGC package (correctly) exited with an error on most trials. We reported this as "unsuitable data" in Figure 7D & E.

When σ_m and σ_p are both 0, the inferred spectral radius of the stochastic process is close to 1, and the MVGC routines can be prohibitively slow (i.e. when running 1000 trials, the program would hang at an early stage for hours). In this case, the authors note that switching from the package's default single-regression mode to an alternative dual-regression mode may improve runtime [20]. We thus switched to the dual-regression mode when the spectral radius was between 0.9999 and 1 (a spectral radius of 1 or more causes an error). This fix had no effect on benchmark results as long as at least one of σ_m and σ_p was not 0.

Convergent cross mapping

Convergent cross mapping looks for a delay map from X to Y. That is, CCM looks for a map from $[X(t), X(t-\tau), X(t-2\tau), ..., X(t-(E-1)\tau)]$ to $Y(t)$. Thus in order to apply CCM one needs to choose

the delay τ and the vector length (dimension of the delay space) E. E and τ should ideally be "generic" in the sense of Takens' theorem: we want to avoid line-crossing (such as the symbol "∞") in the delay space, because otherwise, Φ^{-1} in Figure A21 does not exist. There are different ways to do this, but no method is obviously the best ([92, 23]).

Following [92] and [18] we chose τ and E to maximize univariate one-step-ahead forecast of the putative causee X. That is, for $X(n)$, we try to predict $X(n+1)$ using the simplex projection method by finding delay vectors in the training data of X that are most similar to $[X(n), X(n-\tau), X(n-2\tau), ..., X(n-(E-1)\tau)]$, and take weighed average of their X values 1 step in the future (i.e. Figure 6A where $X = Y$ and the prediction lag is 1). If the delay space has a line crossing, then at the cross-point, a one-step-ahead forecast may have more than one possible outcome and thus perform poorly. In more detail, we made one-step-ahead forecasts within the time range 201-400 (we did not use time range 1-200 to avoid transient dynamics). As per the field standard, we used leave-one-out cross-validation to do simplex projection. That is, when making a forecast for a time t, we used all times within 201-400 other than t as training data (200 time points). We performed a grid search, varying τ from 1 to 6 and varying E from 1 to 6. We then used the combination of τ and E that maximized the forecast skill (the Pearson correlation between forecasts and true values) for subsequent CCM analysis. Additionally, following [18], if the optimal combination of τ and E failed to give a significantly positive forecast skill, we did not report CCM results for that trial and reported the trial as "unsuitable data". To test whether forecast skill is "significantly positive", we ask whether it is robust to small changes in the training dataset. To do so, we used a naive bootstrap approach to create different versions of training libraries composed of randomly chosen delay vectors (sampling with replacement: some vectors may not be sampled and others may be sampled more than once) from the original training data using the 'random_libs' setting in the rEDM (version 1.5) ccm method. The training library size (the number of delay vectors in the library) was chosen to be 200. We then calculated forecast skills with 300 such libraries and considered the forecast skill "significant" if at least least 95% gave a forecast skill greater than 0.

Having chosen τ and E, we checked three CCM criteria to infer causality (criteria 1-3 in Figure 6) using rEDM version 1.5. We did not use the fourth criterion (the prediction lag test) since its interpretation is unclear for periodic systems (Figure A24). For all three criteria, we used the same cross-validation setting that we used to choose τ and E. The first CCM criterion is that cross map skill is greater than 0. Thus, we computed cross map skill using the maximum possible number of distinct delay vectors $(200 - (E-1)\tau)$ and compared this value to 0.

The second CCM criterion is that the cross map skill from causee to causer with real data must be greater than the cross map skill when the putative causer is replaced with surrogate data. To test this criterion, we first computed cross map skill using the same training and testing time points as before to obtain a single

cross map skill value. We then repeatedly (1000 times) computed cross map skill in the same way, but now with the putative causer time series replaced with random phase surrogate data. Random phase surrogate data were generated by Ebisuzaki's method as implemented in the rEDM function make_surrogate_data. We then computed the p-value according to Eq. 5. A putative causal link would pass this criterion if the p-value was less than or equal to 0.05.

The third CCM criterion is that cross map skill increases with more training data. Following [92], we again used a naive bootstrap approach to test for this criterion. Specifically, we computed the cross map skill with a training library composed of randomly chosen delay vectors sampled with replacement from the original training data time points. We used either a large library with $200 - (E-1)\tau$ available training vectors as used previously, or a small library with 15 training vectors. For each of 1000 bootstrap trials, we compared the cross map skill from a randomly chosen small library to the cross map skill from a randomly chosen large library. We said that the cross map skill increased with training data if the cross map skill of the large library was greater than that of the small library in at least 95% of the 1000 bootstrap trials.

For "alternative" CCM testing, we only changed how the third CCM criterion (cross map skill increases with more training data) were tested. Here, instead of using the bootstrap test of [92], we tested the third CCM criterion using Kendall's τ test as suggested in [103]. To do this, we varied the library size from a minimum of 15 vectors to the the maximum library size ($200 - (E-1)\tau$), in increments of 3 vectors. For each library size, we computed cross map skill using 50 libraries randomly sampled without replacement (e.g. the 50 libraries would be identical at the maximal library size). We then computed the median cross map skill for each library size. Finally we ran a 1-tailed Kendall's τ test for a positive association between library size and median cross map skill. We used the function stats.kendalltau from the Python package SciPy to compute a 2-tailed p-value, and then divided this p-value by 2 to get a 1-tailed p-value. We said that cross map skill increased with training data if the τ statistic was positive and the 1-tailed p-value was ≤ 0.05.

Methodological details for Figure A10

The original subpopulation distributions are normal distributions with standard deviation of 10 and mean of 100 (male) or 130 (female). Each sampling plot shows 300 samples randomly drawn from the appropriate mixture distribution.

Methodological details for Figure A21

To generate data for panels C-J, the system of panel A was numerically integrated using the ode45 method in Matlab with a time step of 0.005 and with the initial condition that X, Y, Z, $\frac{dX}{dt}$, $\frac{dY}{dt}$ were all set to 1 at

$t = 0$. Panels C , E, F, G and I show data from a single period. For panel H the system was integrated for about 5 periods to more clearly visualize the lack of a continuous delay map. For panel J, inset the system was integrated for 1 period for the main figure and about 12 periods (to increase sampling density) for the inset. This allows us to better see the separated legs of the curve upon zooming in. Panels C, D, F, H, and J were colored $\mathrm{mod}(t, T)$. That is, they were colored by the remainder of t (time) after dividing by T (here $T = 2\pi$). $\tau = 3.6$ was used for all delay spaces.

Methodological details for Figure A22

All systems were discretized from $t = 1$ to $t = 20$ on an evenly spaced grid of 200 points for visualizing delay spaces.

The dynamics in the top row were generated from the equations:

$$X(t) = \sin(t) + 0.5t$$
$$Y(t) = \sin(1.3t)$$

A delay time of 12 time indices (i.e. $\tau = 12(20-1)/(200-1)$) was used for constructing delay spaces.

The dynamics in the second row were generated from the equations:

$$X(t) = \sin(1.3t)$$
$$Y(t) = Y(t-\delta)(3.77 - 3.77 * Y(t-\delta))$$

with $\delta = (20-1)/(200-1)$ and the initial condition $Y(1) = 0.3$. A delay time of 25 time indices (i.e. $\tau = 25(20-1)/(200-1)$) was used for constructing delay spaces.

In the third row the dynamics are identical to the first row, except that X and Y are switched, and $\tau = 25$ time indices was used for constructing delay spaces.

Methodological details for Figure A23

Top row: For this system, we used the initial conditions $W(0) = \dot{W}(0) = X(0) = Y(0) = \dot{Y}(0) = 1$. We numerically integrated this system using ode45 in Matlab with a time step of 0.03. We composed delay vectors of length $E = 3$ with a delay of $\tau = 3.6$. We visualized the delay space using data from $t = 0$ through $t = 29.97$ (time indices 1-1000). For CCM with temporally separate training and testing sets, we used data

from $t = 0$ through $t = 14.97$ (time indices 1-500) for training data and data from $t = 15$ through $t = 29.97$ (time indices 501-1000) for testing. Specifically, in the rEDM (version 0.7.2) ccm method we set the lib argument to "c(1, 500)" and set the pred argument to "c(501, 1000)". We used rEDM version 0.7.2 for this analysis because we found that it more easily produced distinct training and test sets than later versions (on a computer running MacOS 11.6 and R version 4.0.2). For CCM with temporally interspersed training and testing sets, we set both the lib and pred arguments to "c(1,1000)". This setting instructs rEDM to use leave-one-out cross-validation.

Second row: Ground truth data generation and analysis were the same as in the top row, except that the roles of X and Y were swapped.

Third row: For this system, we used the initial conditions $W(0) = 0, \dot{W}(0) = 1/1.6, X(0) = 1.3, Y(0) = 1.5$. We numerically integrated this system using ode45 in Matlab with a time step of 0.1. We visualized the delay space using data from $t = 0$ through $t = 40$ (time indices 1-401). We used the delay vector parameters ($E = 3, \tau = 3.0$). For CCM with temporally separate training and testing sets, we used data from $t = 0$ through $t = 19.9$ (time indices 1-200) for training data and data from $t = 20$ through $t = 39.9$ (time indices 201-400) for testing. For CCM with temporally interspersed training and testing sets, we used cross-validation over the entire range $t = 0$ through $t = 39.9$.

Bottom row: We discretized this system with a time step of 0.05. We visualized the delay space using data from $t = 0$ through $t = 20$ (time indices 1-401). We used the delay vector parameters ($E = 2, \tau = 2.5$). For CCM with temporally separate training and testing sets, we used data from $t = 0$ through $t = 9.95$ (time indices 1-200) for training data and data from $t = 10$ through $t = 19.95$ (time indices 201-400) for testing. For CCM with temporally interspersed training and testing sets, we used cross-validation over the entire range $t = 0$ through $t = 19.95$.

For convergent cross mapping, we used the same τ and E as for visualizing delay spaces (see above). "Training data size" on the horizontal axis is the number of delay vectors in the training library. Each dot in these CCM plots represents the average forecast skill over 300 randomly chosen libraries. Error bars represent the 95% confidence interval as calculated by the bias-corrected and accelerated bootstrap (1000 bootstraps) as implemented in Matlab's bootci function. Error bars are the same color as the dots and so are not visible when they fit inside the dots.

Methodological details for Figure A24

Top row: For this system, we used the initial conditions $X(1) = 0.2, Y(1) = 0.4$ and composed delay vectors of length $E = 2$ with a delay of $\tau = 2$. We visualized the delay space using data from time points 501-2000.

We used points 801-1000 for training data and points 1001-2000 for testing cross map predictions.

Second row: For this system, we used the initial conditions $X(1) = 0.4, Y(1) = 0.2$ and the delay vector parameters $(E = 2, \tau = 1)$. We visualized the delay space using data from time points 501-2000. We used points 801-1000 for training data and points 1001-2000 for testing cross map predictions.

Third row: For this system, we used the initial conditions $X(1) = 0.2, Y(1) = 0.4$ and the delay vector parameters $(E = 3, \tau = 2)$. We visualized the delay space using data from time points 501-2000 (time points 1-6 $\times 10^5$ for the zoomed-in inset). We used points 801-1000 for training data and points 1001-2000 for testing cross map predictions.

Fourth row: For this system, we used the initial conditions $W(0) = Y(0) = 0$ and $X(0) = \dot{W}(0) = \dot{Y}(0) = 1$. We numerically integrated this system using ode45 in Matlab with a time step of 0.1. We visualized the delay space using data from $t = 50.1$ through $t = 200$ (time indices 501-2000). We used the delay vector parameters $(E = 3, \tau = 7.2)$. We used data from $t = 70.1$ through $t = 100$ (time indices 701-1000) for training data and data from $t = 100.1$ through $t = 200$ (time indices 1001-2000) for testing cross map predictions.

Fifth row: For this system, we used the initial conditions $X(1) = 0.2, Y(1) = 0$ and composed delay vectors of length $E = 2$ with a delay of $\tau = 2$. We visualized the delay space using data from time points 501-2000. We used points 801-1000 for training data and points 1001-2000 for testing cross map predictions.

Sixth row: For this system, the "initial" conditions specified the first 3 timepoints since we included a lag of 3. Thus, for $k = 1, 2, 3$, $W(k) = 0.2$, $X(k) = 0.4$, and $Y(k) = 0.3$. We composed delay vectors of length $E = 3$ with a delay of $\tau = 1$. We visualized the delay space using data from time points 501-2000. We used points 801-1000 for training data and points 1001-2000 for testing cross map predictions.

For convergent cross mapping (in rEDM version 0.7.2), we used the same τ and E as for visualizing delay spaces. The training data size is the number of delay vectors in the training library. For the plots in the fourth column, we chose 300 random libraries of training delay vectors with variable training data size, and used the standard prediction lag of 0. Delay vectors were chosen without replacement. Note that at large training data size, some or all of the 300 random libraries can be identical. Each dot in these CCM plots represents the average forecast skill over all 300 randomly-chosen libraries. Error bars represent the 95% confidence interval as calculated by the bias-corrected and accelerated bootstrap (1000 bootstraps) as implemented in Matlab's bootci function. Error bars are the same color as the dots and so are not visible when they fit inside the dots.

In all rows, the cross map skill for the putative causer Y was greater than for at least 95% of random phase surrogate time series (purple dot). The 5% cutoff value was computed for the maximum library size (156 for row 4 and ~ 200 for all other rows) by running the CCM procedure after replacing the putative causer

Y with 500 random phase surrogate time series generated using the rEDM function make_surrogate_data.

For the plots in the fifth column we used the full library contained within the training data window (156 delay vectors for row 4 and ~ 200 for all other rows) and varied the prediction lag. We did not use random libraries for these plots.

Methodological details for Figure A25

To generate randomized parameter sets, we randomly selected R_X, R_Y, A_{XY} and A_{YX} from uniform distributions. We also randomly selected the initial conditions $X(1)$ and $Y(1)$ from uniform distributions. To make systems in the "friendly" parameter regime, we drew R_X and R_Y independently from the range $3.7 - 3.9$, we drew A_{XY} and A_{YX} independently from the range $0.05 - 0.1$, and we drew $X(1)$ and $Y(1)$ independently from the range $0.01 - 0.99$. These are the same parameters used in the randomized numerical simulations of [96]. Next, to make systems in the "pathological" parameter regime, we drew R_X from the range $3.7 - 3.9$, we drew R_Y from the range $3.1 - 3.3$, we drew A_{XY} and A_{YX} independently from the range $0.15 - 0.2$, and we drew $X(1)$ and $Y(1)$ independently from the range $0.01 - 0.99$. For both parameter regimes we randomly chose 250 sets of parameters and ran the system for 3000 time points. Occasionally a randomly chosen system would leave the basin of attraction and reach large values, represented on the computer as positive or negative infinity, or "not a number". When this occurred, we discarded the data and resampled parameters.

To apply CCM (in rEDM version 0.7.2) on each system, we generated a training library of delay vectors of X by randomly selecting 200 vectors from among time points 100-2000. We then evaluated cross map skill from delay vectors of X to values of Y at points 2001-3000. Following [96], we used delay vectors of length $E = 2$ and a delay duration of $\tau = 1$. We evaluated cross map skill with a prediction horizon of -8 through 8.

Part II

A rigorous and versatile statistical test for correlations between time series

A version of this part is available on the bioRxiv preprint server with the title "An exactly valid and distribution-free statistical significance test for correlations between time series."

1 Introduction

Researchers routinely look for correlations between variables to identify potentially important relationships, or to use as a starting point for downstream modeling and experiments. In fields such as climatology, ecology, and physiology, data are often collected as time series, and so correlations between time series are common.

Interpreting a correlation between time series can be challenging because it is easy to obtain a seemingly high correlation between two time series that have no "meaningful" relationship [121, 122]. For example, the population densities of replicate exponentially-growing bacterial cultures may be correlated over time, but this correlation is driven by a temporal trend rather than any causally meaningful relationship. To avoid spurious correlations, it helps to distinguish between the concepts of "correlation" and "dependence", and how each relates to causation. In time series research, "correlation" is often used procedurally [121, 36, 33]. That is, a correlation function is any function that takes two time series and produces a statistic, which is usually interpreted as a measure of similarity or relatedness. Examples include Pearson's correlation coefficient, local similarity [37], and cross-map skill [18]. Whereas a correlation statistic is a *summary* of an observed dataset, statistical dependence (or independence) is a *hypothesis* about the relationship between variables.

Two variables x and y are (statistically) dependent if the probability distribution of x while statistically controlling for y (the conditional distribution of x given y) differs from the distribution of x while not controlling for y (the marginal distribution of x). For example, lung cancer and smoking are dependent if the probability of lung cancer among smokers is different from that in the entire population. Importantly, dependence is linked to causality (as defined in the usual sense: x causes y if perturbations in x can alter y). The link between dependence and causality is due to Reichenbach's common cause principle, which states that if two variables are dependent, then they are causally related: Either they share a common cause, or one variable causes the other (possibly indirectly) [29, 5]. Thus, before seizing upon causal explanations, it is useful to to first test whether the observed correlation is strong enough to indicate dependence.

In the simpler (non-temporal) case where measurements of two variables are independent and identically distributed (iid, see Appendix A1 of Part I), the permutation test provides a general way to test for dependence between the two variables. Specifically, let (x_i, y_i) be the ith pair of measurements of variables x and y, and let θ be the observed correlation between the x and y measurements. The permutation test randomly shuffles the index of one of the variables, and then recomputes the correlation. This process is then repeated many times, essentially producing a null distribution. Under the null hypothesis that x and y are independent, correlations obtained from the original and the shuffled data follow the same distribution. Thus, a p-value can be calculated as (see, for example, section 6.2.5 of [123]):

$$\frac{N_{\geq} + 1}{N_{randomized} + 1} \tag{6}$$

where $N_{randomized}$ is the total number of "randomized correlations" (i.e. correlations obtained from shuffled data), and where $N_{\geq}$ is the number of randomized correlations that are at least as strong as the original correlation. The "+1" terms account for the original correlation. This test has three especially desirable properties. First, the test is *valid*: If we infer dependence only when p is less than some significance level α, then our false positive rate (i.e. the chance of erroneously reporting dependence) will be no more than α [123, 124]. Second, the test is distribution-free: It does not require that the variables or the correlation statistic follow a particular probability distribution [125]. Lastly, the test is without critical parameters: Its validity does not depend on any parameters that must be estimated or chosen by the user.

Dependence testing is less straightforward when applied to a pair of time series. Although a time series can have many data points, these data are not independent of each other in the sense that they are often autocorrelated (e.g. what occurs today influences what occurs tomorrow). A permutation test carried out by shuffling data within time series will generally not be valid. This is because temporal shuffling destroys autocorrelation, which often leads to artificially weak randomized correlations and thus an unacceptably high false positive rate (Figure 2 in Part I). If multiple independent and identical systems (trials) are available, we can instead perform a valid test by comparing within-trial correlations to between-trial correlations [43, 126, 127]. However, many important questions focus on a pair of single time series only (i.e. they have the "n-of-one" challenge). For instance, global-scale environmental studies are necessarily n-of-one because replicate Earths are unavailable, and the n-of-one perspective has been advocated in psychology because statistical patterns within one individual might differ from patterns in other individuals [128].

One way to address the n-of-one challenge in testing for a significant correlation between time series is to use parametric tests that account for autocorrelation [40]. However, these tests are limited to a particular correlation statistic (such as Pearson's correlation coefficient) [41, 40] because each test relies on

the availability of analytical results tailored to a specific statistic under the null hypothesis of independence. Consequently, the parametric testing approach may not be applicable to some increasingly popular statistics (e.g. cross-map skill in environmental sciences [18, 86, 129, 130, 89, 90]; see also [48, 43] for a broader overview of nonlinear dependence statistics). Parametric tests also require parametric assumptions (e.g. the sample correlation coefficient follows a prespecified distribution) and/or require that the user correctly estimate some parameter (e.g. the autocovariance matrix of the time series) [42, 41, 40]. Alternatively, if a correct model of the autocorrelation is known, the model can sometimes be used to remove the autocorrelation ("prewhitening") so that standard correlation tests may be applied [131]. Yet, prewhitening can remove between-variable dependence ([131], p. 120) and does not work for all data types [132].

When parametric tests and prewhitening are unavailable or inappropriate, dependence between time series is typically tested by the surrogate data testing approach [48]. Specifically, one begins with two time series $\{x_1, x_2, ...x_n\}$ and $\{y_1, y_2, ...y_n\}$ (abbreviated $\{x_t\}$ and $\{y_t\}$), and computes some measure of correlation between them. Next, one uses a computer to simulate how independent replicates of $\{y_t\}$ might have looked. These simulated $\{y_t\}$ time series are called "surrogate" $\{y_t\}$ series. Finally, one computes the correlation between $\{x_t\}$ and each surrogate $\{y_t\}$. A p-value is then given by the proportion of surrogate $\{y_t\}$ series that produce a correlation equal to or larger than the real $\{y_t\}$. More precisely, the p-value is given by Eq. 6, but where where $N_{\geq}$ is now the number of surrogates that produce a correlation statistic at least as large as the original $\{y_t\}$ series and $N_{randomized}$ is the total number of surrogates [47]. The permutation test is a special case of surrogate testing when data are iid instead of time series.

Several procedures have been used to generate time series surrogates, each with different strengths and limitations. For instance, the random phase procedure decomposes the $\{y_t\}$ time series into sine waves, randomly shifts these sine waves horizontally, and finally sums them up to produce surrogates [42, 129] (Appendix 3 Figure 1 of [127]). This procedure is valid when $\{y_t\}$ is a process that is Gaussian (any subsequence following a multivariate Gaussian distribution), stationary (this distribution does not change over time), and linear (future values depending linearly on past values and past random events) [48, 43]; see [49] for precise validity conditions. A more general version of the random phase procedure, called the iterative amplitude-adjusted Fourier transform (IAAFT) procedure, is valid when $\{y_t\}$ is a Gaussian, stationary, and linear process that is observed through an invertible but possibly nonlinear observation function [47, 48]. Yet even this more general version would exclude processes that are inherently nonlinear. Surrogates can also be produced by a block bootstrap method in which random subsequences are selected from $\{y_t\}$ and joined together [133, 50]. However, junctions between the blocks can produce disruptions, rendering the test inexact [133]. A sophisticated variant of the block bootstrap method, called the twin method, attempts to position blocks so that the disruptions are minimized [51]. Yet even in this method, performance depends on

"embedding parameters" [134] that must be appropriately chosen by the user, a potentially difficult task [118]. Overall, such surrogate procedures do not embody the three desired properties (being valid, distribution-free, and without critical parameters).

The procedure we propose in this Part is most closely related to time-shifted surrogates [135, 136, 137, 130, 50, 43]. The essential idea is to produce surrogates of $\{y_t\}$ by shifting the original $\{y_t\}$ in time. One can then calculate a p value according to Eq. 6. However, Bartlett [138] noted that the test is generally invalid because the surrogates are statistically dependent on each other.

Here we describe, analyze, and apply the truncated time-shift (TTS) test, a valid test for dependence between two time series. The TTS test is valid as long as one of the two time series is strict-sense stationary. Like the permutation test applied to iid data, the TTS test is compatible with essentially any correlation statistic and its validity does not require the assumption of a particular probability distribution, nor does it require that a user correctly select some parameter. Although the statistical power of the TTS test can depend on user-selected parameters, we demonstrate using simulations that with sufficient data, simple guidelines for parameter choices allow high power to be achieved. We first illustrate the TTS procedure and mathematically prove that it correctly controls the false positive rate. We then compare this test with other surrogate data tests in numerical experiments. We also consider how to set test parameters to increase the power with which it detects genuine dependence. Lastly, we demonstrate the use of the TTS test by applying it to real data from climatology, animal behavior and microbiome science.

We note that after we uploaded an earlier version of this Part to the bioRxiv preprint server [139], we happened to discover an arXiv preprint that independently conceived and proved an equivalent test [132]. The preprint, whose primary focus is a description and proof of the TTS test, additionally shows that the TTS test is not *excessively* conservative. That is, using the same procedure with a less stringent cutoff will always produce an invalid test for finite data (although a less stringent cutoff does become valid in the limit of infinite data). We nevertheless provide our version of the proof in Appendix A1 because (1) our proof is more complete in the sense that each statement is justified; (2) our proof applies more directly to finite-time processes, as we use a definition of stationarity that applies explicitly to finite time series (instead of stationarity in standard stochastic process literature which is defined only for infinite time series [140, 141, 142]); and (3) our proof is intended to be relatively accessible, with graphical illustrations of intermediate lemmas and relevant background concepts (Appendix A1.2).

2 Results

The truncated time shift (TTS) test

The truncated Time Shift (TTS) test is based on time-shifted surrogates, and requires shifting the original $\{y_t\}$ series in time without altering its length. One way to achieve this is to use cyclic permutations [130, 50]. That is, if the original $\{y_t\}$ series were $\{1, 2, 3, 4\}$, then there would be 3 surrogates, given by $\{2, 3, 4, 1\}$, $\{3, 4, 1, 2\}$, and $\{4, 1, 2, 3\}$. However, these surrogates artificially force the first and final points of the original $\{y_t\}$ series to become neighbors, which can distort the dynamics [48].

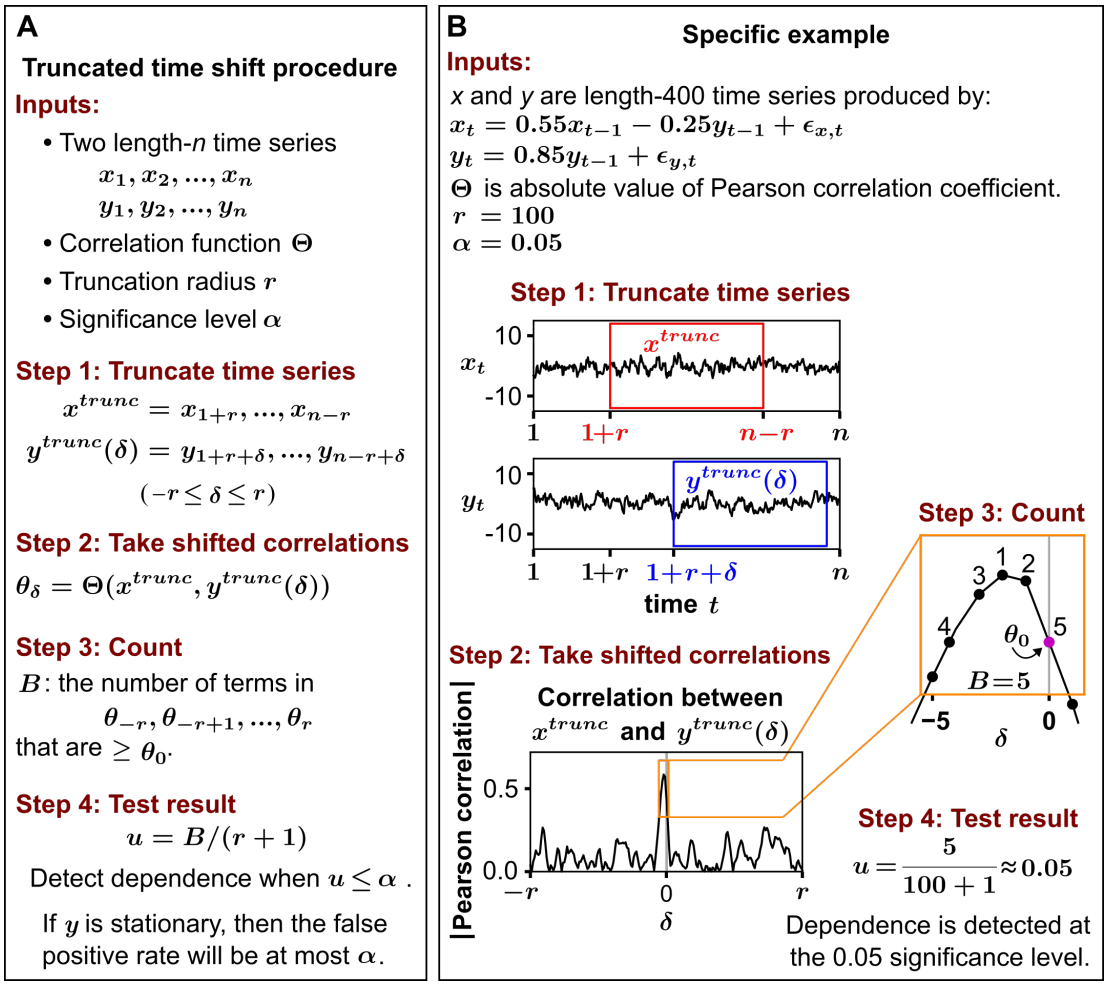


Figure 26: The truncated time shift procedure. (**A**) Stepwise description of the procedure. (**B**) A worked example. In the example, the process noise terms $\epsilon_{x,t}$ and $\epsilon_{y,t}$ are independent and identically distributed normal random variables with variance of 1 and zero mean. u can exceed 1 (e.g. the maximum B can be $2r+1$ in which case we do not reject the null hypothesis of independence). If one insists on reporting a u value of between 0 and 1, $\min(u, 1)$ instead of u can be used, giving what is sometimes called a "superuniform" p-value [143, 144].

Instead, we will truncate time series and then shift them to generate surrogates [48]. Starting with

$\{x_1, x_2, ...x_n\}$, we truncate r time points from each end of the sequence, and obtain:

$$x^{trunc} = \{x_{1+r}, x_{2+r}, ..., x_{n-r}\}.$$

We call r the "truncation radius". We then define a collection of truncated and shifted y time series, which all have the same length as x^{trunc}:

$$y^{trunc}(\delta) = \{y_{1+r+\delta}, y_{2+r+\delta}, ..., y_{n-r+\delta}\}$$

where the shift δ can take on integer values between $-r$ and r (Fig 26A-B, step 1). Note that when $\delta = 0$, x^{trunc} and $y^{trunc}(\delta)$ are aligned. Thus, we can think of $y^{trunc}(0)$ as our original time series and $y^{trunc}(\delta)$ for $\delta \neq 0$ as our surrogate time series. For each value of δ between $-r$ and r, we compute the correlation between x^{trunc} and $y^{trunc}(\delta)$ (Fig 26A-B, step 2). We then define B (for "Bigger") as the number of shifts δ that produce a correlation at least as large as when $\delta = 0$ (Fig 26A-B, step 3). B is bounded between 1 and $2r + 1$: We have $B = 1$ if the strictly greatest correlation is obtained when $\delta = 0$. Conversely, B is equal to $2r + 1$ if the lowest correlation (or a correlation that is tied for lowest) is obtained when $\delta = 0$.

If we were to naively apply the traditional logic of surrogate data testing (Eq 6), we could then write down a p-value as the proportion of correlations (shifted or not) that are at least as large as as the unshifted correlation:

$$p_{naive} = B/(2r + 1). \tag{7}$$

As Bartlett [138] noted, p_{naive} is not a valid p-value because the surrogate y series are not independent of each other (e.g. two consecutive shifts are nearly identical). Instead, our approach relies upon the following statistic:

$$u = B/(r + 1). \tag{8}$$

We refer to this procedure as the truncated time shift (TTS) test. Although u is not a p-value in the usual sense (e.g. $u > 1$ is possible), u can be used in the same way to establish evidence against the null hypothesis. That is, if the null hypothesis is true, then the probability of $u \leq \alpha$ is no more than α (Fig 26A-B, step 4). In Appendix A1 we prove that this property holds under the assumption that y is stationary. Roughly speaking, a temporal process is stationary (also called strict-sense stationary) if its probability distribution does not change over time (see Appendix A1.2 for a precise mathematical definition). Stationary processes

include many equilibrium processes, noise processes, chaotic processes, and periodic processes with random phases.

The above mathematical result may also provide a touch of comfort to analyses performed using the naive test: Comparing the formulas for u and p_{naive} (Eq. 8 and Eq. 7), we can see that as long as the requirement of the TTS test is satisfied (i.e. the time series used to generate surrogate data is stationary), the false positive rate of the naive test will not be inflated above the significance level by more than a factor of 2. However, we note that many applied studies do not use the naive TTS test as we have described it, but instead use a number of variations on it ([135, 136, 137, 50]). In Appendix A2, we consider two possible variants of the naive TTS test and use simulation examples to show that these may in fact be miscalibrated by far more than twofold.

The TTS test correctly controls the false positive rate when other surrogate data tests fail

Here, using simulated systems where two time series are independent, we compare the false positive rates for the TTS test and several existing surrogate data tests. Whereas all other tests fail in at least one stationary system, the TTS test performs correctly in all stationary systems (as expected). The TTS test also performs well in the two nonstationary systems considered here.

Specifically, Fig 27Ai-ii shows a first order autoregressive process and a logistic map, which are two stationary systems commonly used for benchmarking (e.g. [42, 18]). Fig 27Aiii-vi shows four stationary systems with a combination of periodic dynamics and noise designed to challenge existing tests. Fig 27Avii-viii show two biologically-inspired nonlinear systems: a stochastic FitzHugh-Nagumo neuronal model [145] and a competitive Lotka-Volterra system with chaotic behavior [146]. These two systems are likely to be stationary, although formal proofs are generally difficult for multivariate nonlinear systems [147]. Fig 27Aix-x show two systems known to be nonstationary: a random walk and a first order autoregressive process (same as A) with an additive term that increases over time. In all cases, the two time series are independent. See Appendix A3 for mathematical details.

Two different correlation statistics were used: Pearson correlation strength (absolute value of sample Pearson correlation coefficient; Fig 27B, left half), and an estimator of mutual information [148], which is a popular nonlinear form of correlation (Fig 27B, right half). We do not use statistics based on the Granger causality framework as correlation statistics in this Part. This is because the Granger causality framework requires tests of conditional dependence [17], whereas the TTS test and most surrogate data procedures provide tests of (unconditional) dependence and thus are generally inappropriate for Granger

causality testing (i.e. Figure 6 of [77]).

We compared the following surrogate data tests: the IAAFT procedure [47], the stationary block bootstrap [149, 133, 50], the twin surrogate procedure [51], cyclic permutation time-shifted surrogates, the naive TTS test (Eq. 7) and the TTS test (Fig 26). For the first four tests, we use circularization to reduce a potential discontinuity caused by wrap-around effects (Methods), as recommended by [47, 43]. We also tested four autocorrelation-aware parametric tests: A t-test for Pearson correlation strength (Methods) given by [150], a modified version of the same test using a procedure suggested by [41] (Methods), and two recently derived tests for Pearson correlation and mutual information in linear systems [151] respectively. In this benchmark, the latter three tests exceeded the 5% false positive rate threshold more often than the first test and performed similarly to each other (supplementary data file 1). For this reason, we show only the original test of [150] in Fig 27, and use only the same test as our parametric test for all subsequent comparative analyses.

The TTS test has a false positive rate below 5% for all stationary systems. All procedures other than the TTS test mistakenly detect dependence at rates above 5% in one or more stationary systems (Fig 27B). For the cyclic permutation test, a relative of the TTS test that has been used in practice, we go beyond simply showing failure and explore why cyclic permutations break down (whereas the TTS procedure does not) in Appendix A3.4. The TTS test also showed low false positive rates with the two nonstationary systems (Fig 27B, last two rows), although the TTS test is not guaranteed to be valid when surrogates are generated from a nonstationary process (e.g. Fig A42).

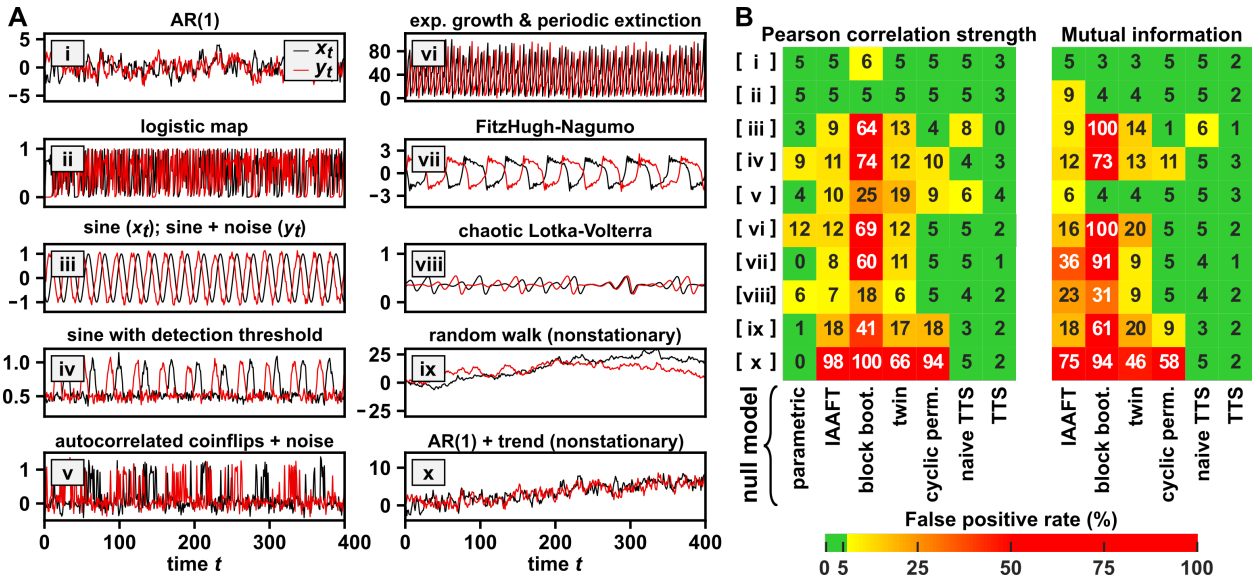


Figure 27: The truncated time shift (TTS) test controls the false positive rate in stationary systems and some nonstationary systems. (**A**) Example time series from different benchmark systems, some of which (i.e. systems i-vi) can be proven to be stationary (Appendix A3). Systems vii-viii are probably stationary, but it is in general difficult to prove strict stationarity for multivariate nonlinear stochastic systems [147]. Systems ix-x are not stationary. A system can have process noise (noise whose effect can propagate to subsequent time steps) and/or measurement noise (whose effect does not propagate). (i) First-order autoregressive process: current values depend linearly on values one step ago and process noise. (ii) Logistic map: a deterministic discrete-time model of population dynamics with growth and a carrying capacity [152]. (iii) Two sine waves, one with measurement noise whose strength varies with a slow "sawtooth wave" [153]. (iv) A sine wave with a detection threshold and measurement noise. (v) A series of coin flips with measurement noise (with 'heads' and 'tails' coded as 1 and 0 respectively), where coins are autocorrelated because the probability of a 'heads' varies over time in a stationary way. (vi) A simple model of a population with exponential growth, periodic extinction events, and constant immigration. (vii) A stochastic discrete-time FitzHugh-Nagumo model, which is a nonlinear oscillator inspired by neural voltage dynamics. (viii) Chaotic Lotka-Volterra model: an ecological model where species engage in intra- and inter-species competition. (ix) A random walk with Gaussian steps (i.e. $x_t - x_{t-1}$ follows a zero-mean normal distribution). (x) The same process as in (i), but with an additive temporal trend. (**B**) False positive rates of dependence tests calculated from 1.5×10^4 trials. Two independent time series were generated from each system and correlated using either the absolute value of the Pearson correlation coefficient or an estimate of mutual information. Each surrogate data test was then used to test for significance (at the 0.05 level) of the correlation under the null hypothesis of independence. The labels "block boot." and "cyclic perm." are shorthand for stationary block bootstrap and cyclic permutation surrogates. See Appendix A3 for further details.

For time series that can be decomposed into a deterministic "trend" and a stationary component (e.g. Fig 27Ax), the TTS test can be modified to be rigorous by first detrending followed by retrending (similar to [154]). The basic idea is to: (1) extract the stationary component by removing the trend, (2) generate surrogates from the stationary component, and then (3) add the trend back to the surrogates. In Appendix A4, we demonstrate that the detrending-retrending TTS procedure produces a valid surrogate data test for time series that are nonstationary due to a deterministic trend.

In summary, whereas the previous section established that the TTS test is universally valid for dependence

testing in stationary time series, here we have shown using simulation examples that popular surrogate data tests are not universally valid.

Simple guidelines for choosing TTS parameters to achieve high statistical power

Having shown that the TTS test always correctly controls the false positive rate (as long as one series is stationary), we now consider its true positive rate (detection power; the probability of detecting a true dependence). Apart from the choice of correlation statistic, the power of the TTS test is sensitive to two parameters. One parameter is the truncation radius r, which is specific to the TTS test. The second parameter (the pre-shift amount s), which we introduce later, arises from the general problem of delayed coupling (e.g. when the effect of one variable on another occurs after a lag). However, simple guidelines for choosing r and s allow high statistical power to be achieved when sufficient data are available, as we demonstrate below using simulation examples.

The truncation radius r simultaneously determines the length of the truncated time series (which is $2r$ less than the total number of time points) and the number of time-shifted surrogates (which is $2r$). A rearrangement of Eq. 8 gives $r = B/u - 1$. If r is exactly 19, significance would be detected at the $u = 0.05$ level only when $B = 1$, meaning that the original correlation would need to be strictly greater than all shifted correlations. If $20 \leq r \leq 38$, significance at the 0.05 level still requires $B = 1$ (since if $B = 2$ and $r = 38$, then $u = B/(r+1) = 2/39 > 0.05$). Thus, choosing $20 \leq r \leq 38$ will never achieve higher power than setting $r = 19$. In general, for a significance level α, power is maximized when r is one less than an integer multiple of $1/\alpha$. Note that the data length needs to be more than $2r$, and consequently more than $2/\alpha$. As r grows larger, the TTS test will be able to detect dependence with progressively greater values of B. However, if r is too large, the truncated time series will be so small that correlation estimate becomes noisy.

Lagged dependence arises when one variable affects another after a delay, or when a common driver affects two variables with different delays (Fig 28A black curves, x lagging behind y). Detecting lagged dependence can be challenging for many statistical tests (i.e. all tests in Fig 27), because although two variables may have a strong shifted correlation (between x_{t+s} and y_t for some shift s), the unshifted correlation (which may at first be the natural statistic for a dependence test) might be very low. The problem is especially severe for tests that rely on shifted correlations as their null models (including the TTS test). This is because the existence of a coupling lag means that some of the null (shifted) correlations will exceed the original (unshifted) correlation (Fig 28B black curve; many black dots above the red dot), potentially leading to low power.

The problem of low power associated with lagged dependence can be mitigated by pre-shifting the $\{x_t\}$ series - the series which is *not* used to produce surrogates (Fig 28 A, light teal curves). In the examples explored here (Fig 28), simply pre-shifting by an amount similar to the coupling lag produced a powerful test. That is, when the pre-shift amount was similar to the coupling lag, correlation at $\delta = 0$ reached a high value (Fig 28B, $s = 2$), enabling detection of dependence. If we know a range of likely values for the coupling lag, we can perform a "multi-shift" test: Test for dependence between $\{x_{t+s}\}$ and $\{y_t\}$ for several different values of s, and perform a Bonferroni correction to account for multiple tests. If m different pre-shifts are used, then we report dependence if any of the m tests is significant at the α/m level. Correspondingly, the truncation radius will need to be 1 less than an integer multiple of m/α (Fig 28 C). As the range of possible lags becomes larger, the minimum r value will increase and longer time series may be needed.

To demonstrate how pre-shifting can improve the statistical power of the TTS test, we use an "unforgiving" case (an autoregressive process where y influences x with a delay of $\tau = 2$ time steps in Fig 28Di). This process is unforgiving in the sense that the shifted correlation plot exhibits a sharp peak, and consequently the unshifted correlation is not higher than most of the shifted correlations (Fig 28Dii). Consequently, the TTS test does not obtain a significant Pearson correlation (Fig 28Diii, first row). The challenge is not entirely unique to the TTS test, as the IAAFT and parametric tests also suffer low power. Pre-shifting x by $\tau = 2$ allows all three tests (including the TTS test with different values for the truncation radius r) to detect the dependence with high power (Fig 28Diii, second row). If the pre-shift amount s is too large, the power again declines as expected, and compared to the IAAFT and parametric tests, the power of the TTS test is more sensitive to an incorrect pre-shift (Fig 28Diii, third row). The multi-shift approach described in the previous paragraph (Fig 28 C) achieved high power (Fig 28Div). With sufficient data, the multi-shift strategy also enabled high detection power when two variables affect each other with different and uncertain lags, as long as the range of pre-shifts covers the maximal lag uncertainty (Appendix A5.2).

Pre-shifting may not be necessary in "forgiving" cases. One such case is a coupled bivariate logistic map, a nonlinear system whose coupling can be readily detected using the "cross-map skill" correlation statistic [18, 129, 97] (Fig 28E, i). As before, the two processes are coupled with a lag of 2. The shifted correlation plot exhibits a peak that is broader than the coupling lag (Fig28E, ii). In this case, since $B = 4$ (as three of the shifted correlations are greater than the unshifted one), a truncation radius of $r = 4/5\% - 1 = 79$ or above can detect dependence even without any pre-shifting (Fig 28E, iii first row).

It is inappropriate to visually inspect the plot of shifted correlations and then decide which pre-shift to use, since this uses test output to select test parameters and will invalidate the test. However, we do note that similar to Fig 28E, other studies have also found broad peaks when using nonlinear correlation statistics to detect coupling between nonlinear deterministic systems [48, 96]. Additionally, if at least two replicate

systems are available for study, one may estimate the peak breadth (and perhaps also position) from one dataset, and use this knowledge to perform the test on the other dataset.

As shown in Fig 28 and in supplementary data and Appendix A5.1, for the autoregressive process, TTS power is generally lower than that of the IAAFT, parametric test, and other tests in Fig 27 (although the multi-shift procedure dramatically improves power relative to individual pre-shifts of 0 or 4; Fig 28Diii). For the logistic map process, TTS power is similar to that of all other nonparametric tests, and far superior to the parametric test. These conclusions hold when we varied parameters such as the time series length, interaction strength, and autocorrelation (Appendix A5.1).

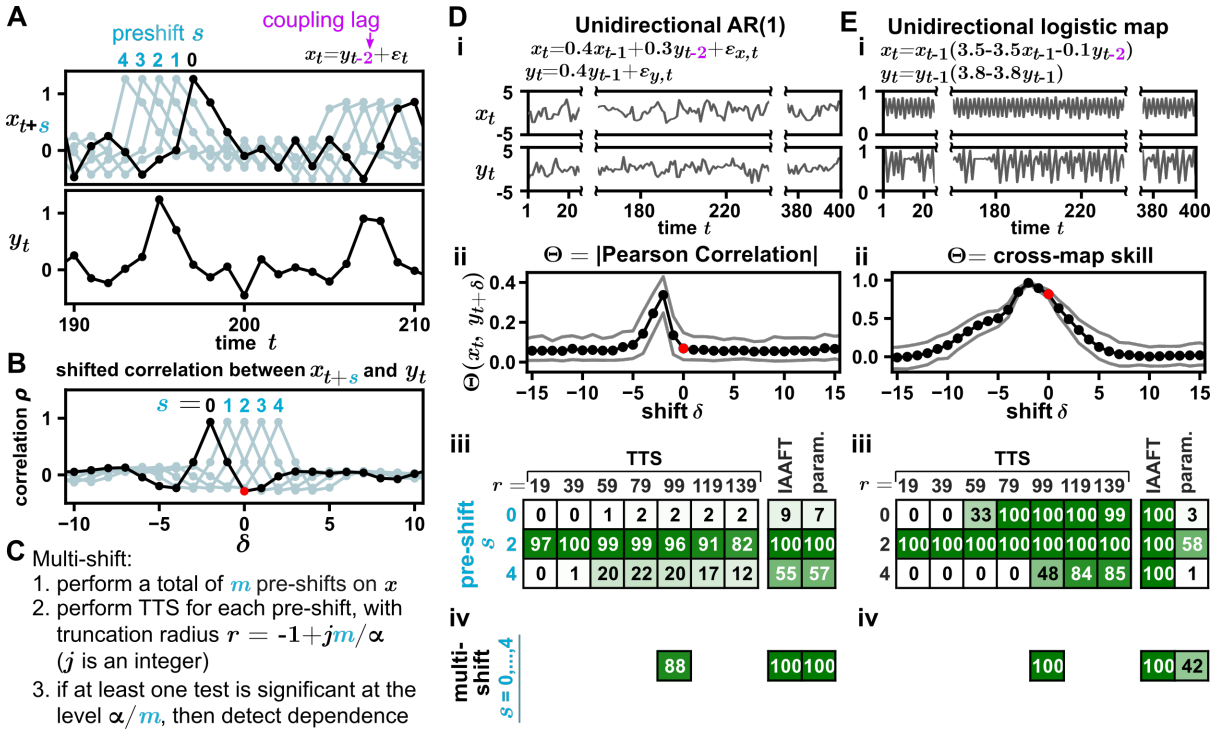


Figure 28: Strategies for increasing power in the challenging setting of a coupling lag. (**A-C**) Conceptual illustration of pre-shift and multi-shift. (**D, E**) Various pre-shifting strategies to account for coupling lag increase the detection power of the TTS test. (**i**) Time series in which y influences x with a lag of 2. We compare a coupled linear autoregressive process (D) and a nonlinear logistic map (E). (**ii**) Shifted correlation plots. The absolute value of the Pearson correlation (D) or cross-map skill (E) was computed between x_t and $y_{t+\delta}$ for various δ shifts. D is more challenging than E since the unshifted correlation (red dot) is at the foothill in D but near the maximum in E. To calculate cross-map skill, we used x to estimate y (as appropriate when y influences x [18]), and the embedding dimension and the embedding lag were set to 2 and 1 respectively following prior works [18, 96]. Grey lines indicate the middle 80% of correlations from 100 trials obtained with a truncation radius of 79. (**iii**, **iv**) Power comparison. Tests in Diii and Eiii used the correlation statistic specified in Dii and Eii respectively, except for the parametric test, which always used Pearson correlation. Prior to performing the tests, the x series was pre-shifted by s (i.e. testing for dependence between $\{x_{1+s}, ..., x_n\}$ and $\{y_1, ..., y_{n-s}\}$, where $s = 0, 2$, or 4 for single pre-shift at significance level 0.05 (iii) or $s = 0, 1, 2, 3, 4$ for multi-shift at a Bonferroni-corrected significance level $0.05/5 = 0.01$ (iv). Detection power was computed as the proportion of 10^4 simulations in which dependence was detected. Note that with multi-shift, we choose $r = 5/0.05 - 1 = 99$.

In the next three sections, we apply the TTS test to existing data sets obtained from systems in climatology, microbiome science, and animal behavior science. These case studies serve as examples of real systems wherein the TTS test is sufficiently powerful to detect dependence relationships, some of which were detected by the original authors of the data sets.

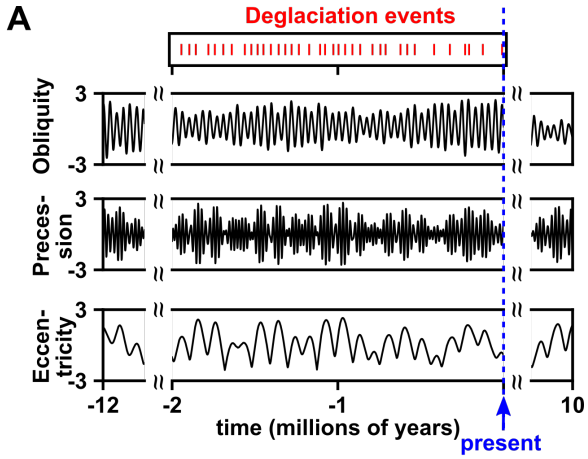
An example from climate science

Pre-industrial climatological change is widely understood to have been driven largely by variability in the Earth's orbit around the sun. The Earth's orbit is characterized by three "orbital parameters" known as eccentricity, obliquity, and precession. Eccentricity describes the shape of the orbit (which varies from nearly circular to slightly elliptical over approximately $96,000$-year cycles); obliquity is the angle between Earth's rotational axis and the normal of the orbital plane (which cycles over roughly $41,000$ years in a band roughly bounded between $22°$ and $24.5°$); and precession describes how the rotational axis of the earth rotates around the line normal to the orbital plane (roughly $21,000$ years/cycle) [155, 156]. Each of these parameters is thought to play a role in Earth's climate, although some parameters may be more influential than others, and the extent of a parameter's influence may change over time [157, 158, 156]. The climate record is characterized by repeated episodes of cooling followed by warming events called deglaciations. Until about one million years ago, deglaciations occurred with a period of about 41 kiloyears, which is the period of obliquity cycles. Because of this, obliquity is often said to "pace" glacial cycles [159, 160]. Yet, two time series with shared periodic elements can be statistically independent (e.g. Fig 27Aiii).

Using the TTS procedure, we tested for dependence between orbital parameters and deglaciations with only the assumption that the time series of the three orbital parameters are stationary (Fig 29A). We used the entire past 2 Myr of deglaciation series as our "truncated" x time series, and generated surrogates using orbital parameter time series spanning from -12 Myr to 10 Myr (i.e. with a truncation radius of 10 Myr; Fig 29A). We did not pre-shift time series because 1) we did not have a prior expectation of the coupling delay, and 2) pre-shifting may not be necessary, because past values could be used to accurately predict the future values within each series (Appendix A6.2). Hence, we expected that small-to-moderate pre-shifts before the TTS test would be unlikely to severely change the correlations with the deglaciation series (see final paragraph of Appendix A6.2 for full argument; see also Fig 28E and related discussion of broad peaks in coupled nonlinear systems).

The TTS test detected a dependence between deglaciation times and obliquity ($u < 0.05$), but not the other two orbital parameters (Fig 29B), similar to Huybers' original period-based analysis [160]. For the TTS test, we used the absolute value of the Pearson correlation coefficient so that both positive and negative

correlations could be detected. We also used a prediction-based nonlinear correlation function (Appendix A6.2; Fig 27C), which can better detect dependence in scenarios where Pearson correlation may be low (e.g. if deglaciations occur at the midpoints between peaks and troughs of orbital series). Both detected the same dependence relationship. A two-tailed parametric Pearson correlation test (see Methods) did not detect the obliquity-deglaciation dependence (Fig 29B). The TTS test is arguably more appropriate than the parametric test in this setting because periodic processes can potentially render the parametric test invalid (Fig 27Biv).



B

	Pearson correlation	TTS *u* value	Parametric *p* value
Obliquity	0.12	0.02	0.052
Precession	-0.04	0.49	0.19
Eccentricity	0.02	0.64	0.29

C

	Nonlinear estimation MSE	TTS *u* value
Obliquity	0.0175	0.03
Precession	0.0177	0.31
Eccentricity	0.0177	0.20

Figure 29: The TTS test detected dependence between deglaciation and obliquity, but not between deglaciation and precession or eccentricity. (**A**) Time series of deglaciation events (from [160]) and the three orbital parameters (estimated from the model in [155]). To convert the 36 deglaciation events ([160]) in the last 2 million years, we used a "sampling frequency" of 1000 years by assigning a 1 to the deglaciation variable if a kiloyear contained a deglaciation event and a 0 otherwise. Our deglaciation time series thus has 2000 time points. We did not use a higher sampling frequency due to uncertainty in deglaciation timing, and avoided a lower sampling frequency (e.g. 10 kiloyears) to adequately capture the shapes of the obliquity and precession cycles. To estimate obliquity, precession, and eccentricity, we used the numerical solution from [155], which provides accurate estimates of orbital parameters over at least tens of millions of years (and also predicts future values). Orbital values are standardized to a mean of 0 and variance of 1. To obtain unshifted correlations, we used truncated time series with times between -1999 kyr and 0 kyr (present time), yielding a total of 2000 time points. Orbital parameter time series were used to generate time-shifted surrogates, with a truncation radius of 10 million years ($20,000$ time-shifted surrogates). (**B, C**) Testing for dependence between orbital parameters and deglaciation. The correlation statistic for the TTS test is either the absolute value of Pearson correlation (B) or mean squared error (MSE) when using orbital parameters to estimate deglaciation events via a state space-based technique (C). The latter, a nonlinear statistic, is similar to the cross-map skill statistic used in other largely deterministic nonlinear systems (Appendix A6.2). Note that when one variable is continuous and the other variable is binary (as here), the Pearson correlation is also known as the "point-biserial correlation" (see pg. 294 in [161]). The TTS tests as well as the two-tailed parametric test using Pearson correlation all suggest that obliquity and glaciation are dependent (u or p values below 0.05).

An example from human microbiome science

The human microbiome is highly spatial, with different body sites playing host to distinct microbial communities [162]. Caporaso et al. performed a longitudinal study in which daily microbial surveys were conducted for over a year on a male subject at four body sites: the left palm, right palm, tongue, and gut (via feces sampling) [163]. Since this study measured relative abundance of microbial taxa (not absolute population

size), it is difficult to test for dependence between x_{mouth} and y_{mouth}, or between x_{mouth} and x_{skin} where x_{site} is the absolute abundance of species x at location "$site$" (Fig A44; see also [164]). This is because relative abundance values sum to a constant, and are thus trivially dependent. Addressing this challenge in the context of time series is beyond the scope of the present work. However, we can more easily test for overall dependence between the microbial communities of two body sites. Intuitively, this is because if the absolute abundance values of microbial species on two body sites are independent, then the relative abundance must also be independent. As this statement's contrapositive, if the relative abundance of microbial species on two body sites are dependent, then the absolute abundance must also be dependent (see Appendix A7.2 for a formal argument).

We applied the TTS test to the time series from [163] to look for dependence between the microbial communities living on the left palm, right palm, tongue, and gut. We first obtained OTU (operational taxonomic unit) relative abundance tables from the data in [163] using the online Qiita platform [165] (see Appendix A7.5). We then preprocessed data (Fig 30A) by removing or filling gaps in time series, and by removing OTU abundance time series that were either mostly absent or potentially nonstationary. Analyzing nonstationary processes is an important problem, but requires prior information, and is outside the scope of this example. After pre-processing, the number of remaining OTUs ranged from 180 (tongue) to 507 (right palm). Finally, we performed a TTS test between each pair of body sites. Note that the TTS test is valid for testing dependence between two sets of data where each set consists of a multivariate time series (Appendix A1), as in this case study (Fig 30B). To correlate datasets from two body sites (Fig 30B), we first listed all of their shared OTUs. Then for each shared OTU, we computed the sample Pearson correlation coefficient between the relative abundance series of that OTU in the two sites. Our correlation statistic Θ was the median of these correlation coefficients across all shared OTUs. We emphasize that we did not perform a separate TTS test for each OTU since the correlation between body sites was summarized by a single univariate statistic (Fig 30B). We did not pre-shift time series because we expect dependence among body sites to be largely driven by migration on time scales likely faster than the sampling period of one day.

We detected dependence between the microbial communities on the left and right palms, and between the palms and the tongue (Fig 30C). This result was the same if instead of Pearson correlation, we used local similarity, a correlation statistic designed to detect transient temporal correlations that is popular in microbiome science [37, 166]. These results reveal more cross-site dependences than the original analysis of [163], which detected dependence only between the left and right palms. The original study first computed the phylogenetic distance between temporally adjacent microbiomes within body sites, and then calculated correlations between the phylogenetic distances of different body sites [163]. Whereas that analysis relied on parametric tests with assumed null distributions, our analysis relies on correlation statistics that lack a

readily-computed parametric sampling distribution (i.e. the median of many Pearson correlation coefficients or local similarities). Yet, due to the flexibility of the TTS test, we are nevertheless able to perform a valid statistical test, assuming that abundance time series are stationary.

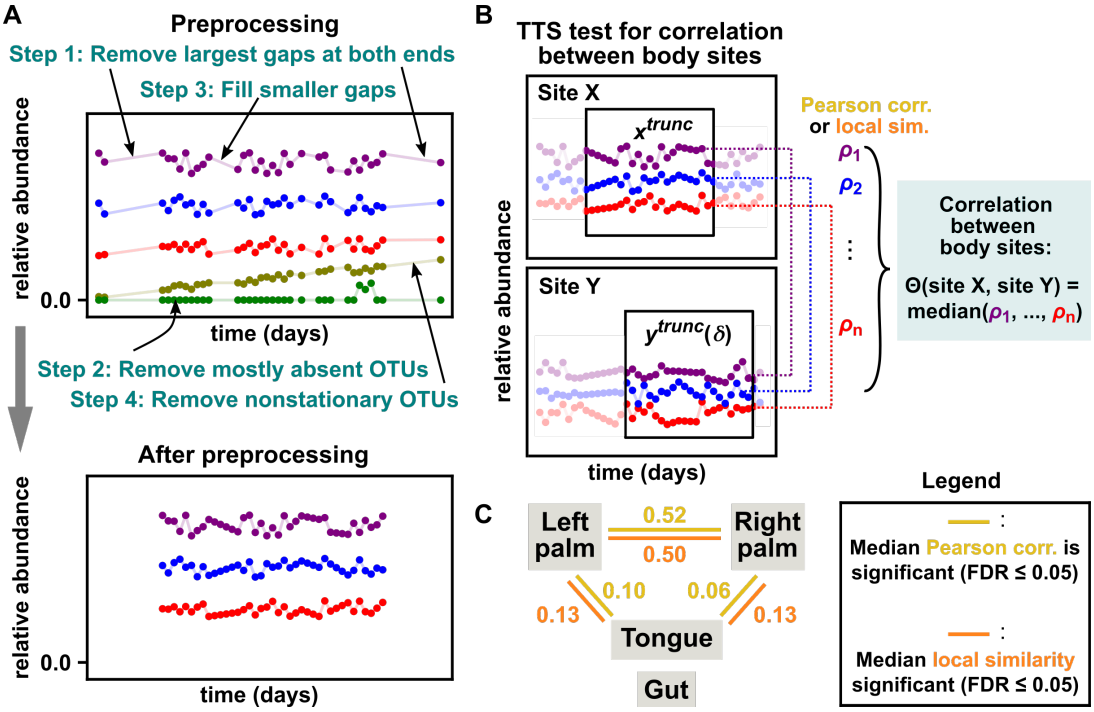


Figure 30: The TTS test applied to longitudinally sampled microbiomes from four body sites detects dependence between body sites. (A) Data preprocessing. To remove long (> 6-day) gaps at the beginning and end of time series, we only used measurements from day 42 to day 418. Remaining gaps were filled by linear interpolation or by randomly resampling abundance values. OTUs were removed if they were absent in over half of measurements or if they were not considered stationary (at the 0.05 significance level) by an augmented Dickey-Fuller (ADF) test implemented in the Statsmodels Python package [67]. OTUs removed from one body site were not necessarily removed from the other body sites so that correlations could still be computed between the other sites. (B) TTS test procedure for correlation between body sites. We used an intermediate truncation radius r of 79 days (inspired by Fig 28C). We quantified correlation between two body sites as follows: For each shared OTU i, we computed ρ_i (the Pearson correlation or local similarity score of OTU i between the two sites). We do not take the absolute values of these statistics for hypothesis testing because, following [163], we expect correlations to arise from mixing of biological material between body sites, and therefore be mostly positive. We then chose the median of $\{\rho_1, ..., \rho_m\}$ (where m is the total number of shared OTUs) to be the between-site correlation (i.e. our "Θ" in the notation of Fig 26). This setup avoids the need to perform a separate test for each OTU. (C) The TTS test detected dependence between the two palms and between palms and the tongue. Numbers in gold (orange) denote the median intraspecies correlation as measured by Pearson correlation (local similarity) when gaps were filled by linear interpolation. All links shown were detected with a significance level of 0.05 after a Benjamini-Hochberg false discovery rate (FDR) adjustment for performing 6 tests [167, 168]. Note that the gut shares few species with the other sites (Fig A45). The same network of significant correlations was obtained for (C) regardless of either the gap-filling method in (A) or whether Pearson correlation versus local similarity was used in (B). Additionally, for each correlation, we obtained the same result regardless of which body site was used to generate surrogates. See Appendix A7.5 for more details.

An example from animal behavior science

A major goal of animal behavior research is to understand the rules that govern how animals act, at the levels of both individuals and groups (e.g. swarms of insects or shoals of fish). Video tracking techniques enable measurements of variables such as an individual's position and velocity [169]. These quantitative measurements have helped researchers detect subtle or complex behaviors, enable useful analogies between animal group behavior and materials physics [170, 171], and connect individual-level and group-level phenomena [169, 172, 173, 174].

We applied the TTS test to a set of zebrafish trajectories recorded by Romero-Ferrero et al. [175]. 100 juvenile zebrafish were placed in a shallow circular tank and tracked by an overhead video camera at 32 frames per second, yielding 2-dimensional trajectories of each fish (height was not measured; Fig 31A). We here use the TTS test to ask whether there is a dependence between the speed of a fish and its direction of motion.

To define a coordinate system for direction of motion, we noticed that large groups of fish often swam parallel to the perimeter of the tank (Fig 31A). To capture this behavior, we define an individual's direction of motion φ as the angle between the individual's position vector and velocity vector (Fig 31B).

For an arbitrary fish, different directions appear to correspond to different speeds (Fig 31C). It is unnatural to quantify this association using the Pearson correlation between speed and direction φ because speed is a linear quantity whereas direction is a circular quantity. We can make linear correlation more appropriate by first transforming the direction variable φ to $|\sin(\varphi)|$, which is the largest when $\varphi = 90^\circ$ or 270° (swimming parallel to the perimeter), and the smallest when $\varphi = 0^\circ$ or 180° (swimming away from or toward the center), and then compute the Pearson correlation between $|\sin(\varphi)|$ and speed. Using this statistic and the TTS test, we detected dependence between speed and direction in 15 out of 44 fish (at the 0.05 level after a Benjamini-Hochberg false discovery rate correction), and in pooled data ($u = 0.02$; 31D). Alternatively, we can use mutual information as our statistic, which is not limited to linear dependence. In this case, we detect dependence between direction and speed in most of the fish that we analysed (35 out of 44; Fig 31D). We did not pre-shift time series because we had no prior expectation of a coupling delay between swimming speed and direction.

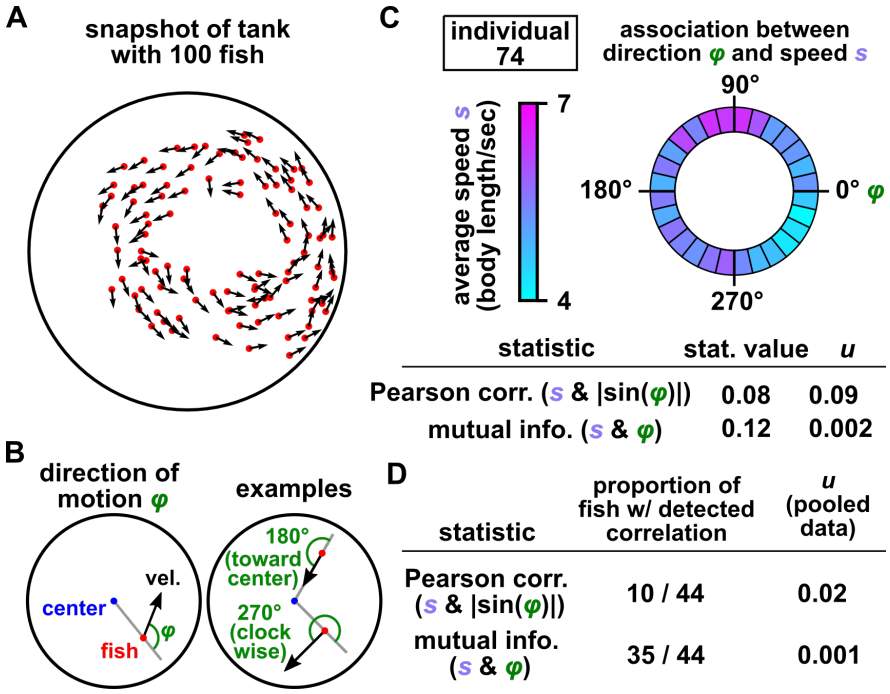


Figure 31: (A) A snapshot of fish positions in a 100-fish tank. (B) A fish's direction of motion φ is defined as the angle between the fish's position vector and velocity vector. Direction φ values of 0° and 180° correspond to motion exactly away from the center or toward the center respectively. Direction values of 90° and 270° correspond to motion exactly counterclockwise and clockwise respectively. (C, D) Association between direction φ and speed s for an arbitrary individual ("individual 74") and for 44 fish in which recorded trajectories did not contain gaps. See Fig A47 for time series. The TTS procedure tests for dependence using the absolute value of the Pearson correlation coefficient between speed and $|\sin(\varphi)|$, or using the mutual information between speed s and direction φ. Mutual information detects correlations more readily in this case. When testing for correlations in 44 fish (D), detections were made at a significance level of 0.05 with a Benjamini-Hochberg FDR correction [167, 168] (middle column). We also performed a TTS test that incorporated all 44 trajectories (right column; "pooled data"). This test was analogous to the test of Fig 30B. That is, for each time shift, we calculated the overall correlation between speed and direction as the median correlation among all 44 trajectories. For all tests we used data from the first of the three replicate videos in [175] and limited our analysis to the first $10,000$ frames (about 5 minutes) since this data segment appeared approximately stationary by visual inspection, although fish speeds are likely nonstationary overall (see Appendix A8.2 and [176]). We used the speed variable for time-shifted surrogates and used a truncation radius of one tenth the total time series length ($1,000$ frames). See Appendix A8 for further details.

3 Discussion

A statistical hypothesis test for dependence between two time series usually requires at least two key ingredients: (1) a correlation statistic that quantifies the strength of the relationship between the two time series, and (2) a null model that can be used to determine the probability distribution of that statistic if the two time series are in fact independent. These two ingredients seem to have received different levels of attention over the past couple of decades. Recent years have witnessed the development and rapid adoption of new correlation statistics that can detect transient or nonlinear forms of dependence, some of which even attempt to

infer the direction of causation [37, 166, 18, 96, 23, 43, 127]. In practice however, these correlation statistics have often been paired with surrogate data null models which assume a Gaussian process (i.e. random phase surrogates), or even independent and identically distributed data (e.g. the permutation test). Analyzing nonlinear processes with a null model that requires a linear process can indeed be dangerous (Fig 27B rows ii-viii, IAAFT columns; see also [48, 92]). Overall the general problem of assigning statistical significance to nonlinear correlations between time series does not appear to have a broadly-accepted solution [92, 103, 5].

The TTS can serve as a (provably) conservative solution to this problem since it is valid as long as one of the time series is stationary (or can be made stationary by detrending; Appendix A4). This is a minimally restrictive requirement among valid nonparametric tests of dependence between time series. This test was sufficiently powerful to verify the previously observed dependence between obliquity and deglaciation timing (Fig 29) as well as dependence between the microbiome compositions of the left and right palms (Fig 30). In the microbiome dataset, we could even use it to identify additional relationships that went undetected by the original analysis of Caporaso et al. [163]. Importantly, the TTS can be applied in cases where linear parametric tests are particularly unnatural such as when correlating fish swimming speed to direction (which exists in an inherently nonlinear space, i.e. a circle). Since surrogate data tests are presently used in disciplines ranging from earth system science to physiology [42, 177, 135, 136, 129, 88, 43, 89, 178], we expect the TTS test to find utility in diverse application domains.

Although TTS test is one-tailed, it may still be used to detect high or low correlations (or both). For example, if the correlation function is the Pearson correlation coefficient, then correlations are deemed "impressive" if they are positive and large. Alternatively, we primarily used the absolute value of the Pearson correlation coefficient for which correlations would be deemed impressive if they are extreme (i.e. large magnitude positive or negative). A third possibility is to simply carry out two TTS tests, one using the Pearson correlation and one using the negative Pearson correlation, and then perform a Bonferroni correction for two tests. If the distribution of correlations is symmetric and centered around zero, then the last two strategies are similar .

The main limitation of the TTS test relative to other tests is that the TTS test tends to require more data or stronger coupling to achieve high detection power. Under the TTS test, the required length of the time series increases with the reciprocal of the desired significance level. Even more data are required if the optimal coupling delay is uncertain. Moreover, whereas other surrogate data tests often specify the desired false positive rate (at least approximately), the TTS test only specifies an upper bound on the false positive rate, and the actual false positive rate is often substantially lower (Fig 27). We can therefore expect other tests for dependence to typically achieve high detection power with fewer data than the TTS test. Indeed, this is what we have found in numerical experiments (Appendix A5). Thus, the TTS test is ideally used

when a nonparametric test is required, when time series are long, when the coupling lag can be pinned down to a small range, and when one is unable to verify the assumptions necessary for other surrogate data tests (which typically include additional requirements beyond stationarity [48, 43]). Additionally, since the naive TTS test seems to usually perform well (Fig 27), we think it is reasonable for authors to choose the naive TTS test for stationary time series, provided they note that under pathological conditions the naive TTS test may underestimate the false positive rate by up to twofold.

Since the stationarity assumption is so central in dependence testing, it would be convenient to have a statistical test for stationarity. However, a single time series can in theory be described by either a stationary process (e.g. Fig A48A) or a nonstationary one (e.g. Fig A48B). Nevertheless, there are many statistical methods that attempt to test for stationarity or a similar property in a single time series, albeit with various modeling assumptions or other caveats [179, 180, 181, 182]. For example, although the popular augmented Dickey Fuller (ADF) test [179, 182] is sometimes used as a pragmatic means of assessing stationarity (e.g. Fig 4; [53, 183]), its null hypothesis is not exactly nonstationarity. In fact, a rejection of the ADF test's null hypothesis indicates that a time series is free from some sources of nonstationarity (e.g. a random walk), but other sources of nonstationarity (e.g. time-varying parameters) may in principle still be present. Overall, statistical tests cannot guarantee that a time series is stationary, but can provide supporting evidence. Background knowledge of the process can also be used to support the stationarity assumption: Mathematical work has shown that the stationarity condition is often met by stochastic processes that tend to relax toward a stable equilibrium [184]. Periodic processes, measurement noise processes, and chaotic processes can also be stationary (e.g. Fig 27B), as can combinations of these processes.

The subject of conditional dependence has been conspicuously absent from our discussion. Tests of conditional dependence (i.e. whether two variables are dependent after we statistically control for a third variable) can help to rule out possible common-cause explanations, and can sometimes even be used to reveal the direction of causation [3, 127]. We initially motivated the TTS test by noticing that it enjoys the rigor and generality of the permutation test, but applies to time series rather than iid data. Could a test for conditional dependence between time series be devised with the same rigor and generality? This seems difficult. Even for continuous iid data, it has been proven under general conditions that no valid test for conditional dependence can both avoid making assumptions and have statistical power [185]. Thus, we expect most tests of conditional dependence among time series to be less rigorous or require more assumptions than the TTS test. Nevertheless, there have been promising recent advances on this front, such as a test based on constrained shuffling [78, 4] and the "knockoff" testing approach, which has been recently applied to sequential data [186, 187]. Exploring methods that can robustly test conditional dependence between time series is an important future direction.

In sum, an important and longstanding problem is that of nonparametrically testing the statistical significance of a correlation between autocorrelated datasets such as time series. The TTS test provides an approach that imposes relatively minimal requirements onto both the correlation statistic and the data-generating process. This gives researchers the freedom to apply a large arsenal of correlation statistics to a wide array of processes without sacrificing rigor. This freedom will become more valuable in the future as both correlation techniques and data availability continue to proliferate across diverse fields of research.

4 Methods

Surrogate data tests for comparative benchmarks

For surrogate data tests based on the IAAFT, stationary block bootstrap, or twin procedures, 499 surrogates were used for the empirical null distribution, unless specified otherwise. We used custom Python scripts to generate stationary block bootstrap surrogates, cyclic permutation surrogates, TTS surrogates, and naive TTS surrogates. We used the Pyunicorn package [188] to generate IAAFT and twin surrogates.

For IAAFT surrogates [47], we used the 'refined_AAFT_surrogates' function in the Pyunicorn package. We set the 'n_iterations' argument to 200 and the 'output' argument to 'true_spectrum'. For stationary bootstrap surrogates, we set the parameter known as p in [149] to 0.05. Cyclic permutation surrogates were generated by shifting in time with a wraparound. If the original time series was of length n, then $n-1$ cyclic permutation surrogate were produced. For example, if the original time series was $\{x_1, x_2, x_3, x_4\}$, then there would be 3 cyclic permutation surrogates: $\{x_2, x_3, x_4, x_1\}$, $\{x_3, x_4, x_1, x_2\}$, and $\{x_4, x_1, x_2, x_3\}$.

For twin surrogates [51], we used the 'twin_surrogates' function in the Pyunicorn package. This function requires four parameter arguments: the minimum temporal distance between twins, the delay vector lag, the delay vector embedding dimension, and the "recurrence threshold". We set the minimum distance between twins to 1. To choose the delay vector lag and embedding dimension, we used the function 'takens_embedding_optimal_parameters' in the Python package giotto-tda [189]. Briefly, this function first selects the delay lag τ so that the mutual information between values τ steps apart is minimized. The function then selects the embedding dimension by the widely used 'false nearest neighbors' algorithm, which is difficult to describe concisely, but is explained clearly by its inventors [190]. The function "takens_embedding_optimal_parameters" requires two arguments, the maximum delay lag and the maximum embedding dimension, and these were both set to 8. Finally, we chose the recurrence threshold parameter by a method advocated by the original authors of the twin procedure [51], which is to select the recurrence threshold that causes the recurrence plot to have between 5% and 20% "black points". We used 12% black

points as this is in the middle of the recommended range.

Following previous works [47, 43], we preprocessed time series to reduce a potential mismatch between the earliest and latest times before applying certain surrogate data tests. We used this preprocessing step (henceforth called 'circularization') for four surrogate procedures: IAAFT, cyclic permutation, stationary bootstrap, and twin. Circularization is recommended for IAAFT surrogates to avoid artifacts due to the periodic nature of Fourier components, and it is recommended for cyclic permutation surrogates because they directly join the beginning and ends of the time series [43]. We also used circularization for the stationary bootstrap and twin methods because these techniques also sometimes join the extremes of the time series.

To circularize a time series $\{y_1, y_2, ..., y_n\}$, we truncate it to $\{y_{k_{start}}, y_{k_{start}+1}, ..., y_{k_{end}-1}\}$, where k_{start} and k_{end} are chosen to minimize the mismatch between the beginning and end of the truncated time series using a formula quoted in [43]:

$$(k_{start}, k_{end}) = \underset{(k_1,k_2)}{\operatorname{argmin}} \left(\sum_{i=0}^{L} (y_{k_2+i} - y_{k_1+i})^2 \right). \tag{9}$$

We used $L = 10$. Additionally, to ensure that the circularized time series is not too short, we require that k_1 and k_2 be near the beginning and end of the time series respectively. Specifically, we impose the constraints $k_1 \leq 40$ and $n - L - k_2 + 1 \leq 40$.

When circularization was used, k_{start} and k_{end} were chosen based on the time series that was used to generate surrogates, but both the x and y time series were circularized using the same choice of k_{start} and k_{end}. Additionally, both the original and surrogate correlations were calculated from circularized time series. Circularization generally improved the false positive rates of tests (compare Fig 27 to Fig A38).

Parametric significance test

We used a parametric test described by [150]. Under the null hypothesis that two stochastic processes $\{x_1, ..., x_n\}$ and $\{y_1, ..., y_n\}$ are independent, [150] estimate the variance of the sample Pearson correlation coefficient (denoted $\hat{\sigma}_\rho^2$) as:

$$\hat{\sigma}_\rho^2 = \frac{\sum_{k=0}^{n-1} n_k \hat{C}_x(k) \hat{C}_y(k)}{n^2 \hat{\sigma}_x^2 \hat{\sigma}_y^2}$$

where $\hat{C}_x(k)$ is the estimated autocovariance of the time series x at lag k:

$$\hat{C}_x(k) = \frac{1}{n-k} \sum_{t=1}^{n-k} (x_t - \bar{x})(x_{t+k} - \bar{x}); \;\; \bar{x} = \frac{1}{n} \sum_{t=1}^{n} x_t$$

and $\hat{\sigma}_x^2$ is the estimated variance of time series x:

$$\hat{\sigma}_x^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2$$

and n_k is the number of entries A_{ij} in an n-by-n matrix A such that $|i-j| = k$. In other words, $n_k = n$ if $k = 0$ and $n_k = 2(n-k)$ if $0 < k < n$.

Since $\hat{\sigma}_\rho^2$ is estimated from finite data, it can under some circumstances be negative, which is nonsensical. If this occurs, $\hat{\sigma}_\rho^2$ is simply set to $1/n$, as is often recommended [150, 191, 41]. Note that $\hat{\sigma}_\rho^2 = 1/n$ corresponds to the case without autocorrelation (to see this, plug into the equation for $\hat{\sigma}_\rho^2$ the following: $\hat{C}_x(k) = \hat{\sigma}_x^2$ if $k = 0$ and $\hat{C}_x(k) = 0$ otherwise).

Next, the autocorrelation-corrected "effective sample size" is given by $\hat{m} = 1 + \hat{\sigma}_\rho^{-2}$, and a standard t-test of the Pearson correlation is performed using $\hat{m}$ in place of n. That is, the test statistic is $T = \hat{\rho}(\hat{m}-2)^{1/2}/(1-\hat{\rho}^2)^{1/2}$, where $\hat{\rho}$ is the sample Pearson correlation coefficient, and a two-tailed p-value is computed by comparing T to a Student's t-distribution with $\hat{m} - 2$ degrees of freedom [150]. This test is what is referred to as the "parametric test" unless otherwise stated.

Note that in the climatology example (Fig 29), time series had differing lengths, and we implemented the parametric test in a way that uses all of the data. We used a deglaciation event series between -2 kyr (past) and 0 kyr (present day), as well as orbital parameter time series between -12 kyr (past) and $+10$ kyr (future). For the parametric test, $\hat{\rho}$ was estimated from the overlapping data (from -2 to 0 kyr), the deglaciation autocovariance values were estimated from the entire deglaciation time series (from -2 to 0 kyr), and the autocovariance terms of the orbital parameters were estimated from the entire orbital parameter time series (from -12 to $+10$ kyr).

In supplementary data file 1, we compare the false positive rate of this test to those of three other parametric tests. In that file, "Test 1" refers to the test described above. "Test 2" refers to a variant of this test in which $\hat{C}_x(k)\hat{C}_y(k)$ is set to zero for $k > n/4$, as suggested by [41, 40] (but otherwise, all steps are the same as in Test 1). The rationale for this is that estimating $\hat{C}_x(k)$ for large values of k is difficult. "Test 3" and "Test 4" refer to recently described tests from [151] for Pearson correlation and mutual information respectively.

Mutual information

To estimate mutual information (except in Fig 31; see next paragraph), we used the internal function '_compute_mi_cc' in the Scikit-Learn package [192] (available at: `github.com/scikit-learn/scikit-learn/blob/2beed55847ee70d363bdbfe14ee4401438fba057/sklearn/feature_selection/_mutual_info.py#L18`),

which implements "estimator $I^{(1)}$" of [148]. The estimator requires a choice of distance metric for each of the variables being correlated, and one parameter (the number of neighbors). We used 3 neighbors and regular (i.e. Euclidean) distance for both variables.

For the fish behavior example in which we correlated speed with direction (Fig 31), the circular nature of the direction variable required a slightly different mutual information estimator. Specifically, we again used the $I^{(1)}$ estimator of [148] with 3 neighbors and used Euclidean distance for speed. For direction, we used angular distance so that, for instance, the angles of 0.1π and 1.9π would have a distance of 0.2π rather than 1.8π. Mathematically, we took the distance between two direction angles α and β to be:

$$\arccos\left(\cos(\alpha-\beta)\right) = \arccos\left(\cos(\alpha)\cos(\beta)+\sin(\alpha)\sin(\beta)\right).$$

This restricts the range of possible angular distances to $(0,\pi)$. We implemented the mutual information estimator using this distance metric as a custom Python script accelerated with the Numba compiler [193].

Appendix to Part II

A1 Mathematical justification of the truncated time-shift test

Appendix A1.1 provides a justification for the claim that the TTS test correctly controls the false positive rate as long as one of the time series under study is strictly stationary. Appendix A1.2 provides definitions and auxiliary results that support the arguments in Appendix A1.1, as well as a graphic introduction (Fig A34) to some important background concepts from probability theory.

A1.1 Proof of the truncated time-shift test

To make this proof more accessible, we have replaced some of the most tedious aspects of mathematical notation with two terms: "r-neighborhood" and "r-locally top-b". We begin with definitions for these terms.

Definition 1 *r-locally top-b*

- For a point a_k in a sequence, the "r-neighborhood" of a_k is the subsequence $\{a_{k-r}, ..., a_{k+r}\}$. That is, a_k's r-neighborhood contains all points that are no more than r steps away from a_k.
- A point a_k in a sequence is "r-locally top-b" if a_k is among the top b points within a_k's r-neighborhood. That is, a_k is r-locally top-b if no more than b points in a_k's r-neighborhood (including a_k itself) are greater than or equal to a_k. The edge case of a_k being r-locally top-0 (i.e. at most zero points in a_k's r-neighborhood are at least as large as a_k) never occurs.

Figure A32A-B illustrates these definitions and how they behave in the presence of ties. Note that in order for these definitions to make sense, the r-neighborhood must not "fall off the edge" of the sequence. For instance, if our sequence is $\{a_1, a_2, ..., a_{10}\}$, then it does not make sense to talk about the 3-neighborhood of a_8, as this would contain a_{11}, which does not exist. To avoid this problem, when writing about the r-neighborhood of an element a_k of a sequence $\{a_1, ..., a_m\}$, we will require $1 + r \leq k \leq m - r$.

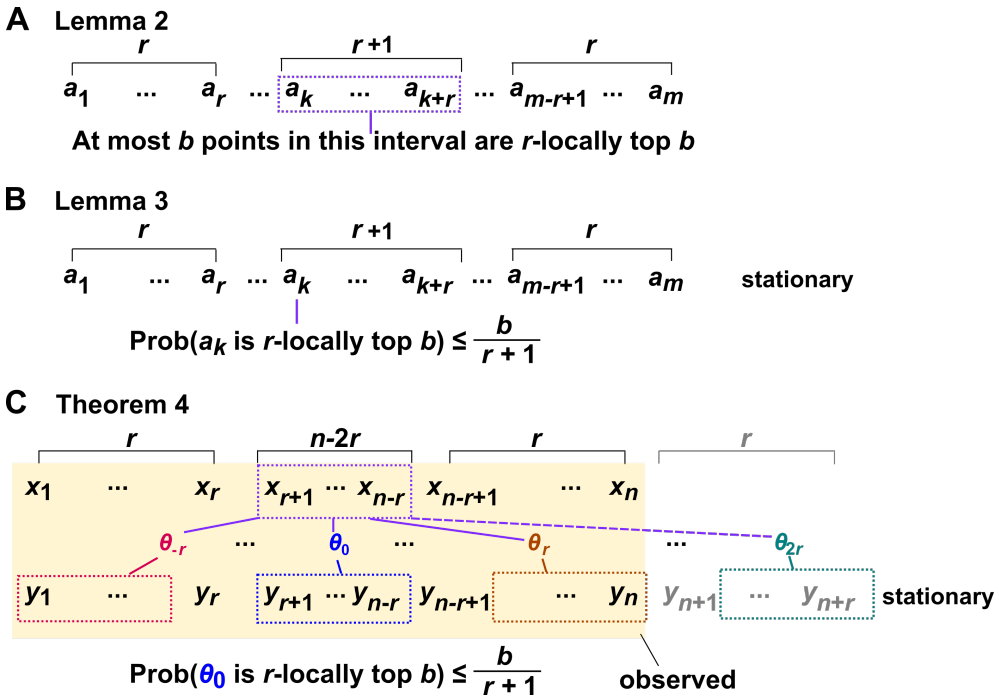


Figure A33: Illustration of elements of the proof.

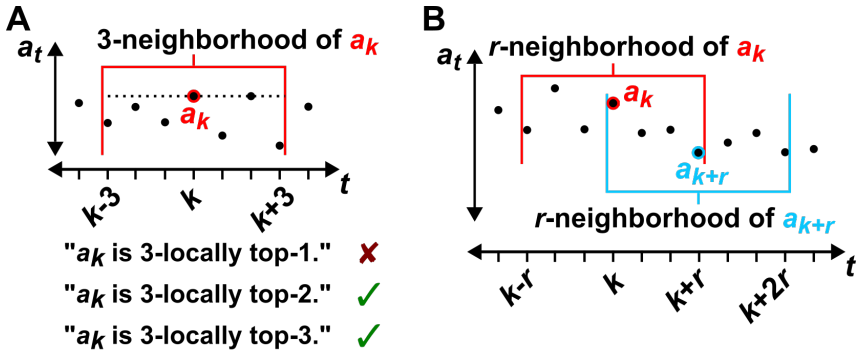


Figure A32: Illustration of "r-locally top-b". (A) The 3-neighborhood of a_k includes a_k itself as well as the three points to the left or right of a_k. In this example, a_k is not 3-locally top-1, but is 3-locally top-2, 3-locally top-3, 3-locally top-4 etc. (B) a_k is the leftmost point whose r-neighborhood contains a_{k+r} and a_{k+r} is the rightmost point whose r-neighborhood contains a_k. Thus, the r-neighborhood of any point between a_k and a_{k+r} contains the sequence $\{a_k, ..., a_{k+r}\}$. This fact is used in the proof of lemma 2.

Lemma 2

Let $\{a_1, a_2, ...a_m\}$ be a sequence. For nonnegative integer r, consider a subsequence $\{a_k, a_{k+1}, ..., a_{k+r}\}$ such that $1 + r \leq k \leq k + r \leq m - r$ (i.e. $1 + r \leq k \leq m - 2r$). Let b be a non-negative integer. Then, at most b points within the subsequence $\{a_k, a_{k+1}, ..., a_{k+r}\}$ are r-locally top-b.

This lemma is represented graphically by Fig A33A.

Proof: We prove the lemma by contradiction. Suppose that at least b+1 points within $\{a_k, a_{k+1}, ..., a_{k+r}\}$ are r-locally top-b. Let a_s be the smallest of these $b+1$ points (possibly tied for smallest). Then, $\{a_k, a_{k+1}, ..., a_{k+r}\}$ is within the r-neighborhood of a_s. (In detail, $s - r \leq k \leq k + r \leq s + r$ since $k \leq s \leq k + r$). Since there are at least $b+1$ points greater than or equal to a_s in a_s's r-neighborhood, a_s is not r-locally top-b. Thus, we have a contradiction, which completes the proof. We required $1+r \leq k \leq m-2r$ to ensure that the r-neighborhoods of $\{a_k, a_{k+1}, ..., a_{k+r}\}$ are properly defined.

Lemma 3

Let $\{a_1, a_2, ...a_m\}$ be a stationary sequence of random variables. Consider an element a_k such that $1 + r \leq k \leq m - 2r$ for some nonnegative integer r. Then, for some nonnegative integer b, the probability that a_k is r-locally top-b has an upper bound of $b/(r + 1)$.

This lemma is represented graphically by Fig A33B.

Proof: Define the variable $T_i(r, b)$ to indicate whether the ith point is r-locally top-b. Specifically, $T_i(r, b) = 1$ if a_i is r-locally top-b and $T_i(r, b) = 0$ otherwise. Since $1 + r \leq k \leq m - 2r$, lemma 2 says that at most b points within $\{a_k, a_{k+1}, ..., a_{k+r}\}$ are r-locally top-b. Written in terms of $T_i(r, b)$, this condition is:

$$b \geq \sum_{i=k}^{k+r} T_i(r, b).$$

Since $\{a_1, a_2, ...a_m\}$ is stationary, $\{T_k(r, b), T_{k+1}(r, b), ..., T_{k+r}(r, b)\}$ must also be stationary. (This fact is intuitive, but a rigorous proof is given by lemma 13.) Denote by $E[\cdot]$ the expected value of a random variable (i.e. the mean). Since the mean of a stationary sequence is independent of time, we have:

$$b = E[b] \geq E\left[\sum_{i=k}^{k+r} T_i(r, b)\right] = \sum_{i=k}^{k+r} E\left[T_i(r, b)\right] = (r + 1)E\left[T_k(r, b)\right].$$

So $E\left[T_k(r, b)\right] \leq b/(r + 1)$. But $E\left[T_k(r, b)\right]$ is the probability that a_k is r-locally top-b, so the proof is complete.

We are now ready to turn our attention to the theorem which justifies the TTS test. Since we will need to frequently mention specific sequences, we will make use of the abbreviation $\{a_1, a_2..., a_n\} = \{a_t\}_1^n$.

The statement of the following theorem parallels the TTS procedure closely, but with two main differences. The first difference is that we specify that the x_t and y_t terms are in $\mathbb{R}^l$ and $\mathbb{R}^k$. Essentially, this just means

that the time series are allowed to be multivariate and are not required to have the same dimension. In the simple case where we are dealing with two univariate time series, we can just set both l and k to 1. Alternatively, it is possible to correlate time series with different dimensions (i.e. $l \neq k$) using correlation functions based on distances or predictions ([194, 18, 127]). An example where this may be useful is in Fig A43, where a series of one-dimensional x_t terms is correlated to a two-dimensional time series whose terms are $v_t = (z_t, z_{t-1})$. In this example, a correlation is high if pairs of points that are nearby in v_t space have similar x_t values.

The second difference between the theorem below and the TTS procedure in the main text is that in the theorem we require the existence of $\{y_t\}_1^{n+r}$, despite only having data from $\{y_t\}_1^n$. This is a theoretical requirement. We do not need to obtain data from $\{y_{n+1}, ..., y_{n+r}\}$, but we must assume that the subsequence $\{y_{n+1}, ..., y_{n+r}\}$ exists in order to apply lemma 2.

Theorem 4 *Validity of the truncated time-shift surrogate test*

Let n ("the observed data length") and r ("the truncation radius") be integers such that $0 < 2r < n$. Let $\{x_t\}_1^n$ be a (possibly nonstationary) stochastic sequence where each term x_t is in $\mathbb{R}^l$. Let $\{y_t\}_1^{n+r}$ be a stationary stochastic sequence where each term y_t is in $\mathbb{R}^k$. Let the two sequences be independent of each other. Let Θ (the "correlation function") be a real-valued function maps a pair of sequences (the "truncated x series" $\{x_{1+r}, ..., x_{n-r}\}$ and the "shifted truncated y series" $\{y_{1+r+\delta}, ..., y_{n-r+\delta}\}$) to a real number. Moreover, let Θ be $(\mathbb{R}^{(l+k)\times(n-2r)}, \mathcal{B}_{(l+k)\times(n-2r)})$-measurable[1]. Define the "shifted correlations" θ_δ as

$$\theta_\delta = \Theta(\{x_{1+r}, ..., x_{n-r}\}, \{y_{1+r+\delta}, ..., y_{n-r+\delta}\}).$$

Let B be the number of terms in the sequence

$$\theta_{-r}, \theta_{-r+1}, ..., \theta_r$$

that are greater than or equal to θ_0. Then

$$P(B \leq b) \leq \frac{b}{r+1}$$

[1]Although we have tried to avoid potentially difficult concepts in this proof, measurability does appear here as a necessary condition. We give a brief explanation here and a formal definition in section A1.2. Here, the phrase "Θ is a $(\mathbb{R}^{(l+k)\times(n-2r)}, \mathcal{B}_{(l+k)\times(n-2r)})$-measurable function" can be translated as: "for any pair of stochastic sequences $\{x_{1+r}, ..., x_{n-r}\}$ and $\{y_{1+r+\delta}, ..., y_{n-r+\delta}\}$ where each term x_t is in $\mathbb{R}^l$ and each term y_t is in $\mathbb{R}^k$, it is guaranteed that $\Theta(\{x_{1+r}, ..., x_{n-r}\}, \{y_{1+r+\delta}, ..., y_{n-r+\delta}\})$ is a random variable with a well-defined cumulative distribution function". If Θ is not measurable, then we may not be able to define the probability distribution of $\Theta(\{x_{1+r}, ..., x_{n-r}\}, \{y_{1+r+\delta}, ..., y_{n-r+\delta}\})$, which is a serious problem for statistical analysis. Measurability is a theoretical requirement, not a practical one: Essentially any function we choose in a practical correlation analysis will be measurable unless we intentionally engineer it to be otherwise.

for any non-negative integer b.

This theorem is represented graphically by Fig A33C.

Proof: Since $\{x_t\}_1^n$ and $\{y_t\}_1^{n+r}$ are independent, and $\{y_t\}_1^{n+r}$ is stationary, the sequence

$$\theta_{-r}, \theta_{-r+1}, ..., \theta_{2r}$$

is also stationary (by theorem 12). Since θ_0 is a member of a stationary sequence and is flanked by r values to its left and $2r$ values to its right, we may apply lemma 3 to obtain:

$$\frac{b}{r+1} \geq P(\theta_0 \text{ is } r\text{−locally top}-b) = P(B \leq b)$$

which completes the proof.

Why does theorem 4 justify the TTS test? Recall that the TTS procedure defines the u statistic to be:

$$u = \frac{B}{r+1}$$

with B being the number of correlations (θ_δ's) that are greater than or equal to the unshifted correlation (θ_0) as in theorem 4. In the main text, we claimed that we can use u like a p-value. That is, for a significance level α, we have $P(u \leq \alpha) \leq \alpha$. To see why this is true, choose $b = \text{floor}(\alpha(r+1))$, where floor($a$) rounds a down to the nearest integer. Then, as long as the requirements of theorem 4 are met, we have:

$$\begin{aligned} P(u \leq \alpha) &= P\left(\frac{B}{r+1} \leq \alpha\right) \\ &= P(B \leq \alpha(r+1)) \\ &\leq P(B \leq \text{floor}(\alpha(r+1))) \\ &= P(B \leq b) \leq b/(r+1) \leq \alpha \end{aligned}$$

where we applied theorem 4 and the fact that $b \leq \alpha(r+1)$ in the last line.

A1.2 Background definitions and supporting proofs

Here we review relevant definitions and properties from probability theory. Note that our definition of stationarity departs from much of the literature in that we work with a notion of stationarity that applies to finite time series. The final two results of this section, theorem 12 and lemma 13, are auxiliary results that support the arguments made in Appendix A1.1. Definitions 5, 6, and 8 are illustrated graphically in Fig A34.

Definition 5 *Probability spaces (chapter 2 of [195])*

A probability space is any triple $(\Omega, \mathcal{F}, P)$ where Ω, $\mathcal{F}$, and P have the following names and meet the following requirements:

- Ω is called the sample space and is a nonempty set.
- $\mathcal{F}$ is called the σ-algebra and is a set of subsets of Ω. $\mathcal{F}$ contains Ω itself and the empty set $\emptyset$. $\mathcal{F}$ is closed under the formation of complements and countable unions, meaning that $\mathcal{F}$ contains the complements and countable unions of any elements of $\mathcal{F}$.
- P is called the probability measure and is a mapping from $\mathcal{F}$ to $[0, 1]$. $P(\emptyset) = 0$, $P(\Omega) = 1$, and P is countably additive.

In essence, $\mathcal{F}$ (also known as the "event space") is the set of events (each event being a set) to which we can assign a probability value.

We will now introduce the concept of Borel field on $\mathbb{R}$, a special type of σ-algebra generated from intervals on the real number line. This concept will be used when defining the random variable.

Definition 6 *The Borel fields on $\mathbb{R}$ and $\mathbb{R}^k$ (definitions 3.18 and 3.19 in [196]).*

The Borel field on $\mathbb{R}$, denoted $\mathcal{B}$, is the smallest set of sets that includes:

1. all intervals $\{b : -\infty < b \leq \alpha\}$ where $\alpha \in \mathbb{R}$;
2. the complement B^c of any set B in $\mathcal{B}$;
3. the countable union of any sequence $\{B_i\}$ in $\mathcal{B}$.

Elements of a Borel field are called Borel sets.

The Borel field on $\mathbb{R}^k$ (where $k < \infty$), denoted $\mathcal{B}_k$, is the smallest collection of sets that includes:

1. all intervals $\{b : -\infty < b \leq \alpha\}$ where $b \in \mathbb{R}^k$ and $\alpha \in \mathbb{R}^k$;

2. the complement B^c of any set B in $\mathcal{B}_k$;

3. the countable union of any sequence $\{B_i\}$ in $\mathcal{B}_k$.

Above, in the multivariate case, the "$\leq$" and "$<$" inequality symbols apply element-wise.

The Borel fields are also closed under intersections. This fact will be useful later.

Lemma 7

If $B_1, B_2, ..., B_n$ (where n is a positive integer) is a sequence of sets in $\mathcal{B}_k$, then the intersection $\cap_{i=1}^n B_i$ is also in $\mathcal{B}_k$.

Proof: For two sets A and B, A^C, $A \cup B$, and $A \cap B$ denote the operations of complementation, union, and intersection respectively. From De Morgan's laws we know that

$$(A \cap B)^C = A^C \cup B^C.$$

Equivalently,

$$A \cap B = (A^C \cup B^C)^C. \tag{10}$$

Thus, since $\mathcal{B}_k$ is closed under complementation and union, if sets A and B are in $\mathcal{B}_k$, then so is $A \cap B$. This is the special case for two sets. The lemma itself then follows from the inductive generalization of the two-set case. Let $B_1, B_2, ...$ be a sequence of sets in $\mathcal{B}_k$. Let $\mathcal{C}(n)$ be the condition $\cap_{i=1}^n B_i \in \mathcal{B}_k$. We will show by induction that $\mathcal{C}(n)$ holds for all $n = 1, 2, ...$ The base case of $\mathcal{C}(1)$ is immediately satisfied as $\cap_{i=1}^1 B_i = B_1$ is already in $\mathcal{B}_k$. For the inductive step, assume that $\mathcal{C}(s-1)$ holds, meaning that $\cap_{i=1}^{s-1} B_i \in \mathcal{B}_k$. Then, since $B_s \in \mathcal{B}_k$, we have:

$$\begin{aligned} \cap_{i=1}^s B_i &= (\cap_{i=1}^{s-1} B_i) \cap B_s \\ &= ((\cap_{i=1}^{s-1} B_i)^C \cup B_s^C)^C \in \mathcal{B}_k \end{aligned}$$

where the second line applied Eq. 10. Thus $\mathcal{C}(s)$ holds, completing the inductive step. We then conclude by induction that $\mathcal{C}(n)$ holds for $n = 1, 2, ...$, which proves the lemma.

Note that the empty set $\emptyset$ is in $\mathcal{B}_k$ as $\emptyset$ is the intersection of any set and its complement.

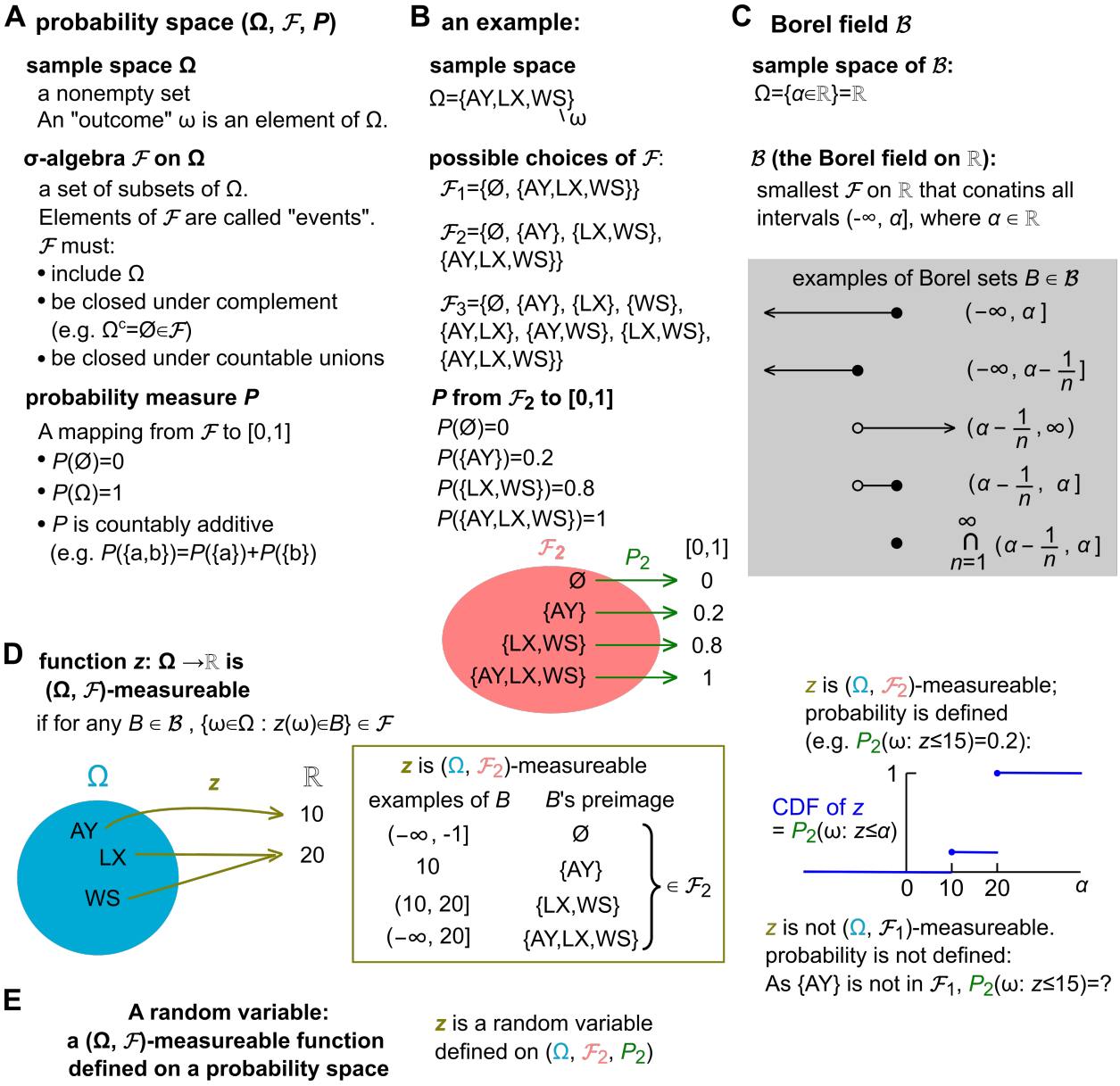


Figure A34: Some background concepts from probability theory. (**A**) A probability space includes a sample space Ω, a σ-algebra $\mathcal{F}$, and a probability measure P. (**B**) An example of a probability space. Note that the "event" {LX,WS} occurs if either of the "outcomes" LX or WS occurs. We encourage readers to check these examples against the definitions in (A). (**C**) The Borel field $\mathcal{B}$ and some elements of $\mathcal{B}$ (which are called Borel sets). α is a real number, and n is an integer. Rows 1 and 2 are elements of $\mathcal{B}$, as is seen immediately from $\mathcal{B}$'s definition. Row 3 is the complement of row 2. Row 4 is the intersection of rows 1 and 3. Row 5, the countable intersection of intervals $(\alpha - 1/n, \alpha]$ where n ranges from 1 to ∞ (see, for instance, section 1.4 of [197]). This demonstrates that a real number is also an element of $\mathcal{B}$. (**D**) Measurement function: definition and illustration. Although not pictured, z is also $(\Omega, \mathcal{F}_3)$-measurable. (**E**) Definition of a random variable.

The following concepts allow us to ensure that our models of probabilistic phenomena behave as they should (e.g. that their cumulative distribution functions [CDF]s exist; Fig A34D-E).

Definition 8 *Measurable functions, random variables, and stochastic sequences*

- A real-valued function $z : \Omega \to \mathbb{R}$ is called measurable on $(\Omega, \mathcal{F})$, or $(\Omega, \mathcal{F})$-measurable, if for any set B in $\mathcal{B}$,

$$\{\omega \in \Omega : z(\omega) \in B\} \in \mathcal{F}.$$

In other words, for any B in $\mathcal{B}$, its preimage (i.e. the set of all those ω in the sample space whose $z(\omega)$ value is in B) is an element of $\mathcal{F}$. The above definition of measurable functions is useful for proving results about functions already known to be measurable. However, another definition is equivalent (see page 29 of [195] or proposition 2.1.9 of [198]) and particularly useful for showing that a function is in fact measurable: A real-valued function $z : \Omega \to \mathbb{R}$ is measurable if for all $\alpha \in \mathbb{R}$,

$$\{\omega \in \Omega : z(\omega) \in (-\infty, \alpha]\} \in \mathcal{F}.$$

- A random variable is a $(\Omega, \mathcal{F})$-measurable real function defined on a particular probability space.
- A stochastic sequence is a family of random variables $\{z_1, z_2, ..., z_n\} = \{z_t\}_{t=1}^n$ where all such random variables are defined on the same probability space (i.e. they all share the same Ω and $\mathcal{F}$). Since in this work, we are dealing with finite-length stochastic sequences, we can alternatively think of a stochastic sequence as a vector-valued random variable.

Definition 9 *Stationary stochastic sequences*

A stochastic sequence $\{x_1, ..., x_n\}$ is stationary if for all triples (i, j, τ) such that $1 \leq i \leq i + \tau \leq n$ and $1 \leq j \leq j + \tau \leq n$, the joint distribution of $\{x_i, ..., x_{i+\tau}\}$ is the same as the joint distribution of $\{x_j, ..., x_{j+\tau}\}$.

Theorem 10 *Applying a measurable function to a stationary sequence produces a stationary sequence*

Let $\{x_1, ..., x_n\}$ be a stationary sequence of real-valued random variables. If $y_t = f(x_t, ..., x_{t+s})$ (where $1 \leq t \leq t + s \leq n$) for some $(\mathbb{R}^{s+1}, \mathcal{B}_{s+1})$-measurable function f, then $\{y_1, ..., y_{n-s}\}$ is stationary.

Proof: Let $B_0, B_1, ..., B_\tau$ be Borel sets in $\mathcal{B}$. We want to show that the joint distribution of y, $P\left((y_t \in B_0), (y_{t+1} \in B_1), ..., (y_{t+\tau} \in B_\tau)\right)$, is independent of time t (i.e. identical between $t = i$ and $t = j$).

$$\begin{aligned} P\left((y_i \in B_0), (y_{i+1} \in B_1), ..., (y_{i+\tau} \in B_\tau)\right) =& P\left(\left((x_i, ..., x_{i+s}) \in f^{-1}B_0\right), ..., \left((x_{i+\tau}, ..., x_{i+\tau+s}) \in f^{-1}B_\tau\right)\right) \\ =& P\left(\left((x_j, ..., x_{j+s}) \in f^{-1}B_0\right), ..., \left((x_{j+\tau}, ..., x_{j+\tau+s}) \in f^{-1}B_\tau\right)\right) \\ =& P\left((y_j \in B_0), ..., (y_{j+\tau} \in B_\tau)\right) \end{aligned}$$

The second line follows from the stationarity of $\{x_1, ..., x_n\}$. Note that $f^{-1}(\cdot)$ denotes the preimage of f, not the inverse of f. The indexing terms i, j, τ, s are defined to prevent subsequences from falling off the edge (i.e. $1 \leq i \leq i + \tau \leq n - s$ and $1 \leq j \leq j + \tau \leq n - s$).

The last two results of this section are not background theory. Rather, they are auxiliary results that are necessary to fully detail the arguments in Appendix A1.1. Specifically, theorem 12 is used in the proof of theorem 4. Lemma 13 is used in the proof of lemma 3. In other words, we do not attempt to motivate the statements given below, and simply provide them to be used as a reference for readers working through the proofs of theorem 4 and lemma 13.

Theorem 11 *Functions of independent random variables are independent (Proposition 3.2.3 of [195])*

Let $x \in \mathbb{R}^{m \times n}$ and $y \in \mathbb{R}^{m \times n}$ be random variables (or vectors of random variables, or matrices of random variables) such that x and y are independent. Let $f : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ and $g : \mathbb{R}^{m \times n} \to \mathbb{R}^{m \times n}$ be functions that are measurable on $(\mathbb{R}^{m \times n}, \mathcal{B}_{m \times n})$. Then, $f(x)$ and $g(y)$ are independent.

Proof: Let S_x and S_y be Borel sets in $\mathcal{B}_{m \times n}$. Suppose that x and y are independent. Then, considering the joint distribution of $f(x)$ and $g(y)$, we have:

$$\begin{aligned} P(f(x) \in S_x, g(y) \in S_y) &= P(x \in f^{-1}S_x, y \in g^{-1}S_y) \\ &= P(x \in f^{-1}S_x)P(y \in g^{-1}S_y) \\ &= P(f(x) \in S_x)P(g(y) \in S_y) \end{aligned}$$

Since the joint distribution of $f(x)$ and $g(y)$ is equivalent to the product of the marginal distributions, $f(x)$ and $g(y)$ are independent. This proof is given in [195] for the univariate case, but we repeat it here to stress that the same logic applies in the multivariate case.

Theorem 12

Let $\{x_1, ..., x_m\}$ and $\{y_1, ..., y_n\}$ be independent stochastic sequences, where $\{y_1, ..., y_n\}$ is stationary. The two sequences may be multivariate (i.e. $x_t \in \mathbb{R}^k$ and $y_t \in \mathbb{R}^r$). Construct a sequence $\{z_1, ..., z_{n-s}\}$ such that $z_t = f(\{x_1, ..., x_m\}, \{y_t, ..., y_{t+s}\})$ where $1 \leq t \leq t + s \leq n$ and f is $(\mathbb{R}^{mk+r(s+1)}, \mathcal{B}_{mk+r(s+1)})$-measurable. That is, each term z_t is a function of the entire x sequence and of a sliding window of size $s + 1$ within the y sequence (e.g. a correlation function of two time series). Then $\{z_1, ..., z_{n-s}\}$ is stationary.

Proof: The proof is similar in spirit to that of theorem 10. The difference is that here, f is a function of two series, and thus the preimage of f will consist of two sets. For a Borel set B in $\mathcal{B}$, let its preimage $f^{-1}B = A$, where A's x component looks like $\{x_1, ..., x_m\}$ (i.e. $A^x \subseteq \mathbb{R}^{km}$) and A's y component looks like $\{y_1, ..., y_n\}$ (i.e. $A^y \subseteq \mathbb{R}^{r(s+1)}$).

As before, we proceed by showing that the joint distribution of $\{z_i, ..., z_{i+\tau}\}$ is the same as the joint distribution of $\{z_j, ..., z_{j+\tau}\}$ (where i and j may differ). Let $B_0, B_1, ..., B_\tau$ be Borel sets in $\mathcal{B}$.

$$\begin{aligned} P\left(\left(z_i \in B_0\right), ..., \left(z_{i+\tau} \in B_\tau\right)\right) &= P\left(\left(\left\{x_t\right\}_1^m, \left\{y_t\right\}_i^{i+s}\right) \in f^{-1}B_0\right), ..., \left(\left(\left\{x_t\right\}_1^m, \left\{y_t\right\}_{i+\tau}^{i+\tau+s}\right) \in f^{-1}B_\tau\right)\right) \\ &= P\left(\left(\left\{x_t\right\}_1^m, \left\{y_t\right\}_i^{i+s}\right) \in A_0\right), ..., \left(\left(\left\{x_t\right\}_1^m, \left\{y_t\right\}_{i+\tau}^{i+\tau+s}\right) \in A_\tau\right)\right) \\ &= P\left(\left(\left\{x_t\right\}_1^m \in A_0^x, \left\{y_t\right\}_i^{i+s} \in A_0^y\right), ..., \left(\left\{x_t\right\}_1^m \in A_\tau^x, \left\{y_t\right\}_{i+\tau}^{i+\tau+s} \in A_\tau^y\right)\right) \end{aligned}$$

Since x and y series are independent, the above equation can be rewritten as

$$P\left(\left(z_i \in B_0\right), ..., \left(z_{i+\tau} \in B_\tau\right)\right) = P\left(\left(\left\{x_t\right\}_1^m \in A_0^x\right), ..., \left(\left\{x_t\right\}_1^m \in A_\tau^x\right)\right) P\left(\left(\left\{y_t\right\}_i^{i+s} \in A_0^y\right), ..., \left(\left\{y_t\right\}_{i+\tau}^{i+\tau+s} \in A_\tau^y\right)\right)$$

Since the y series is stationary, we have:

$$\begin{aligned} P\left(\left(z_i \in B_0\right), ..., \left(z_{i+\tau} \in B_\tau\right)\right) &= P\left(\left(\left\{x_t\right\}_1^m \in A_0^x\right), ..., \left(\left\{x_t\right\}_1^m \in A_\tau^x\right)\right) P\left(\left(\left\{y_t\right\}_j^{j+s} \in A_0^y\right), ..., \left(\left\{y_t\right\}_{j+\tau}^{j+\tau+s} \in A_\tau^y\right)\right) \\ &= P\left(\left(z_j \in B_0\right), ..., \left(z_{j+\tau} \in B_\tau\right)\right) \end{aligned}$$

Thus, z is stationary.

Lemma 13

Let $\{a_1, a_2, ...a_m\}$ be a stationary sequence of random variables. Let b and r be non-negative integers. Consider a subsequence $\{a_{1+r}, ..., a_{m-r}\}$ so that the r-neighborhood of each element is valid. For $1+r \leq k \leq m-r$, define $T_k(r, b) = \psi(a_{k-r}, ..., a_{k+r})$ to indicate whether a_k is r-locally top-b. Specifically, $T_k(r, b) = 1$ if a_k is r-locally top-b and $T_k(r, b) = 0$ otherwise. Then, $\{T_{1+r}(r, b), ..., T_{m-r}(r, b)\}$ is a stationary sequence of random variables.

Proof: According to Theorem 10, we can prove this theorem by showing that the function ψ is measurable on $(\mathbb{R}^{2r+1}, \mathcal{B}_{2r+1})$. To show that ψ is measurable on $(\mathbb{R}^{2r+1}, \mathcal{B}_{2r+1})$, it is sufficient (see definition 8) to show that for all $(-\infty, \alpha]$ where $\alpha \in \mathbb{R}$,

$$\{\overrightarrow{a_k} \in \mathbb{R}^{2r+1} : \psi(\overrightarrow{a_k}) \in (-\infty, \alpha]\} \in \mathcal{B}_{2r+1} \tag{11}$$

where $\overrightarrow{a_k} = \{a_{k-r}, ..., a_{k+r}\}$. We will also refer to $\{\overrightarrow{a_k} : \psi(\overrightarrow{a_k}) \in (-\infty, \alpha]\}$ as "the preimage" of $(-\infty, \alpha]$ under ψ.

There are three cases to consider. First, if $\alpha < 0$, then the preimage is simply the empty set $\emptyset$, since $T_k(r, b)$ is never less than 0. $\emptyset$ is an element of $\mathcal{B}_{2r+1}$, so the condition of Eq. 11 is satisfied. Second, if $\alpha \geq 1$, since $T_k(r, b) \leq 1$, Eq. 11 is again satisfied as the preimage then becomes $\mathbb{R}^{2r+1}$, which is an element of $\mathcal{B}_{2r+1}$.

Lastly, if $0 \leq \alpha < 1$, then the preimage is the set of all $\overrightarrow{a_k}$ where more than b elements are greater than or equal to a_k. But since one of these elements, namely a_k itself, is always equal to a_k, we can rephrase the above statement to be more useful for the subsequent arguments: *The preimage is the set of all $\overrightarrow{a_k}$ where at least b elements (other than a_k itself) are greater than or equal to a_k.* A trivial edge case is where $b = 0$, in which case the preimage is $\mathbb{R}^{2r+1} \in \mathcal{B}_{2r+1}$. When $b > 0$, we will need a more careful strategy to show that the preimage is in $\mathcal{B}_{2r+1}$.

Our strategy will be similar to the one used in Fig A34C (grey box) where we showed that a set consisting of a single real number is a member of $\mathcal{B}$. Analogously to that argument, here we will find some "starting sets" that are known to be in $\mathcal{B}_{2r+1}$ and show that one can arrive at the preimage by taking countable unions and intersections of these starting sets.

To build our intuition, let's first consider the specific setting where $r = 2, b = 3$. The preimage is then the set of all $\overrightarrow{a_k} = (a_{k-2}, a_{k-1}, a_k, a_{k+1}, a_{k+2})$ where at least 3 elements (other than a_k itself) are greater than or equal to a_k. We begin with "starting sets" $\mathcal{S}_i = \{\overrightarrow{a_k} : a_{k+i} \geq a_k\}$, where $i = -2, -1, 1, 2$. For example, $\mathcal{S}_1$ is the set of all $\overrightarrow{a_k}$s wherein $a_{k+1} \geq a_k$. To obtain some geometric intuition for this set, note that if $\overrightarrow{a_k}$ had been simply (a_k, a_{k+1}), then $a_{k+1} \geq a_k$ would correspond to the half plane on and above the line of identity $a_{k+1} = a_k$. $\mathcal{S}_1$ is in $\mathcal{B}_5$, because its complement (the set of all $\overrightarrow{a_k}$s wherein $a_{k+1} < a_k$) is an open set. Note that a set O is called "open" if for every point x in O, there is some positive "neighborhood radius" ϵ such that all of x's "neighbors" (i.e. points less than ϵ away from x) are also in O (e.g. page 385 of [198]). Since any open set is a Borel set (e.g. proposition 1.1.5 of [198]), $\mathcal{S}_1$ is in $\mathcal{B}_5$. By the same logic, all four $\mathcal{S}_i$s are in $\mathcal{B}_5$. Fig A35 shows how we can arrive at the preimage by taking unions and intersections of $\mathcal{S}_{-2}$, $\mathcal{S}_{-1}$, $\mathcal{S}_1$, and $\mathcal{S}_2$. The intersection $\mathcal{S}_{-2} \cap \mathcal{S}_{-1} \cap \mathcal{S}_2$, which is outlined in red and resembles a petal, is the set of all $\overrightarrow{a_k}$ where a_{k-2}, a_{k-1}, and a_{k+2} are all $\geq a_k$. Clearly this is a subset of the preimage because it is one way in which at least 3 elements (other than a_k) are $\geq a_k$. There are three other such triple intersections, which are the three other overlapping "petals" in Fig A35. The preimage itself, shown as the grey shaded area, is

the union of these four petals.

The same reasoning holds for other choices of b and r. In general, the starting sets are $\mathcal{S}_i = \{\overrightarrow{a_k} : a_{k+i} \geq a_k\}$, where $-r \leq i \leq r$ and $i \neq 0$. In words, $\mathcal{S}_i$ is the set of all $\overrightarrow{a_k}$s wherein $a_{k+i} \geq a_k$. By the same reasoning as above, these starting sets $\mathcal{S}_i$ are all elements of $\mathcal{B}_{2r+1}$. Any intersection of b starting sets is a situation where a_k is not r-locally top-b, and the preimage is given by the union of all such intersections. Since the preimage can be constructed from starting sets in $\mathcal{B}_{2r+1}$ by taking countable unions and intersections, the preimage must be in $\mathcal{B}_{2r+1}$ for this (final) case.

Overall, for all $\alpha \in \mathbb{R}$, the condition of Eq. 11 is satisfied, which in turn establishes that

$$\{T_k(r,b), T_{k+1}(r,b), ..., T_{k+r}(r,b)\}$$

is a stationary sequence of random variables, as required.

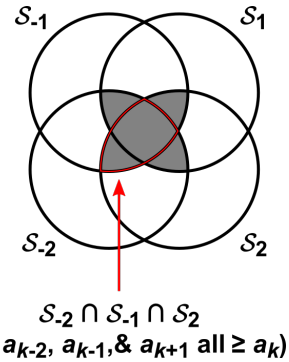


Figure A35: An example of how we can obtain the preimage by taking intersections and unions of starting sets $\mathcal{S}_i$. Here, $r = 2$ and $b = 3$, so the preimage (shaded in grey) is the set of all $\overrightarrow{a_k} = (a_{k-2}, a_{k-1}, a_k, a_{k+1}, a_{k+2})$ where at least 3 elements (other than a_k itself) are greater than or equal to a_k.

A2 Certain variants of the naive TTS test may be miscalibrated by more than twofold

In the main text we pointed out that the naive TTS test (Eq. 7) will never be miscalibrated by more than a factor of 2 as long as the time series used to produce surrogate data is stationary. In practice, the naive TTS test is not typically implemented exactly according to the procedure we describe, but instead several related variants are used [135, 136, 199, 137, 50]. Are variants of the naive TTS test also guaranteed to be miscalibrated by no more than twofold when applied to stationary time series? Here we describe two possible variants of the TTS test, and show that they can be miscalibrated by well over twofold. We use highly contrived examples in this section. Our purpose here is only to show that it is possible for these

variants to fail catastrophically, in contrast to the naive TTS procedure, which will not be miscalibrated by more than a factor of 2 (for stationary data). A study of how likely these failure modes are to occur in realistic systems is beyond the scope of this section.

One way in which applied works modify the time-shifting approach is by the use of what we call an "exclusion radius" [135, 136, 137, 50]. The idea is that the shifted time series that comprise the null model must all be shifted by more than some user-defined amount. We now describe this procedure in the context of the naive TTS test. As in Fig 26, we begin with two time series $\{x_t\}_{t=1}^n$ and $\{y_t\}_{t=1}^n$, choose a flanking radius r, and define the truncated sequences $x^{trunc} = \{x_{1+r}, ..., x_{n-r}\}$ and $y^{trunc}(\delta) = \{y_{1+r+\delta}, ..., y_{n-r+\delta}\}$. Also as in Fig 26, we define shifted correlations $\theta_\delta = \Theta(x^{trunc}, y^{trunc}(\delta))$. At this point we break with the standard recipe and choose an exclusion radius r_{ex}, which is a number that must be less than r. Rather than using all shifted y^{trunc} sequences to compute the naive p-value, we will only use those with a shift larger than r_{ex}. To express this idea as an equation, define $B_-(r_{ex})$ to be the number of shifted correlations within $\{\theta_{-r}, \theta_{-r+1}, ..., \theta_{-r_{ex}-1}\}$ that are greater than or equal to θ_0. Similarly define $B_+(r_{ex})$ to be the number of shifted correlations within $\{\theta_{r_{ex}+1}, \theta_{r_{ex}+2}, ..., \theta_r\}$ that are greater than or equal to θ_0. Finally the empirical p-value (corresponding as always to the null hypothesis that the two time series are independent) is written as

$$p = \frac{B_-(r_{ex}) + B_+(r_{ex}) + 1}{2(r - r_{ex}) + 1}.$$

Note that in special case where $r_{ex} = 0$, this procedure reduces to the naive TTS test (Eq. 7).

We now give a simulation example of a pair of independent stationary systems $\{x_t\}$ and $\{y_t\}$ for which this test (with $r_{ex} > 0$) is severely miscalibrated. The $\{x_t\}$ series is generated according to a periodic process afflicted with periodically varying measurement noise:

$$x_t = (1 - \lambda_t)b_t + \lambda_t \epsilon_t$$

We can think of b_t as a signal with a measurement strength of $(1 - \lambda_t)$ and we can think of ϵ_t as noise with strength of λ_t. The the noise terms ϵ_t are independently and identically distributed continuous uniform random variables between 0 and 1. The noise strength λ_t is given by:

$$\lambda_t = \begin{cases} 0.8 & \text{if } a_t \geq 450 \\ 0.5 & \text{if } a_t < 450 \end{cases}$$

$$a_t = \text{mod}(t + \phi, 950)$$

where $\mathrm{mod}(\alpha, \beta)$ is the modulo operation, which is the remainder obtained from dividing α by β. The phase term ϕ is chosen from $\{0, 1, ..., 949\}$ with equal chance. The initial signal value b_1 is chosen to be 0 or 1 with equal chance and all subsequent b_t values are given by $b_{t+1} = 1 - b_t$ (i.e. they alternate between 0 and 1). ϕ, b_1, and the ϵ_t terms are all independent. The $\{y_t\}$ series is an independent realization of the same process as the $\{x_t\}$ series (i.e. with independent choices of ϕ, b_1, and ϵ_t). Sample realizations are shown in Fig A36A.

To see that $\{x_t\}$ is stationary, note that $\{a_t\}$ is stationary because $\{a_t\}$ is a periodic process whose phase is chosen uniformly from throughout its period. Similarly, $\{b_t\}$ is stationary for the same reason. Next, $\{\lambda_t\}$ is stationary by theorem 10 since $\{\lambda_t\}$ is a static function of $\{a_t\}$. Lastly, since all three of $\{b_t\}$, $\{\lambda_t\}$ and $\{\epsilon_t\}$ are stationary and independent of one another, $\{x_t\}$ must be stationary.

We tested whether the two time series were dependent using a pair of length-2000 time series, a flanking radius of $r = 500$, and an exclusion radius of $r_{ex} = 450$. We used the sample Pearson correlation coefficient as the correlation statistic. In 43% of 1000 trials, dependence was detected at the 0.05 significance level (Fig A36B).

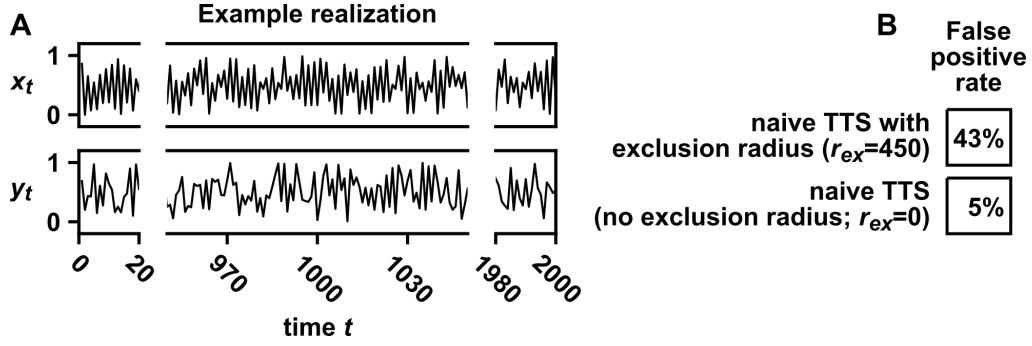


Figure A36: The exclusion radius procedure can cause the naive TTS test to be miscalibrated by more than twofold. (A) Sample dynamics of the benchmark system. See text for details. (B) False positive rates of the naive TTS test with or without an exclusion radius. For both tests, the truncation radius r was set to 500, the correlation statistic was the absolute value of the Pearson correlation coefficient, and significance was evaluated at the 0.05 level.

A second possible variant of the naive TTS test concerns possible delayed coupling. If two time series are coupled but with a delay, then the time shift approach may fail to detect a dependence relationship because shifted $\{y_t\}$ series may correlate more strongly to the $\{x_t\}$ series than the unshifted $\{y_t\}$ series. In the main text, our proposed solution is to pre-shift one of the two series (Fig 28). However, a seemingly natural alternative solution is the following: We define truncation radius r, x^{trunc}, $y^{trunc}(\delta)$, correlation function Θ, and shifted correlation $\theta_\delta = \Theta(x^{trunc}, y^{trunc}(\delta))$ as before, but now we choose a new parameter Δ (the coupling delay we expect to be in the system). Δ is the value of δ at which we expect the correlation θ_δ to be maximized. We then let B_Δ be the number of shifted correlations θ_δ ($-r \leq \delta \leq r$) that are as large as or

larger than θ_Δ. We then write the empirical p-value as:

$$p = \frac{B_\Delta}{2r+1}.$$

We call this the "naive TTS test with an expected coupling delay" and note that the naive TTS test is the special case of this test where $\Delta = 0$. Although we have not found this variant in the literature, we have been asked about it when presenting this work.

We now give a simulation example of a pair of independent stationary processes $\{x_t\}$ and $\{y_t\}$ for which this test is miscalibrated by more than twofold. Both are periodic process afflicted with periodically varying measurement noise. $\{x_t\}$ is given by:

$$x_t = (1 - \lambda_{x,t})b_{x,t} + \lambda_{x,t}\epsilon_{x,t}$$

$$t = 1, 2, ..., 400$$

where

$$\lambda_{x,t} = 0.1 + 0.1\left(\frac{\text{mod}(t + \phi_x, 3000)}{3000}\right)$$

$$b_{x,t} = 1 - b_{x,t-1}; b_{x,1}\text{chosen to be 0 or 1 with equal chance}$$

Next, $\{y_t\}$ is given by similar formulae, but where the equation for $\lambda_{y,t}$ has some differences in parameter choices:

$$y_t = (1 - \lambda_{y,t})b_{y,t} + \lambda_{y,t}\epsilon_{y,t}$$

$$\lambda_{y,t} = 0.3 + 0.5\left(\frac{\text{mod}(t + \phi_y, 3000)}{3000}\right)$$

$$b_{y,t} = 1 - b_{y,t-1}; b_{y,1}\text{chosen to be 0 or 1 with equal chance}$$

Here, ϕ_x and ϕ_y are chosen from among $\{0, 1, ..., 2999\}$ with equal chance. The $\epsilon_{x,t}$ and $\epsilon_{y,t}$ noise terms are independently and identically distributed continuous uniform random variables between 0 and 1. ϕ_x, ϕ_y,

$b_{x,1}$, $b_{y,1}$, and the $\epsilon_{x,t}$ and $\epsilon_{y,t}$ terms are all independent. As with the process of Fig A37, both $\{x_t\}$ and $\{y_t\}$ are stationary here as well; this can be seen using a line of argument analogous to the one used for the process of Fig A37. Sample realizations are shown in Fig A37A.

We tested whether the two time series were dependent using a naive TTS test with a flanking radius of $r = 100$ and various choices for the expected coupling delay Δ. We used the sample Pearson correlation coefficient as the correlation statistic. Fig A37B shows how the false positive rate varies as a function of the expected coupling delay. In this example, strongly negative values of Δ result in a false positive rate as high as 30%, even though detections were made at the 0.05 significance level.

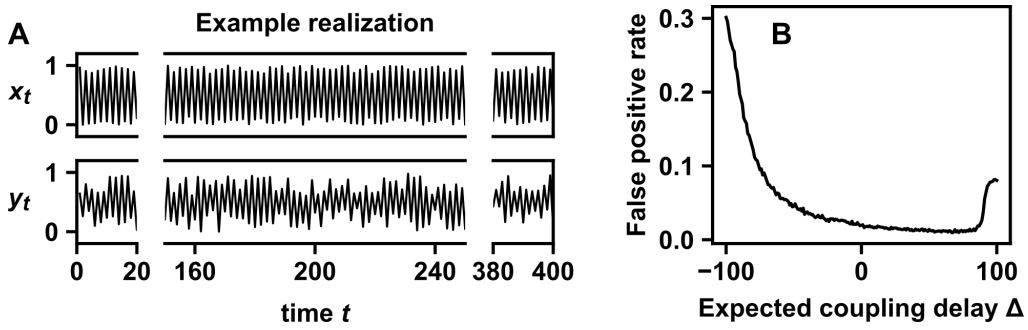


Figure A37: The naive TTS test with an expected coupling delay can be miscalibrated by more than twofold. (A) Sample dynamics of the benchmark system. See text for simulation details. (B) False positive rate of the naive TTS test as a function of the expected coupling delay Δ. The truncation radius r was set to 100 and the correlation statistic was the absolute value of the Pearson correlation coefficient. We varied the expected coupling delay Δ from -100 to 100. Detections were made at the 0.05 significance level. False positive rates were estimated from $10,000$ replicate trials.

A3 Detailed methods and results for the simulation benchmark

Here we describe the data-generating processes, surrogate data tests, and correlation statistics that we used in the benchmark study of false positive rates (Fig 27).

A3.1 Data-generating processes

For all data-generating processes except the sine wave with drifting measurement noise (Fig 27Aiii), the x_t series and y_t series were generated as independent replicates of the same process. Therefore, for all processes except that of Fig 27Aiii, we describe only the procedure to generate the x_t series.

Standard first-order autoregressive process (Fig 27A i)

$$t = 1, 2, ..., 400$$

$$x_{t+1} = 0.7x_t + \epsilon_t$$

where the ϵ_t terms are independent random variables with a standard normal distribution (i.e. with a mean of zero and variance of 1). The initial condition x_1 follows the stationary distribution of the process, which in this case is a normal distribution with a mean of zero and a standard deviation of $(1 - 0.7^2)^{-1/2}$ (see Eq. 20-6 in [142], for instance).

Logistic map (Fig 27A ii)

$$t = 1, 2, ..., 400$$

$$x_{t+1} = 4x_t(1 - x_t)$$

where x_1 is chosen according to the Beta$(0.5, 0.5)$ distribution. This system is stationary (see chapter 1 of [200], for instance).

Sine wave with measurement noise whose strength varies as a sawtooth wave (Fig 27A iii)

$$t = 1, 2, ..., 400$$

$$x_t = \sin\left(\frac{2\pi(t + \phi_x)}{22}\right)$$

$$y_t = \sin\left(\frac{2\pi t + \phi_y}{22}\right) + a_t(\epsilon_t - 0.5)$$

where

$$a_t = \frac{\text{mod}(t + \phi_a, 2800)}{7000}$$

where $\text{mod}(\alpha, \beta)$ is the modulo operation, which is the remainder obtained from dividing α by β. The terms ϕ_x, ϕ_y, ϕ_a and $\{\epsilon_t\}_{t=1}^{400}$ are all independent random variables with the following distributions: ϕ_x and ϕ_y are chosen uniformly at random from $\{0, 1, ..., 21\}$. ϕ_a is chosen uniformly at random from $\{0, 1, ..., 2799\}$. The ϵ_t terms are independent random variables drawn from a Beta$(1/3, 1/3)$ distribution. Note that the phase terms ϕ_x, ϕ_y, and ϕ_a are all time-invariant. More generally, we stress that throughout this section, all terms not indexed by t are time-invariant.

$\{x_t\}$ is stationary since it is a periodic sequence whose initial value is chosen uniformly from among the

points within a period. To see that $\{y_t\}$ is stationary, start with $\{a_t\}$. We know that $\{a_t\}$ is stationary because it is a periodic sequence whose initial value is chosen uniformly from among the points within a period. Since $\{\epsilon_t\}$ are iid (and thus stationary), and since any static function of two independent stationary sequences is stationary, the "$a_t(\epsilon_t - 0.5)$" term in the equation for y_t is also stationary. The sinusoid term in the equation for y_t is also stationary since it is a periodic sequence whose initial value is chosen uniformly from among the points within a period. Finally, $\{y_t\}$ must be stationary since it is the sum of two terms that are independent of each other and also individually stationary.

Sine wave with detection threshold (Fig 27A iv)

This process describes a sine wave with a detection threshold (0.5 and additive noise). See Fig A39 for more details.

$$t = 1, 2, ..., 400$$

$$x_t = a_t + \epsilon_t$$

$$a_t = \max(\sin(\phi + \frac{2\pi t}{35}), 0.5)$$

where ϕ is a continuous uniform random variable drawn from between 0 and 2π, and ϵ_t terms are independent normal random variables with mean of 0 and standard deviation of 0.05.

To see that $\{x_t\}$ is stationary, start with $\{a_t\}$. Note if ϕ is a continuous uniform random variable drawn from between 0 and 2π (as it is here), and if k is a real number, then $\sin(\phi)$ and $\sin(\phi + k)$ have the same distribution due to the periodicity of the sin function. Then it follows that $a_1, a_2, ..., a_n$ has the same distribution as $a_{1+t}, a_{2+t}, ..., a_{n+t}$ for any t. That is, $\{a_t\}$ is stationary. The $\{x_t\}$ process is stationary because it is a static function of two independent stationary processes ($\{a_t\}$ and $\{\epsilon_t\}$).

Coin flips with additive noise and time-varying 'heads' probability (Fig 27A v)

$$t = 1, 2, ..., 400$$

$$x_t = b_t + \epsilon_t$$

where $\{\epsilon_t\}_{t=1}^{400}$ are independent random normal variables with mean of 0 and standard deviation of 0.15. The terms b_t are Bernoulli random variables ('coin flips') with probability:

$$P(b_t = 0) = 1 - a_t$$
$$P(b_t = 1) = a_t$$

where a_t is given by:

$$a_t = \frac{1}{2}\left(\frac{1}{2}\sin\left(\phi_1 + \frac{2\pi t}{75}\right) + \frac{1}{2}\right)^6 + \frac{1}{12}\left(\frac{1}{2}\sin\left(\phi_2 + \frac{2\pi t}{31\sqrt{2}}\right) + \frac{1}{2}\right)$$

and where ϕ_1 and ϕ_2 are independent continuous uniform random variables drawn from between 0 and 2π.

To see that $\{x_t\}$ is stationary, note that both additive terms in $\{a_t\}$ are stationary because they are each periodic functions whose phase is uniformly distributed over the period, similar to the sine wave in system iv. Then, $\{a_t\}$ is itself stationary because it is a function of two independent stationary sequences. Then, $\{x_t\}$ is stationary because it is a function of $\{a_t\}$ and $\{\epsilon_t\}$, which are also two independent stationary sequences.

Exponential growth with periodic extinction (Fig 27A vi) Consider a population with a size of a_t that experiences constant immigration (50 each time step) together with exponential growth punctuated with periodic extinction (every 8 time units):

$$a_{t+1} = \begin{cases} a_t + 0.2a_t + 50 & \text{if } t \text{ is not a multiple of 8} \\ 50 & \text{if } t \text{ is a multiple of 8} \end{cases}$$

$$a_1 = 50.$$

Suppose that the population has already been in existence for a random amount of time when observations begin. In this case, the process during the observation window may be $b_1, ..., b_{400}$:

$$b_t = a_{t+\phi}$$

where ϕ is an integer drawn uniformly at random from among $\{0, 1, ..., 7\}$.

Additionally, only about 10 percent of the population is observed, via a binomial sampling process, which we approximate using a normal distribution:

$$x_t = 0.1b_t + (0.09b_t)^{1/2}\epsilon_t$$

where the ϵ_t terms are independent normal random variables with mean of 0 and variance of 1. As usual, $x_1, ..., x_{400}$ is the reported time series.

To see that $\{x_t\}$ is stationary, note that $\{b_t\}$ is stationary because it is a periodic process whose phase is chosen uniformly at random from among the period. Then, $\{x_t\}$ is stationary because it is a static function of $\{b_t\}$ and $\{\epsilon_t\}$, which are two independent stationary processes.

FitzHugh-Nagumo model with noise (Fig 27A vii)

$$(x_{t+1} - x_t)/0.7 = x_t - x_t^3/3 - w_t + \epsilon_t$$
$$(w_{t+1} - w_t)/0.7 = 0.08(x_t + 0.7 - 0.8w_t)$$

where the ϵ_t terms are independent random variables with standard normal distribution. We started the simulation with the initial condition $(x_1, w_1) = (0, 0)$ and ran the system for 1000 time points to allow it to equilibrate, using points 1001 through 1400 for the statistical benchmark.

Chaotic Lotka-Volterra (Fig 27A viii)

$$\frac{ds_i(t)}{dt} = 1.5r_i \left(s_i(t) - \sum_{j=1}^{4} a_{ij} s_i(t) s_j(t) \right), \; i = 1, 2, 3, 4$$

$$r = \begin{bmatrix} 1 \\ 0.72 \\ 1.53 \\ 1.27 \end{bmatrix}$$

$$a = \begin{bmatrix} 1 & 1.09 & 1.52 & 0 \\ 0 & 1 & 0.44 & 1.36 \\ 2.33 & 0 & 1 & 0.47 \\ 1.21 & 0.51 & 0.35 & 1 \end{bmatrix}$$

Unlike the other systems which are discrete-time and mostly stochastic, this is a system of differential

equations. These parameter choices are due to [146]. Although this system is continuous-time, statistical analysis was performed on discrete time points. We initialized the system by independently choosing $s_1(1)$, $s_2(1)$, $s_3(1)$, and $s_4(1)$ from a continuous uniform distribution between 0.1 and 0.5. We then numerically integrated the system for 2000 time units to allow it to equilibrate and finally used $t = 2001, 2002, ..., 2400$ for the statistical benchmark. The x_t and y_t series were independent realizations of $s_4(t)$.

Random walk (Fig 27A ix)

$$t = 1, 2, ..., 400$$

$$x_{t+1} = x_t + \epsilon_t$$

where the ϵ_t terms are independent random variables with a standard normal distribution. The initial condition x_1 was set to 0.

First-order autoregressive process with trend (Fig 27A x)

This system was generated by adding $t/60$ to the stationary first-order autoregressive process. Specifically,

$$t = 1, 2, ..., 400$$

$$x_t = a_t + t/60$$

where

$$a_{t+1} = 0.7a_t + \epsilon_t$$

Here the ϵ_t terms are independent random variables with a standard normal distribution. The initial condition a_1 is given by a normal distribution with a mean of zero and a standard deviation of $(1-0.7^2)^{-1/2}$.

A3.2 Hypothesis testing

Correlation statistics and surrogate data tests and the parametric tests were implemented as described in Methods. For the naive TTS test (Eq. 7) we set the flanking radius r to 50. For the TTS test (Eq. 8) we set the flanking radius r to 59, since the power of the TTS test is maximized when r is set to one less than a multiple of 20, as in Fig 28.

A3.3 False positive rates of surrogate data tests without circularization

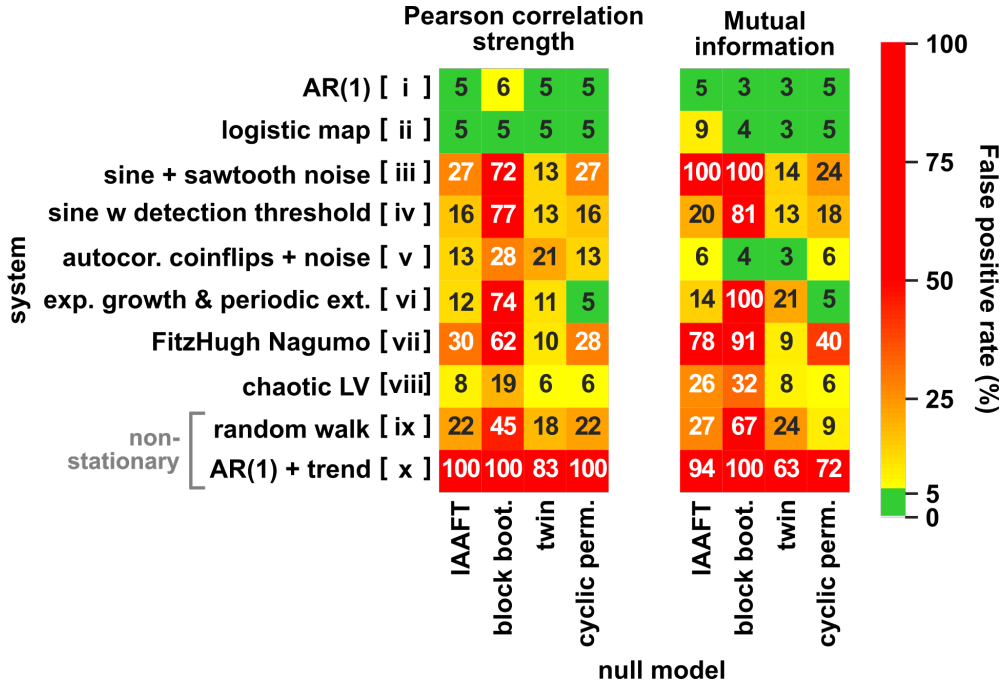


Figure A38: False positive rates of dependence tests without circularization. The same benchmark analysis was performed as in Fig 27 for IAAFT, block bootstrap, twin, and cyclic permutation tests, except that here, no circularization step was performed. Only these four surrogate tests are shown here because they were the only tests that used a circularization preprocessing step 27.

A3.4 Challenges for cyclic permutation surrogates and circularization preprocessing

The cyclic permutation method is fairly similar to time-shift methods, but lacks the guarantee of the TTS test (as demonstrated in Fig 27). Because the two procedures are similar, we here try to gain some intuition for their sometimes different behavior by exploring specific ways in which stationary sequences can cause the cyclic permutation test to become invalid.

The cyclic permutation procedure is valid if: 1) the true process is periodic; 2) the initial condition was chosen at random from among the points within a period with equal chance, because that way, the ground truth process will be reflected by the cyclic permutation process (wherein each position of the period is sampled equally by surrogates); and 3) the length of the truncated time series is an integer multiple of the period.

Since cyclic permutation surrogates directly join the ends of sequences, it is intuitive that problems may arise when the beginning and end of a time series look very different, and therefore we may expect the cyclic permutation procedure to benefit substantially from circularization preprocessing [48, 43]. This occurs most

strikingly in system vii (Fig 27), where the cyclic permutation test of mutual information has a false positive rate of either 5% with circularization (Fig 27) or 41% (Fig A38) without it. Thus, time series that are difficult to properly preprocess may pose challenges to the cyclic permutation method. Below, we illustrate two such examples. In both, we use the cyclic permutation surrogate test with circularization with the same parameters as in Fig 27.

First, a noise process or thresholding process can prevent the circularization procedure from selecting the optimal trimmed sequence length. As an example of this, consider a process in which $\{x_t\}$ and $\{y_t\}$ are generated by stationary sine waves with a detection threshold and additive noise:

$$x_t = \max(\sin(\phi_x + \frac{2\pi t}{35}), \beta) + \epsilon_{x,t}$$
$$y_t = \max(\sin(\phi_y + \frac{2\pi t}{35}), \beta) + \epsilon_{y,t}$$

Here, ϕ_x and ϕ_y are independent uniform random variables between 0 and 2π. The noise terms $\epsilon_{x,t}$ and $\epsilon_{y,t}$ are independent normal random variables with zero mean and standard deviation of 0.05. The parameter β is a threshold term. Example dynamics are shown in Fig A39A.

We generated independent $\{x_t\}$ and $\{y_t\}$ time series from this system and computed the proportion of trials in which the cyclic permutation test reported a significant Pearson correlation coefficient strength at the 0.05 significance level (Fig A39B purple). As the threshold increased, so did the false positive rate. This is presumably because at higher threshold values such as 0.5, much of the periodicity of the series is masked by the threshold, and so it is more difficult to determine the optimal length of the trimmed sequence via circularization (Eq. 9). Indeed, because the period of $\{y_t\}$ is 35, it would be natural to trim the sequence to a length that is a multiple of 35; this is nearly always achieved for low threshold values, but rarely for high threshold values (Fig A39C). We can confirm that the increase in the false positive rate is due to a suboptimal trimming length because when we analyze only the trials in which the trimmed length was a multiple of 35, the false positive rate is correctly set at 0.05 regardless of the threshold value (Fig A39B, red). In this particular case, one may try to avoid this problem by estimating the period by other means, but this example is intended to illustrate the general idea that detection thresholds and noise processes may pose a challenge to circularization methods.

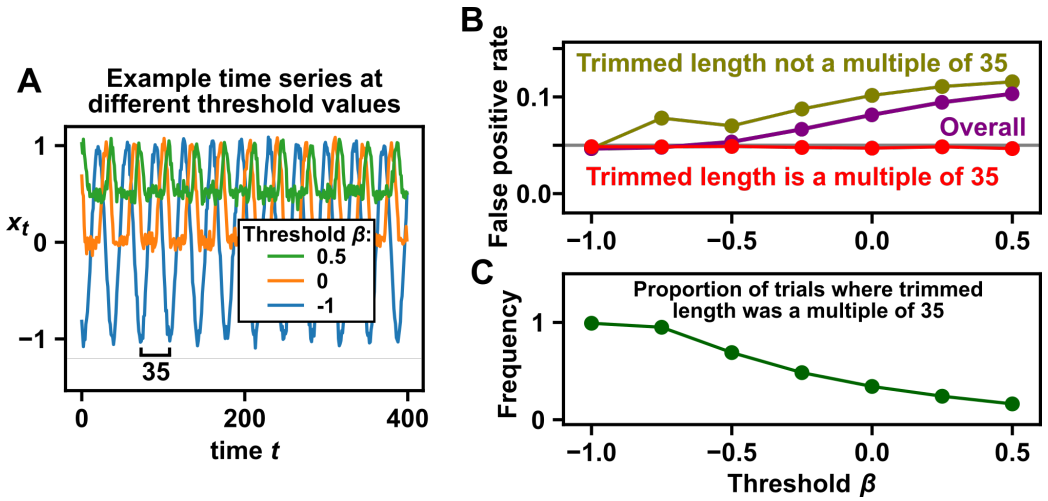


Figure A39: A process with a detection threshold is difficult for setting trimming length and thus poses a challenge for the cyclic permutation surrogate procedure. (A) Example time series of the thresholded sine process. A representative time series is shown for each of three possible threshold values. (B) False positive rate as a function of the detection threshold and the length of the trimmed time series. Specifically, we performed 10^5 trials in which two time series were generated with the same threshold β. We used the cyclic permutation procedure (with the circularization trimming step) to test for the significance of the Pearson correlation strength at the 0.05 level. For all thresholds, there were at least 16,000 trials in which the trimmed length was a multiple of 35. (C) The proportion of trials in which the trimmed (i.e. circularized) time series had a length that was a multiple of 35. The trimmed length is optimally a multiple of 35 because the dynamics have a period of 35.

A second potential challenge occurs when the method used to select the cutoff points (k_{start} and k_{end} in Eq. 9) is not appropriate for the time series under study. For instance, Eq. 9 chooses (k_{start}, k_{end}) as the values of (k_1, k_2) that minimize the score:

$$\sum_{i=0}^{L} (y_{k_2+i} - y_{k_1+i})^2.$$

Although this score will identify subsequences at the beginning and the end of a time series that share similar values, these subsequences may not share other aspects of their distribution.

As a simple illustrative example, consider a system that toggles between a random even integer upper-bounded by the even number β and a random odd integer upper-bounded by the odd number $\beta - 1$. Specifically, let the phase term ϕ be either 0 or 1 with equal chance. Then, for $t = 1, ..., 400$, if t has the same even/odd status as ϕ (i.e. t and ϕ are either both even or both odd), x_t will be a random even integer between 0 and β; alternatively, if t does not have the same even/odd status as t, x_t will be a random odd integer between 1 and $\beta - 1$. Lastly, because we plan to use mutual information to correlate two realizations of this process, we add a tiny amount of measurement noise (a uniform random variable between -10^{-8} and 10^{-8}) to each data point, as recommended by [148] for data sets with "tied" values. Fig A40A shows two

examples of this process for $\beta = 6$. Note that this process is stationary.

We estimated the mutual information between an independent pair of such processes with the same value of β and used the cyclic permutation procedure to test for significance at the 0.05 level. Since the process has a period of two, the circularization step should ideally trim the time series to an even length; doing so is both necessary and sufficient to ensure a false positive rate of 5% (Fig A40C). When $\beta = 2$, the possible even values (either 0 or 2) can be easily distinguished from the possible odd value (1 only), and thus the circularization procedure successfully trims the series to an even length, leading to a well calibrated false positive rate (Fig A40B, $\beta = 2$). As β becomes larger (e.g. 6), the possible even values (0, 2, 4, or 6) cannot be easily distinguished from the possible odd values (1,3, or 5), and thus circularization will sometimes trim the series to an odd length, leading to a high false positive rate (Fig A40B, $\beta = 6$). When β exceeds 10, the false positive rate begins to decline. This decline is not due to a decrease in odd-trimmed sequences (green curve in B), but instead due to a decrease in the false positive rate among odd-trimmed sequences (olive curve in C). The explanation for this decrease is omitted here because it is complicated and is beside the main point of this example.

Overall, when time series have important fluctuations in aspects of their distributions that are not tied to distances between points, it may be difficult for the circularization scores such as Eq. 9 to select the optimal trimming length.

We re-emphasize that both of these examples used stationary processes, and are therefore guaranteed to be appropriate for the TTS test. Thus, these challenges are relevant to the circular permutation approach (and perhaps other techniques that rely on circularization), but will not result in inflated false positive rates under the TTS procedure.

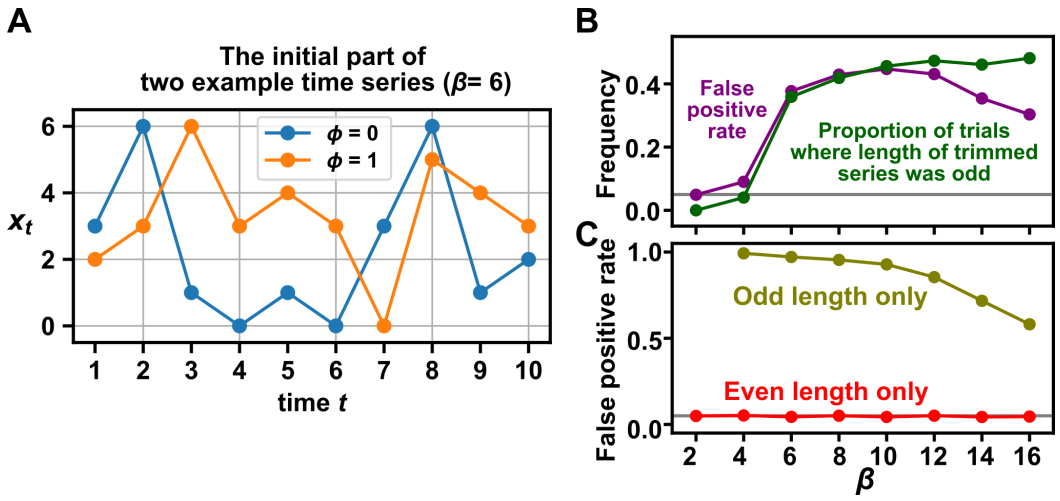


Figure A40: A process with important dynamics that are not closely tied to interpoint distances frustrates a circularization procedure based on minimizing the distance between the beginning and the ending subsequences of a series. (A) A process x toggles between a random even number (between 0 and β) and a random odd number (between 1 and $\beta - 1$) at each time step. See text for details. (B) We simulated an independent pair of realizations from this process at different values of β and tested for significant mutual information at the 0.05 level using the cyclic permutation procedure (including the circularization step). We performed 10^4 trials for each β. The false positive rate and the frequency of obtaining an odd-length trimmed sequence are shown. (C) The false positive rate as a function of the length of the trimmed time series. In both (B) and (C), a grey line indicates the 0.05 level.

A4 Surrogates for some nonstationary time series: The detrend-retrend TTS test

In this section, we describe a modified TTS procedure to generate surrogates for certain nonstationary time series. We first give some intuition for the procedure and when it is valid. We then describe the procedure precisely and show why it results in a valid test for independence. Lastly we show a simulation example in which it outperforms simpler approaches.

Although the TTS test as described in Fig 26 requires that surrogates are generated from a stationary time series, it is possible to modify the TTS test for nonstationary time series that can be correctly decomposed into a deterministic trend component and a stationary component. The basic idea is that we can: (1) obtain the stationary component by subtracting the deterministic trend, (2) generate time-shifted surrogates from the stationary component, and (3) add the trend back to each of the time-shifted surrogates (Fig A41). This idea has previously been applied to random phase surrogates [154].

A4.1 Detailed description of the detrend-retrend TTS test

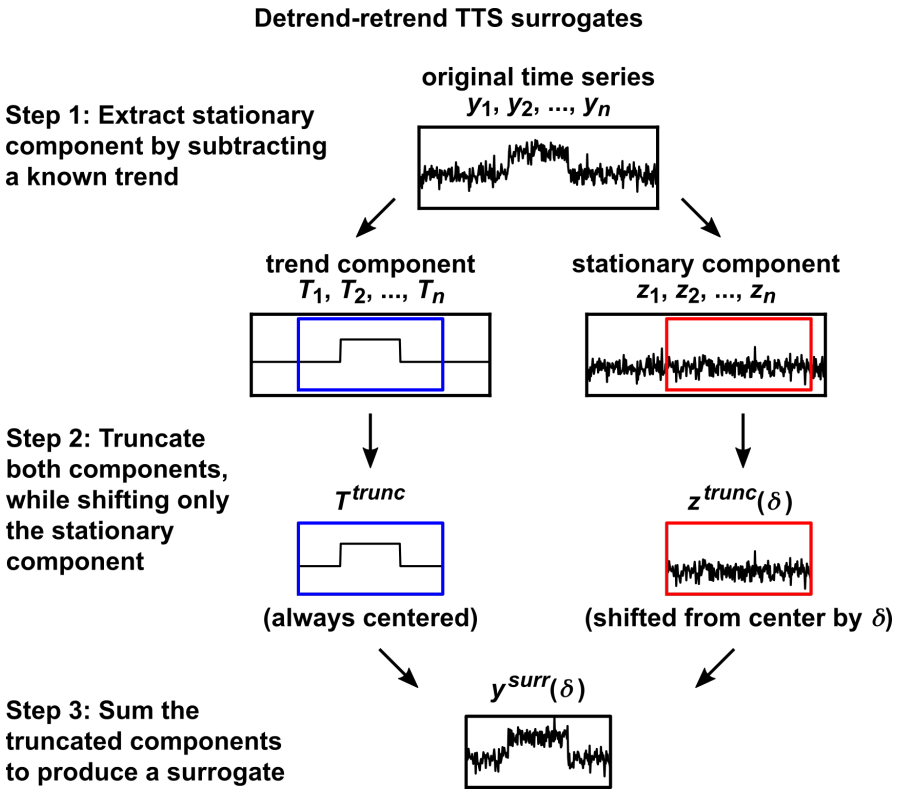


Figure A41: Procedure to generate detrend-retrend TTS surrogates.

We now describe this procedure (the "detrend-retrend TTS test") in detail. Suppose we have two time series $\{x_t\} = \{x_1, x_2, ..., x_n\}$ and $\{y_t\} = \{y_1, y_2, ..., y_n\}$, and we wish to test whether they are independent. As in the regular TTS test (Fig 26), we choose a truncation radius r, a correlation function Θ, and a significance level α. Suppose now that $\{y_t\}$ is not stationary, we know how to decompose it into a deterministic trend component and a stationary component. That is,

$$y_t = T_t + z_t \text{ for all } t = 1, 2, ..., n$$

where $\{T_1, ..., T_n\}$ is deterministic, $\{z_1, ..., z_n\}$ is stationary. We now produce truncated sequences from $\{x_t\}$, $\{T_t\}$ and $\{z_t\}$:

$$x^{trunc} = \{x_{1+r}, ..., x_{n-r}\}$$

$$T^{trunc} = \{T_{1+r}, ..., T_{n-r}\}$$

$$z^{trunc}(\delta) = \{z_{1+r+\delta}, ..., z_{n-r+\delta}\}$$

where, as before, δ takes on integers between $-r$ and r. Next, we produce y surrogates $y^{surr}(\delta)$ as the element-wise sum of $z^{trunc}(\delta)$ and T^{trunc} (Fig A41). That is, for each value of δ, $y^{surr}(\delta)$ is the $(n-2r)$-length sequence whose ith element is given by:

$$T_{i+r} + z_{i+r+\delta}.$$

(We call these surrogates $y^{surr}(\delta)$ instead of $y^{trunc}(\delta)$ because they cannot be obtained by simply truncating $\{y_t\}$ at different shifts.) Next, we use these surrogates in the same way as before: Obtain the shifted correlation values

$$\theta_\delta = \Theta(x^{trunc}, y^{surr}(\delta)).$$

Let B be the number of terms in the sequence $\{\theta_{-r}, ..., \theta_r\}$ that are greater than or equal to θ_0. Let

$$u = B/(r+1).$$

Finally, the test detects dependence between $\{x_t\}$ and $\{y_t\}$ at a significance level of α if $u \leq \alpha$.

Let us see why the detrend-retrend TTS procedure is a valid test of dependence between $\{x_t\}$ and $\{y_t\}$ (as long as $\{y_t\}$ is the sum of a known deterministic component and a stationary component): First, note that the above procedure is equivalent to applying the (regular) TTS test for dependence between $\{x_t\}$ and the stationary component $\{z_t\}$, but with a special correlation function $\tilde{\Theta}$:

$$\tilde{\Theta}(x^{trunc}, z^{trunc}(\delta)) = \Theta(x^{trunc}, T^{trunc} + z^{trunc}(\delta)).$$

Thus, since $\{z_t\}$ is stationary, this procedure rigorously tests for dependence between $\{x_t\}$ and $\{z_t\}$. Next, since $\{y_t\} = \{T_t\} + \{z_t\}$ and since $\{T_t\}$ is nonrandom by definition, we know that $\{x_t\}$ and $\{z_t\}$ are dependent if and only if $\{x_t\}$ and $\{y_t\}$ are dependent. (Since this fact may not be immediately obvious, we have formally proven it below as lemma 14.) Thus, by testing for dependence between $\{x_t\}$ and $\{z_t\}$, the detrend-retrend TTS procedure also tests for dependence between $\{x_t\}$ and $\{y_t\}$, as promised.

Lemma 14 *Let $\{x_t\}$, $\{y_t\}$, and $\{z_t\}$ be sequences of n random variables such such that $\{y_t\} = \{T_t\} + \{z_t\}$, where $\{T_t\}$ is a sequence of nonrandom real numbers. Then, $\{x_t\}$ and $\{y_t\}$ are independent if and only if $\{x_t\}$ and $\{z_t\}$ are independent.*

Proof: Let us begin with a notational change. Since our sequences have a finite length, we may instead represent them as vector-valued variables. Thus, let us rewrite our respective sequences as the length-n random vectors $\vec{x}$, $\vec{y}$, $\vec{z}$, and the length-n nonrandom vector $\vec{T}$. The proof essentially consists of applying 11 from Appendix A1.2. We first show the forward direction. To do so, suppose that $\vec{x}$ and $\vec{y}$ are independent. Define the functions $f(\vec{a}) = \vec{a}$ and $g(\vec{a}) = \vec{a} - \vec{T}$. By theorem 11, $f(\vec{x})$ and $g(\vec{y})$ are independent, but $f(\vec{x}) = \vec{x}$ and $g(\vec{y}) = \vec{z}$, so $\vec{x}$ and $\vec{z}$ are independent, as required. For the second reverse direction, suppose that suppose that $\vec{x}$ and $\vec{z}$ are independent. Define the function $h(\vec{a}) = \vec{a} + \vec{T}$. By theorem 11, $f(\vec{x}) = \vec{x}$ and $h(\vec{z}) = \vec{y}$ are independent, as required.

A4.2 With nonstationary data, the detrend-retrend TTS test is sometimes, but not always, superior to simpler alternatives.

We compared various flavors of the TTS test to simulated systems in which one or both series are nonstationary. We compared detrend-retrend TTS to two simpler strategies: directly applying the TTS test, or applying the TTS test after detrending the $\{y_t\}$ series (without retrending). The detrend-retrend test was superior in the case where both $\{x_t\}$ and $\{y_t\}$ are nonstationary, whereas the detrending-only strategy was superior where $\{x_t\}$ is stationary and $\{y_t\}$ is nonstationary. Below, we describe the details of the simulations.

We first simulated a system where $\{x_t\}$ and $\{y_t\}$ have related nonstationary behavior. We avoided using a system where the key problem addressed by the detrend-retrend TTS test is trivial: If $\{x_t\}$ and $\{y_t\}$ share a known common trend, then one could simply subtract this shared trend from both time series and proceed with the TTS test, obviating the need to "retrend". Avoiding that trivial case, $\{y_t\}$ has a visually apparent additive trend, but $\{x_t\}$ does not (Fig A42A).

For $t = 1, 2, ..., 400$, we generated $\{x_t\}$ and $\{y_t\}$ as follows:

$$H_t = \begin{cases} 5 & \text{if } 150 < t \leq 250 \\ 1 & \text{otherwise} \end{cases}$$

$$a_t = \begin{cases} 5 & \text{with probability } H_t/25 \\ 0 & \text{otherwise} \end{cases}$$

$$x_t = a_t + \epsilon_{x,t}$$

$$y_t = H_t + \epsilon_{y,t}$$

where the measurement noise terms $\epsilon_{x,t}$ and $\epsilon_{y,t}$ are normal random variables with a mean of 0 and

variance of 1. The $\epsilon_{x,t}$ terms at different times are independent, and so are the $\epsilon_{y,t}$ terms. The covariance of the $\epsilon_{x,t}$ and $\epsilon_{y,t}$ terms was set to either 0 (so that $\{x_t\}$ and $\{y_t\}$ are independent) or 0.3 (so that $\{x_t\}$ and $\{y_t\}$ are dependent). Both $\{x_t\}$ and $\{y_t\}$ are nonstationary. Whereas $\{y_t\}$ has a trend component given by $\{H_t\}$, the $\{x_t\}$ process does not have a deterministic trend component (Fig 26A).

We simulated $\{x_t\}$ and $\{y_t\}$ time series and tested whether they are dependent using various flavors of time-shift tests. Specifically, we either (1) tested for dependence between $\{x_t\}$ and $\{y_t\}$ using the (regular) TTS test, (2) tested for dependence between $\{x_t\}$ and the detrended series $\{\epsilon_{y,t}\}$ using the (regular) TTS test, or (3) tested for dependence between $\{x_t\}$ and $\{y_t\}$ using the detrend-retrend TTS test (i.e. detrending and retrending $\{H_t\}$ on $\{y_t\}$). For the different tests, we report the proportion of trials in which dependence was detected when the true covariance between $\epsilon_{x,t}$ and $\epsilon_{y,t}$ was either 0 (Fig A42B, "False positive rate") or 0.3 (Fig A42B, "True positive rate"). Directly applying the TTS test to $\{x_t\}$ and $\{y_t\}$ is invalid (since $\{y_t\}$ is nonstationary) and failed to control the false positive rate at the 0.05 significance level (Fig A42B top row). The other two tests are valid and correctly controlled the false positive rate. Of these, the detrend-retrend TTS test had higher detection power.

For the system where $\{x_t\}$ is stationary and $\{y_t\}$ is nonstationary (Fig A42C), we simulated an equivalent system except where x_t is no longer determined by H_t:

$$a_t = \begin{cases} 5 & \text{with probability } 1/25 \\ 0 & \text{otherwise} \end{cases}$$

$$x_t = a_t + \epsilon_{x,t}$$

All other aspects of the system in Fig A42C are equivalent to those in Fig A42A. In this case, directly applying the TTS test did not result in a high false positive rate, and simply detrending gave grater detection power than the detrend-retrend TTS test (Fig A42D).

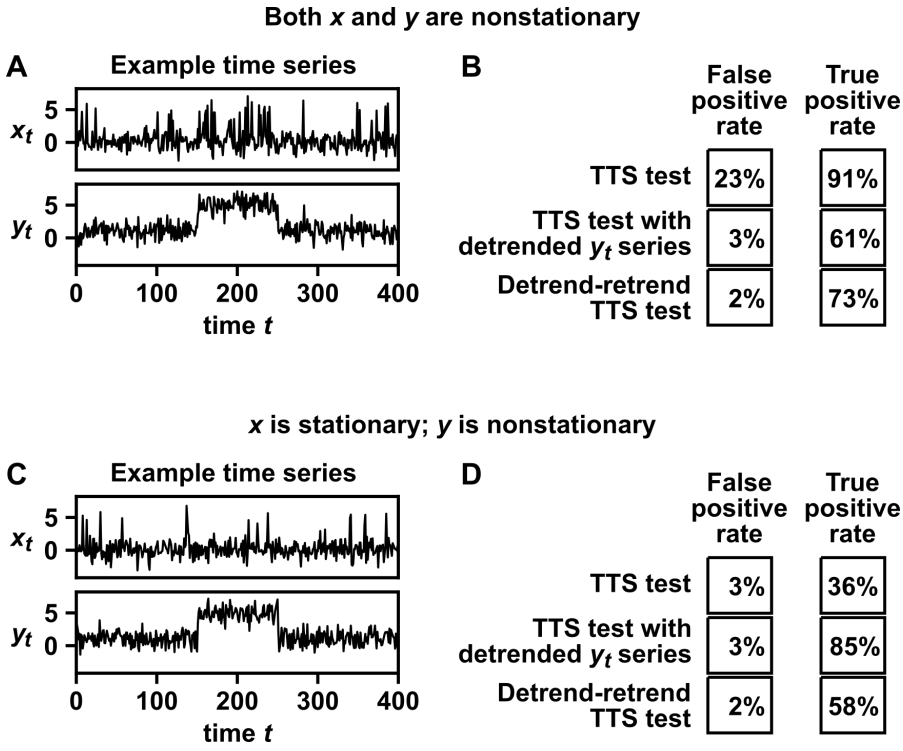


Figure A42: The detrend-retrend TTS test correctly controls the false positive rate in a benchmark simulation. (A) A simulated benchmark system in which both time series are nonstationary and where one time series ($\{y_t\}$) can be decomposed into a deterministic trend component and a stationary component. See text for simulation details. (B) Performance comparison. Different procedures were used to test whether $\{x_t\}$ and $\{y_t\}$ are dependent at the 0.05 significance level. Either the TTS test was applied directly, the TTS test was applied after detrending the $\{y_t\}$ series, or the detrend-retrend TTS test was used. To compute true positive rates, we set the covariance between the measurement noise of x_t terms and measurement noise of y_t to 0.3; for false positive rates this covariance was set to 0. For each test, the absolute value of the Pearson correlation coefficient was used as the correlation function, the truncation radius r was set to 79, and true (or false) positive rates were reported as the proportion of 10^4 trials in which dependence was detected. (C-D) Same as (A-B), but for a system where $\{x_t\}$ is stationary and $\{y_t\}$ is nonstationary.

A5 Additional detection power comparisons between TTS and other tests

We performed several additional simulations examining the detection power of the TTS test, and comparing it to those of other surrogate data tests and the parametric test for Pearson correlation. In all of the following benchmark tests, we used the significance level of 0.05 to detect dependence. In many of the following benchmark tests, we pre-shift data as in Fig 28. To fix notation, "pre-shifting by s" means that we test for dependence between $\{x_{1+s}, x_{2+s}, ..., x_n\}$ and $\{y_1, y_2, ..., y_{n-s}\}$ if $s \geq 0$, and test for dependence between $\{x_1, x_2, ..., x_{n-s}\}$ and $\{y_{1+s}, y_{2+s}, ..., y_n\}$ if $s < 0$.

A5.1 Unidirectionally coupled processes

Autoregressive process: We simulated the unidirectionally coupled autoregressive process:

$$x_t = \rho_x x_{t-1} + \rho_{yx} y_{t-1} + \epsilon_{x,t}$$

$$y_t = \rho_y y_{t-1} + \epsilon_{y,t}$$

where the $\epsilon_{x,t}$ and $\epsilon_{y,t}$ terms are independent random variables with a standard normal distribution. Unless otherwise specified, we used $\rho_x = 0.4$, $\rho_y = 0.4$, $\rho_{yx} = 0.3$, and a time series length of 400. We varied each of these four parameters (ρ_x, ρ_y, ρ_{xy}, and the time series length) individually. The initial conditions of x and y were chosen independently at random from a standard normal distribution and the system was allowed to equilibrate for 500 steps before recording measurements used for testing. We tested for dependence between the two time series using the Pearson correlation strength as the correlation statistic, and compared the detection power of each of the tests used in Fig 27. Additionally, we pre-shifted the $\{x_t\}$ series by $s = 1$ time step. For the TTS test, we used two different values for r (79 and 19). We used $r = 79$ because this is an intermediate value as suggested in the main text and we used $r = 19$ in order to use the TTS test for time series lengths as low as 100. We used $r = 80$ for the naive TTS test as this is the multiple of 10 nearest to 79. For all other tests, details are the same as in Fig 28. 10^4 trials were used to compute the true positive rate for each choice of parameters. The true positive rates are given in the file 2_coupled_ar1.xlsx.

Coupled logistic map: We simulated the unidirectionally coupled logistic map process:

$$x_t = x_{t-1}(r_x - r_x x_{t-1} - r_{yx} y_{t-1})$$

$$y_t = y_{t-1}(r_y - r_y y_{t-1})$$

where the initial conditions of x and y were chosen independently at random from a uniform distribution between 0.2 and 0.8, and the system was allowed to equilibrate for 500 steps before recording measurements used for testing. Unless otherwise specified, we used $r_x = 3.5$, $r_y = 3.8$, $r_{yx} = 0.1$ and a time series length of 400. We varied each of these four parameters (ρ_x, ρ_y, ρ_{xy}, and the time series length) individually. We tested for dependence between the two time series using cross-map skill with the same direction and parameters as in Fig 28 (except for when using the parametric test, which relies on Pearson correlation). Additionally, we pre-shifted the $\{x_t\}$ series by $s = 1$ time step. The values of r for the TTS and naive TTS tests are the

same as in the section immediately above. For all other tests, details are the same as in Fig 28. 2000 trials were used to compute the true positive rate for each choice of parameters. The true positive rates are given in the file 3_coupled_logistic.xlsx.

Nonlinearly coupled autoregressive process: We simulated a pair of autoregressive processes that are coupled in a nonlinear way:

$$x_t = \rho x_{t-1} + \cos(\epsilon_t)$$
$$y_t = \rho y_{t-1} + \sin(\epsilon_t)$$

where the ϵ_t terms are independent random variables drawn from a uniform distribution between 0 and 2π. The state variables x and y were initialized at 0, and then the system was allowed to equilibrate for 500 steps before recording measurements used for testing. For the default system we used $\rho = 0.35$ and a time series length of 400. We either varied ρ while keeping the length at 400 or varied the length while keeping ρ at 0.35. We tested for dependence using the mutual information estimator (see Methods) and compared the detection power of each of the tests used in Fig 27. We did not pre-shift time series. The values of r for the TTS and naive TTS tests are the same as in the sections immediately above. For all other tests, details are the same as in Fig 28. 2000 trials were used to compute the true positive rate for each choice of parameters. The true positive rates are given in the file 4_nonlin_coupled_ar1.xlsx.

A5.2 Bidirectionally coupled processes with random coupling delays

Autoregressive process: We simulated the bidirectionally coupled autoregressive process where x and y both influence each other with random strengths and random coupling delays, with iid Gaussian process noise of mean zero and variance 1:

$$x_t = 0.4x_{t-1} + \rho_{yx} y_{t-\tau_{yx}} + \epsilon_{xt}$$
$$y_t = 0.4y_{t-1} + \rho_{xy} x_{t-\tau_{xy}} + \epsilon_{yt}$$

The interaction strengths are chosen randomly from a continuous uniform distribution, and they are

made sum to 0.4:

$$\rho_{yx} \sim \text{Unif}(0, 0.4)$$
$$\rho_{xy} = 0.4 - \rho_{yx}$$

Note that for any particular realization of the process, we have $\rho_{xy} \neq \rho_{yx}$ to mimic real world where variables are rarely symmetric.

The interaction delays τ_{yx} and τ_{xy} are drawn independently from the set $\{1, 2, ..., \tau_{max}\}$ with equal chance. Thus, τ_{max} determines the amount of uncertainty in the coupling delays. Here, we varied τ_{max} from 1 to 5. Additionally, the initial conditions of x and y were chosen independently at random from a standard normal distribution and the system was allowed to equilibrate for 500 steps before recording measurements used for testing.

We tested for dependence using the absolute value of the Pearson correlation between $\{x_t\}$ and $\{y_t\}$. To account for the delay uncertainty, we pre-shifted data by several choices of s and performed a multiple-testing correction. We tried different numbers of pre-shifts centered at $s = 0$. That is, we preshifted by:

$$-s_{max}, ..., s_{max}$$

and performed a Bonferroni correction for the number of pre-shifts ($= 2s_{max} + 1$). We varied s_{max} from 1 to 5. Larger values of s_{max} correspond to an assumption of greater uncertainty in the coupling delay. As in Fig 28, we used the lowest value of r that would enable significance at the 0.05 level *after* the Bonferroni correction (i.e. $r = 59$ for 3 pre-shifts, $r = 99$ for 5 pre-shifts, and so on). In order to handle up to 11 different choices of s (and thus 11 different hypothesis tests to correct for), we used a time series length of 800. We used the parametric test for Pearson correlation and IAAFT surrogate test for comparison. We used 1499 IAAFT surrogates rather than 99, to account for the lower p-values demanded by the Bonferroni correction. To calculate true positive rates, we repeated the trial 10^4 times for TTS and parametric tests, and 500 times for the slower IAAFT test. The true positive rates are given in the file 5_bidirectional_ar.xlsx.

Logistic map: We simulated the bidirectionally coupled logistic map process where x and y influence each other with random coupling delays:

$$x_t = x_{t-1}(r_x - r_x x_{t-1} - 0.1 y_{t-\tau_{yx}})$$

$$y_t = y_{t-1}(r_y - r_y y_{t-1} - 0.1 x_{t-\tau_{xy}})$$

where the initial conditions of x and y were chosen independently at random from a uniform distribution between 0.2 and 0.8, and the system was allowed to equilibrate for 500 steps before recording measurements used for testing. The terms r_x and r_y are randomly drawn from the set:

$$\{3.7, 3.72, 3.74, 3.76, 3.78, 3.8\}$$

without replacement (i.e. $r_x \neq r_y$). We require $r_x \neq r_y$ to avoid pathological synchrony [18, 127]. As in the above section, the interaction delays τ_{yx} and τ_{xy} are randomly drawn independently from the set $\{1, 2, ..., \tau_{max}\}$, and we varied τ_{max} from 1 to 5. Again, we used a time series length of 800 for this benchmark.

We tested for dependence between the two time series using cross-map skill with the same direction and parameters as in Fig 28 (except for when using the parametric test, which relies on Pearson correlation). As in the section above, we used the TTS test with multiple pre-shift values and a Bonferroni multiple test correction, and the IAAFT and parametric correlation tests. All other details of the analysis, such as choices of r and trial numbers, are the same as in the above section. The true positive rates are given in the file 6_bidirectional_logistic.xlsx.

A6 Detailed methods and results for the orbital-climate dependence example

A6.1 Obtaining orbital parameter time series and deglaciation event times.

We obtained obliquity, precession, and eccentricity time series from [155] using the web interface at `http://vo.imcce.fr/insola/earth/online/earth/online/index.php`, with a sampling frequency of 1 kyr. The solar constant was left as the default value of 1368 watts per square meter. Deglaciation event times [160] were obtained from `https://www.ncei.noaa.gov/pub/data/paleo/contributions_by_author/huybers2006/huybers2006.txt`.

A6.2 Testing for dependence using a nonlinear correlation statistic

For Fig 29C we used a correlation function based on state space reconstruction (SSR) with delay vectors. One of the most popular state space reconstruction correlation statistics is the so-called "cross-map skill" of the convergent cross-mapping approach [18]. Our technique, while inspired by cross-map skill, deviates from it in two ways: the weighting function and the number of neighbors used for prediction. We first describe this correlation function and use a simulation to justify why these changes are important for the orbital-climate dependence example.

To obtain a correlation between two time series $\{x_t\}$ and $\{z_t\}$, we begin by constructing delay vectors from the $\{z_t\}$ series:

$$v_t = (z_t, z_{t-D}, ..., z_{t-(E-1)D})$$

where D is the delay amount and E is the embedding dimension (we will discuss the choice of the parameters D and E later). v_t is thus a point in an E-dimensional "delay space". We now have the paired time series $\{x_t, v_t\}_{t=1}^{n}$. Next, we try to predict the $\{x_t\}$ series using the delay vectors $\{v_t\}$. Specifically for each value of $i = 1, 2, ..., n$, we find the k nearest neighbors of v_i in the delay space. Note that v_i cannot be its own neighbor. We use these nearest neighbors to predict x_i:

$$\widehat{x}_i = \frac{\sum_{j=1}^{n} x_j w(v_i, v_j) I_k(i,j)}{\sum_{j=1}^{n} w(v_i, v_j) I_k(i,j)}$$

$$I_k(i,j) = \begin{cases} 1 & \text{if } v_j \text{ is one of } v_i\text{'s } k \text{ nearest neighbors} \\ 0 & \text{otherwise} \end{cases}$$

where $\widehat{x}_i$ is the predicted value of x_i and $w(v_i, v_j)$ is a weighting function which is larger when v_i and v_j are closer together (to be discussed further below). Our correlation statistic is then given by the negative mean squared error of the predictions:

$$-\frac{1}{n}\sum(x_i - \hat{x}_i)^2. \tag{12}$$

We choose negative error (instead of error) for the statistic so that higher values of the statistic (lower error) will correspond to stronger coupling signal, as per the TTS test. The popular cross-map skill procedure [18] typically uses an exponential weighting function:

$$w_{exp}(v_i, v_j) = e^{-|v_i - v_j|/d_i} \tag{13}$$

where $|v_i - v_j|$ is the Euclidean distance from v_i to v_j and where d_i is the Euclidean distance from v_i to its closest neighbor. However, we found that these choices do not perform well for time series that resemble the deglaciation series. Specifically, our $\{x_t\}$ series (i.e. the series being predicted) is a series of 2000 kiloyears in which 36 deglaciations occurred. That is, 36 values of the $\{x_t\}$ series are 1s and the rest are 0s. We have found that in simulations where $\{x_t\}$ is a sparse event time series, the exponential weight function has abysmal detection power. Instead, here we use a simple inverse-distance weighting function:

$$w_{inv}(v_i, v_j) = 1/|v_i - v_j| \tag{14}$$

which we found to have high detection power for sparse event time series.

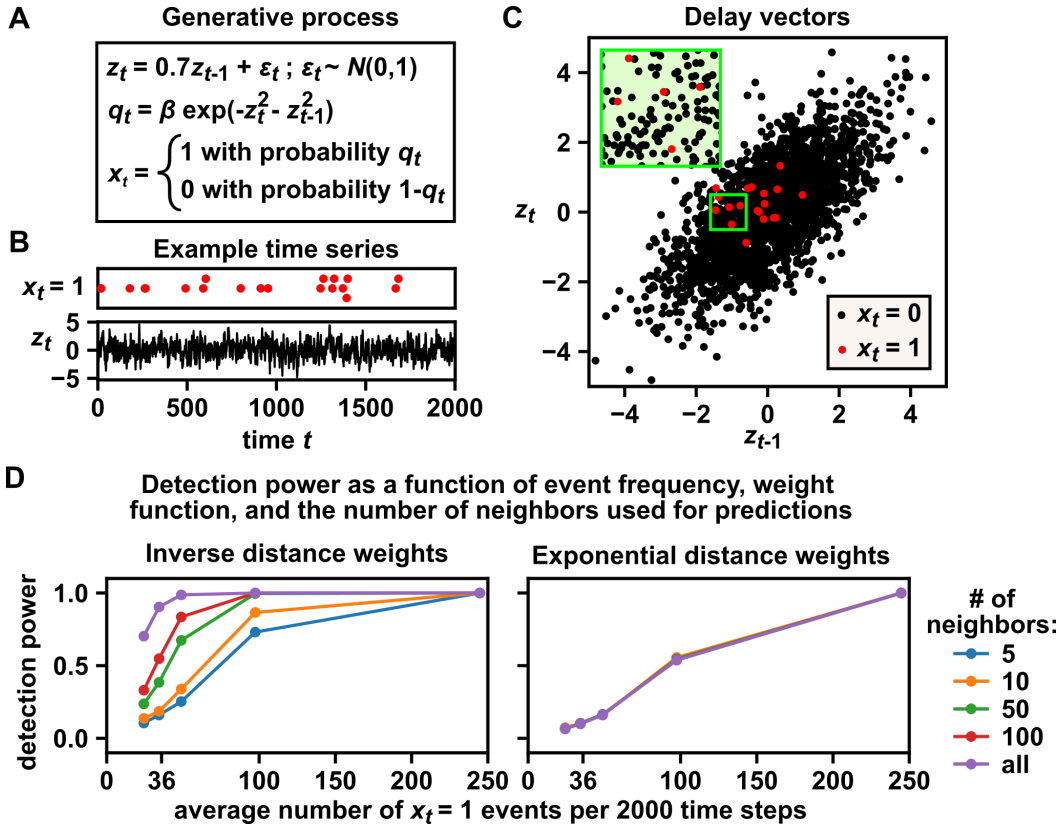


Figure A43: Inverse distance weights and larger numbers of neighbors improve detection power for a simulated sparse event time series. (A) $\{z_t\}$ was generated according to a simple first-order autoregressive process. $\{x_t\}$ is an event time series (x_t is either 1 or 0) and depends on $\{z_t\}$, with 1 occurring more frequently when z_t and z_{t-1} are near the origin. The parameter β determines the frequency of events. (B, C) Time series and delay vector plot of a realization with $\beta = 0.05$. In the event time series, overlapping dots are vertically separated so that they can be distinguished (i.e. the vertical axis is meaningless). Delay vectors were constructed as: $v_t = (z_t, z_{t-1})$. When events are sparse, event points are generally surrounded by non-event points. Note that events are clustered around the origin. Thus, nearest-neighbor prediction techniques with a small number of nearest neighbors may suffer poor prediction accuracy. (D) The TTS test with various SSR-based correlation functions (Eq. 12) was used to detect the dependence between $\{x_t\}$ and $\{v_t\}$. Time series were of length 2198 and r was set to 99, so that truncated $\{x_t\}$ and $\{v_t\}$ series would be of length 2000, mirroring the length of the deglaciation series. The event frequency parameter β was set to 0.05, 0.07, 0.1, 0.2, and 0.5, which correspond to an average of 24, 34, 49, 97, and 245 events within the length-2000 truncated window. (The deglaciation time series has 36 events among its 2000 time points). The detection power (i.e. the proportion of 1000 trials in which the dependence between $\{x_t\}$ and $\{z_t\}$ was detected at the 0.05 significance level) is shown as a function of the average number of events within the truncated $\{x\}$ window. Either inverse distance weights (Eq. 14) or exponential distance weights (Eq. 13) were used and the number of neighbors used for prediction (i.e. k) was varied from 5 to all possible neighbors (1999 in this case). For sparse event series, inverse distance weighting with all possible neighbors provides substantially higher power than all other options. Curves for exponential distance weights are superimposed because the number of neighbors did not affect detection power.

In order to compute correlations using this approach, we must pick embedding parameters (the delay lag D and the embedding dimension E) for each orbital parameter time series. As is typical, we selected embedding parameters that maximized predictions of future values using past values of the same time series.

There are three further considerations in our embedding parameter selection procedure: (1) Which portion of the data did we use as training data to select embedding parameters? (2) If we chose embedding parameters to minimize the error in forecasts q kyr into the future, what was q? (3) Which possible parameter values did we search over and what other hyperparameters were needed? Below we discuss each of these in turn.

Which data did we use for parameter value selection? We randomly selected 25 windows of 2 Myr (the same as the length of the deglaciation time series) from between -12 Myr and $+10$ Myr (where 0 is the present time). For each 2-Myr window (2000 data points), we constructed delay vectors from the orbital parameter time series and used the SSR correlation function to predict future values of the same time series. We thus obtained 25 mean squared error values (one for each window). Embedding parameters were chosen to minimize the average over all 25 mean squared error values.

Embedding parameters were chosen by minimizing the error of future predictions; how far into the future (forecast horizon) did we predict? We chose the forecast horizon to be a fixed proportion of the cycle period instead of a fixed amount of time in order to subject time series with different autocorrelation structures to prediction problems of similar "difficulty". Cycle periods for obliquity, precession and eccentricity, estimated as average interpeak distances, were 41 kyr, 21 kyr, and 99 kyr respectively (broadly consistent with known values [156]). Thus, for embedding parameter selection, we used forecasts of 8, 4, and 20 kyr ahead for obliquity, precession, and eccentricity respectively.

Which possible embedding parameter values did we search over? We scanned embedding dimensions from 2 to 6 and scanned delay lags from 1 to 55 kyr, and chose embedding parameters that minimized average prediction error. Also, since these predictions were on continuous time series (not event data), we used the standard [18] exponential distance weights (Eq. 13) and used $k = 5$ nearest neighbors for predictions.

From this procedure with all of the above ingredients, we obtained delay vector parameters for obliquity ($E = 6,\ D = 31$), precession ($E = 5,\ D = 4$), and eccentricity ($E = 6,\ D = 45$).

We then computed the nonlinear correlation statistic (Eq. 12 and 14) with all neighbors (i.e. all delay vectors except self, or equivalently, $k = 1999$) by predicting deglaciation events based on the delay vectors of orbital parameters. Finally, we used the TTS test to assess the significance of the statistic. For the x^{trunc} window we used the 2000 kyr of deglaciation event data (-1999 to 0 kyr) from [160], with event times rounded to the nearest kyr. We used a flanking radius of $10,000$ kyr.

We did not pre-shift time series because orbital parameter time series could predict their own future values with high accuracy. More specifically, using the same exponential distance weighting prediction method as before (Eq. 13 with $k = 5$) with the chosen delay vector parameters, all three orbital time series could predict their own values 10 kyr into the future with a coefficient of determination R^2 of more than 0.96 over the time stretch of -1999 to 0 kyr. Since orbital time series could predict their own futures with such high accuracy,

we expect that there exists a reasonably high-quality mapping (predictability) between each orbital series and itself shifted by $\lesssim 10$ kyr. Moreover, if there exists a mapping from unshifted obliquity series to shifted obliquity series and there exists a mapping from shifted obliquity series to deglaciation, then we can expect there to be some mapping from unshifted obliquity series to deglaciation. In other words, we expected that small-to-moderate pre-shifts before the TTS test would be unlikely to severely change the correlations with the deglaciation series.

A7 Detailed methods and results for the cross-site microbiome dependence example

A7.1 Testing for dependence between population at the single-species level is impeded by compositional data.

Many past microbiome surveys have relied on 16S ribosomal RNA sequencing approaches in which relative abundance levels of taxa are observed ("compositional data"), but not absolute abundance levels [164]. Because of this, detecting species-level dependence relationships directly from data is difficult (Fig A44). Specifically, it is difficult to make valid species-level inferences about dependence between subpopulations using relative abundance data, both within a body site (Fig A44B-iii) and between body sites (Fig A44B-iv). For this reason, we focus on testing for dependence between body sites as a whole, a task that is not impeded by compositional data, as we prove in the following subsection. For simplicity, we used purely visual arguments in this illustration rather than discussing the details of hypothesis testing.

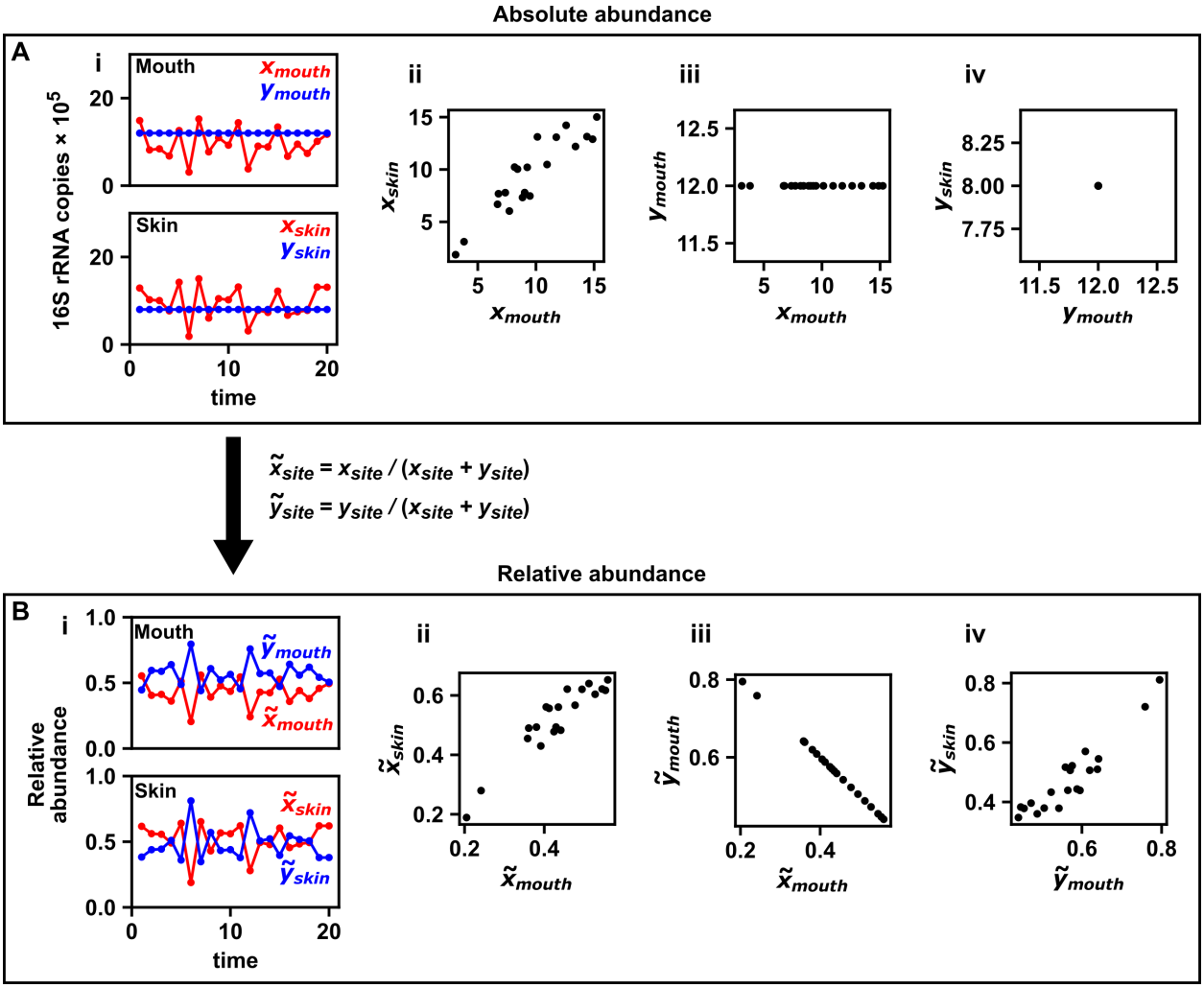


Figure A44: Naive species-level analysis of relative abundance data can result in spurious correlations. (A) Absolute abundance estimates of 16S rRNA copy numbers (a proxy for biomass) of two microbial species (x and y) measured over time at two body sites (mouth and skin). (A-i) shows that x_{mouth} and x_{skin} covary, whereas y_{mouth} and y_{skin} stay constant over time. Scatter plots show that x is correlated between the two body sites (A-ii), whereas there is no evidence of a correlation between the x and y populations in a single site (e.g. A-iii), and there is also no evidence of a correlation between the y populations at the two sites (A-iv). (B) Same as A, but now using relative abundance measurements ($\tilde{x}$ and $\tilde{y}$), calculated as $\tilde{x}_{site} = x_{site}/(x_{site} + y_{site})$. The original correlation still exists (B-ii), but now spurious correlations have appeared, both within (B-iii) and between (B-iv) body sites.

A7.2 Testing for dependence between whole microbial communities at different body sites is not impeded by compositional data

Here we formally show that testing for dependence between the microbial communities of two body sites is not impeded by compositional data. The argument relies primarily upon theorem 11 from Appendix A1.2, which is repeated here for the reader's convenience.

Theorem: Let $x \in \mathbb{R}^{m\times n}$ and $y \in \mathbb{R}^{m\times n}$ be random variables (or vectors of random variables, or matrices of random variables) such that x and y are independent. Let $f : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ and $g : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ be functions that are measurable on $(\mathbb{R}^{m\times n}, \mathcal{B}_{m\times n})$. Then, $f(x)$ and $g(y)$ are independent.

With this key fact under our belt, let the absolute population densities of m taxa at a certain body site be the matrix X where $X_{i,t}$ is the density of the ith taxon on day t. Then, the compositional (or relative abundance) matrix $h(X)$ is given by:

$$(h(X))_{i,t} = \frac{X_{i,t}}{\sum_{j=1}^{m} X_{j,t}}.$$

As long as a nonzero total microbial load is measured on each day, h is continuous and therefore measurable. Thus we can apply the above theorem. In particular, if X and Y are independent random variables, then their compositional counterparts $h(X)$ and $h(Y)$ must also be independent. In fact, the contrapositive of this statement is more useful for statistical testing: If we can show by statistical testing that $h(X)$ and $h(Y)$ are dependent, then it follows that X and Y are also dependent.

A7.3 Obtaining OTU tables

The OTU count table and sample information were obtained from the Qiita platform [165]. The OTU table we used is available in BIOM format at https://qiita.ucsd.edu/download/219436. A file with sample information (such as collection dates, body sites, and the human subject a sample was taken from) is available at https://qiita.ucsd.edu/download/773810. Microbiome survey time series were reported from two human subjects (M3 and F4) [163]. We used data from subject M3 as the time series are longer.

The OTU table was generated by a standard workflow that trimmed sequences to 90 base pairs and then generated an OTU table by a closed reference picking method. This workflow can be viewed in a visual flowchart at https://qiita.ucsd.edu/study/description/550. The OTU table ID is 44973.

A7.4 Local similarity analysis

Local similarity analysis has been implemented in various different ways [37, 166]. We used the original procedure of [37] with the maximum delay parameter set to zero. Specifically, to compute the local similarity score of two time series $s(\{x\}_1^n, \{y\}_1^n)$, we begin by taking the rank of each time series. Let $r_{x,t}$ and $r_{y,t}$ be the rank of x_t and y_t respectively. We dealt with tied points by assigning them the average of the ranks they spanned. For example, if the x series was

$$x_1 = 10, x_2 = 10, x_3 = 5, x_4 = 20$$

then the ranks would be

$$r_{x,1} = 2.5, r_{x,2} = 2.5, r_{x,3} = 1, r_{x,4} = 4.$$

"Normalized" x and y time series are then computed as:

$$\tilde{x}_t = \Phi^{-1}\left(\frac{r_{x,t}}{n+1}\right); \tilde{y}_t = \Phi^{-1}\left(\frac{r_{y,t}}{n+1}\right)$$

for $t = 1, ..., n$, where Φ is the cumulative distribution function of the standard normal distribution. Next, a recursive procedure is used to look for positive (S^+) or negative (S^-) local correlations, which can vary over time:

$$S_1^+ = 0; S_1^- = 0$$

$$S_{t+1}^+ = \max(0, S_t^+ + \tilde{x}_t\tilde{y}_t)$$
$$S_{t+1}^- = \max(0, S_t^- - \tilde{x}_t\tilde{y}_t).$$

We then find the maximum of these local (anti)correlations:

$$S_{\max}^+ = \max(S_1^+, S_2^+, ..., S_{n+1}^+)/n$$
$$S_{\max}^- = \max(S_1^-, S_2^-, ..., S_{n+1}^-)/n.$$

Finally, the local similarity score s is given by the maximum correlation or maximum anticorrelation, whichever stronger:

$$s = \begin{cases} S_{max}^+ & \text{if } \mathrm{S}_{\max}^+ > \mathrm{S}_{\max}^- \\ -S_{\max}^- & \text{if } \mathrm{S}_{\max}^+ < \mathrm{S}_{\max}^- \\ 0 & \text{if } S_{\max}^+ = S_{\max}^- \end{cases}$$

A7.5 Detailed results

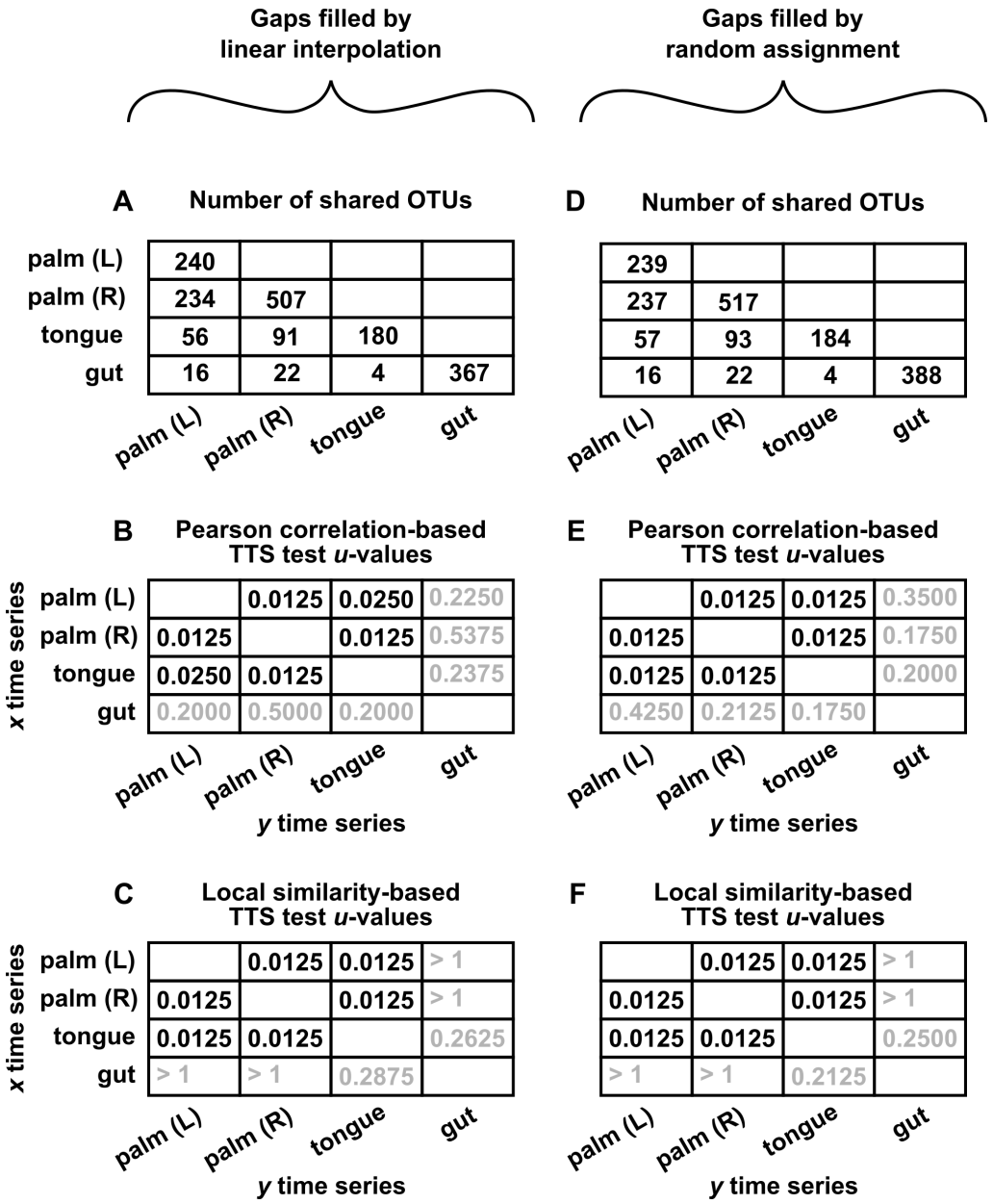


Figure A45: Detailed results from the cross-site microbiome dependence example of Fig 30. We only used measurements from day 42 to day 418 to avoid the longest gaps (> 6 days). Each of the resulting time series contained no more than 41 short (1-day) gaps, no more than 11 medium (2-, 3-, or 4-day) gaps, and no more than 2 longer (5- or 6-day) gaps. Gaps in time series were filled by either linear interpolation (A-C) or random assignment (D-F; each missing taxon abundance value was filled with an abundance value of the same taxon at the same body site from a random time). All random assignments were performed independently (i.e. "with replacement"). (A,D): The number of shared taxa between each body site after removing taxa that were either rare (i.e. absent in half of measurements) or nonstationary (according to an augmented Dickey-Fuller test at the 0.05 significance level). Values along the diagonal denote the total number of taxa at a site after preprocessing. (B,E): TTS test u-values from the test based on Pearson correlation. (C,F): TTS test u-values from the test based on local similarity scores. All u-values shown are raw (i.e. before FDR adjustment). The y time series denotes the series that was used to generate surrogates. Note that tables in B, C, E, and F do not need to be symmetric.

A8 Detailed methods and results for the zebrafish behavior example

A8.1 Obtaining speed and direction time series

We obtained fish trajectory data from the web address at `https://drive.google.com/drive/folders/1UmzlX-yJhzQ5KX5rGry8wZgXvcz6HefD`. We used the file at the path "100/1/trajectories_wo_gaps.npy" (where 100 is for the number of fish per tank, and 1 indicates the first video). This file contains sequences of fish positions indexed by video frame, as well as constants such as frame rate and approximate fish body length. We preprocessed these data using custom scripts inspired by the authors' tutorial at `https://gitlab.com/polavieja_lab/idtrackerai_notebooks/-/blob/master/trajectories_analysis/T1_trajectories_analysis.ipynb`. Where applicable, preprocessing scripts were validated against the authors' trajectorytools package [174].

First, we estimated the position of the center of the tank using the miniball algorithm [201] (as implemented in `https://github.com/weddige/miniball`), and used this to center the data. We then computed fish velocity as

$$\overrightarrow{v}_t = \frac{\overrightarrow{x}_{t+1} - \overrightarrow{x}_{t-1}}{2} R$$

where $\overrightarrow{x}_t$ is a 2-dimensional column vector specifying fish position in units of body length at video frame t, $\overrightarrow{v}_t$ is a 2-dimensional column vector specifying the fish velocity in units of body length per second at video frame t, and R is the frame rate (32 frames per second). Fish speed was simply defined as $|\overrightarrow{v}_t|$. We then computed direction as the angle φ_t between the position and velocity vectors, as described pictorially in Fig 31. Mathematically, φ_t is the angle that satisfies:

$$\cos(\varphi_t) = \left(\frac{\overrightarrow{x}_t}{|\overrightarrow{x}_t|}\right)^\top \frac{\overrightarrow{v}_t}{|\overrightarrow{v}_t|}$$

$$\sin(\varphi_t) = \left(\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \frac{\overrightarrow{x}_t}{|\overrightarrow{x}_t|}\right)^\top \frac{\overrightarrow{v}_t}{|\overrightarrow{v}_t|}.$$

Note that the matrix in the equation for $\sin(\varphi_t)$ is the rotation matrix [202] for a 90° rotation in 2 dimensions. A few fish trajectories had a small number of frames with zero speed. If speed is zero, then direction of movement is, strictly speaking, undefined. To deal with this, we filled in the direction values at zero-speed frames by linear interpolation between the nearest frames with nonzero speeds. Most of the fish trajectories had data gaps. These gaps were coded as "nan" values. Since there were still 44 fish trajectories

without any gaps, we only performed the TTS test on fish trajectories without gaps.

A8.2 Selecting approximately stationary time series

Groups of zebrafish are known to reduce their swim speed in the days after being introduced into a tank [176]. Indeed, visual inspection of the video from [175] appears to reveal a trend of slowing average fish speed throughout the 10 minutes of tracking in the video, suggesting that fish behavior may be nonstationary (Fig A46A). In fact, this trend is replicable (Fig A46B) across the three replicate videos from [175]. Although the TTS was robust to nonstationarity arising from a simple linear trend in Fig 27, we still wished to more closely align the data with the theoretical assumption of stationarity. We therefore restricted our analysis to the first $10,000$ frames (≈ 5 minutes) of video 1 as there is relatively little systematic trend in the average fish speed in this segment (dotted box in Fig A46A).

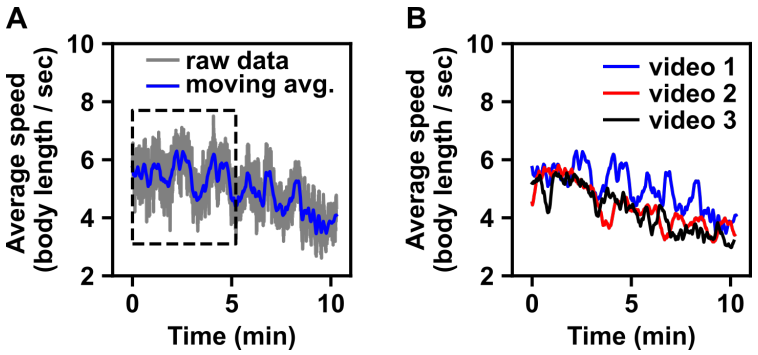


Figure A46: Average fish speeds. (A) Average fish speed in the first video from [175]. The average speed of the 100 fish in the tank is shown in grey. An 11-second moving average of the grey trajectory is shown in blue. The black dotted box shows the first $10,000$ frames, which were selected for analysis. (B) 11-second moving averages of all 3 videos from [175].

A8.3 Detailed results

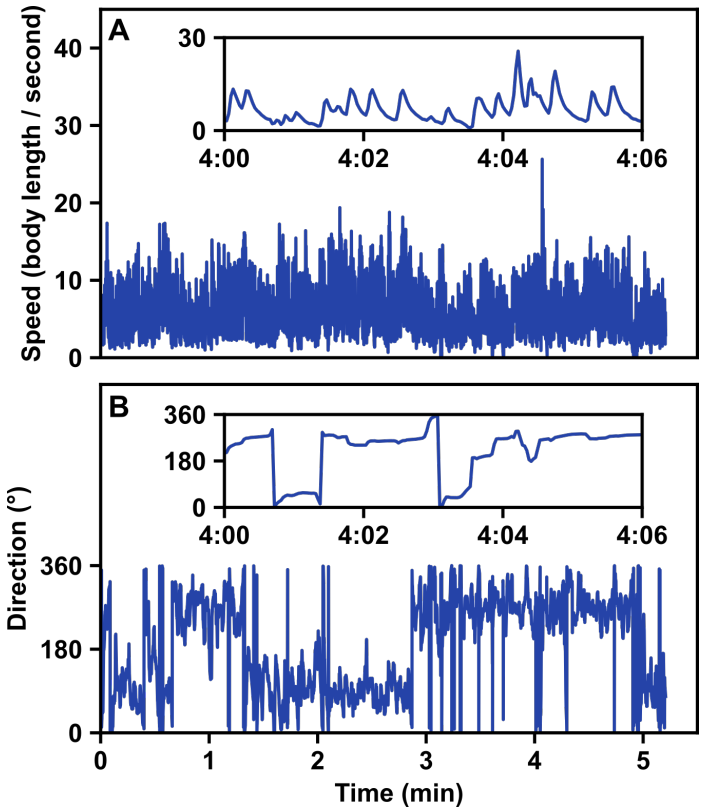


Figure A47: Time series of speed and direction from the fish ("individual 74") analyzed in Fig 31C.

A9 Difficulty of testing for stationarity

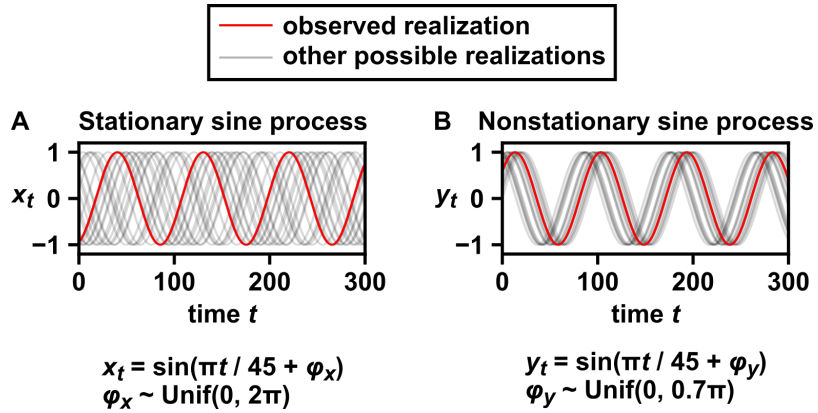


Figure A48: Stationarity is a property of an ensemble of time series, not any single time series. (A) A stationary process that produces a sine wave. Twenty possible realizations of the process are shown. An arbitrary realization is colored red and called the "observed realization". (B) A nonstationary process that produces a sine wave. To see that this process is nonstationary, note that the ensemble mean of processes changes with time. Although only one of the two processes is stationary, a single realization of one process is functionally indistinguishable from a single realization of the other.

Part III

Testing nonparametrically for dependence between nonstationary time series with very few replicates

A version of this part is available on the bioRxiv preprint server with the title "Testing nonparametrically for dependence between nonstationary time series with very few replicates."

Abstract

Many processes of scientific interest are nonstationary, meaning that they experience systematic changes over time. These processes pose a myriad of challenges to data analysis. One such challenge is the problem of testing for statistical dependence between two nonstationary time series. Existing tests mostly require strong modeling assumptions and/or are largely heuristic. If multiple independent and statistically identical replicates are available, a trial-swapping permutation test can be used. That is, within-replicate correlations (between time series of X and Y from the same replicate) can be compared to between-replicate correlations (between X from one replicate and Y from another). Although this method is simple and largely assumption-free, it is severely limited by the number of replicates. In particular, the lowest attainable p-value is $1/n!$ where n is the number of replicates. We describe a modified permutation test that partially alleviates this issue. Our test reports a lower p-value of $1/n^n$ when there is particularly strong evidence of dependence, and otherwise defaults to a regular trial-swapping permutation test. We use this method to confirm the observation that groups of zebrafish swim faster when they are aligned, using an existing dataset with only 3 biological replicates.

1 Introduction

Scientists frequently look for correlations between variables to identify potentially important relationships, or to support conceptual or quantitative models. In disciplines with a focus on dynamics (such as ecology, physiology, and psychology), it is common to measure correlations between temporal processes (i.e. time series correlations). Many important biological processes are nonstationary, meaning that their statistical

properties (e.g. mean, variance, etc.) change systematically across time. For instance, the expansion of an invasive species and a cell's response to a new environmental stress are both likely to be nonstationary processes.

Interpreting a correlation between a pair of nonstationary time series can be highly fraught because it is easy to obtain a seemingly impressive correlation between two time series that have no meaningful relationship [122]. For example, the sizes of any two exponentially-growing populations will be correlated over time due to a shared growth law, even if the populations lack any interaction or shared influences. To avoid spurious correlations, it helps to distinguish between the concepts of "correlation" and "dependence", and how each relates to causality. In time series research, the term "correlation" is often used procedurally [121, 36, 33]. Similarly, here we define a correlation function (ρ) to be any function that takes two time series and produces a statistic, although it is usually interpreted as a measure of similarity or relatedness. For example, two common correlation statistics are Pearson's correlation coefficient and mutual information.

In contrast to correlation, "dependence" is a hypothesis about the relationship between variables, and it has immediate causal implications. Two events A and B are called dependent if the probability that they both occur $P(A, B)$ differs from the product of their individual probabilities $P(A)P(B)$. Similarly, two temporal processes are dependent if the probability of observing any particular pair of trajectories differs from the product of the probabilities of individual trajectories. (The formal definition of dependence extends this idea to continuous variables, in which case discrete event probabilities may, for example, be replaced by probability densities [195, 203].) Importantly, dependence is linked to causality (as defined in the usual sense: X causes Y if perturbations in X can directly or indirectly alter Y). The link between dependence and causality is due to Reichenbach's common cause principle, which states that if two variables are dependent, then they are causally related in the following sense: Either they share a common cause, or one variable causes the other (possibly indirectly) [29, 38]. Thus, it is useful to to first test whether an observed correlation indicates dependence before pursuing specific causal explanations.

There are a handful of systematic approaches to testing for dependence between nonstationary time series. However, most are fairly limited in their scope. First, when possible, a nonstationary time series can be transformed to become stationary (i.e. statistical properties do not change systematically across time). This enables access to a wide arsenal of tests applicable to stationary data. Transformations include subtracting a trend, taking the derivative (more precisely, "differencing" between neighboring points), or choosing a stationary-seeming window of time [204, 205]. However, it is easy to see potential pathologies in each of these three transformations: Taking the derivative of an exponential curve just produces another exponential curve; subtracting a fitted linear trend from the random walk ($X(t) = X(t-1) + \epsilon(t)$ where $\epsilon(t)$ is random noise) does not make it stationary [204]; similarly, searching for a stationary window of the

same random walk process is futile since the variance of X_t increases at each step. As an alternative to data transformation, one may compare the observed correlation between two time series with correlations obtained when one of the time series is replaced by a synthetic replicate, where synthetic replicates are generated in a way that models the form of nonstationarity in the process [206]. However, these require a correct model of how the statistical properties of the process evolve over time. Lastly, there are tests within the econometrics literature that can provide evidence for dependence between random-walk-like nonstationary processes by detecting a property called cointegration, but cointegration only occurs when some linear combination of the time series is stationary [207].

Another approach is possible when there are multiple identically distributed and independent (iid) replicates (or equivalently, iid trials; we use "replicates" and "trials" interchangeably). In this case, one may evaluate the significance of a within-replicate correlation by comparing it to between-replicate correlations. That is, if X_i and Y_i are time series of variables X and Y from replicate i, the correlation of (X_i, Y_i) may be compared to the correlation of $(X_i, Y_{j\neq i})$. If the two variables are dependent, within-replicate correlations should tend to be stronger than between-replicate correlations. This is approach is sometimes called "inter-subject surrogates" [208, 43].

The inter-subject surrogate approach can be used to test for correlations in each trial separately [39]. In this case, a simple nonparametric test of the correlation between, for instance, X_1 and Y_1 can be performed by computing $\rho(X_1, Y_j)$ for $j = 1, ..., n$ and writing down a p-value as the proportion of these correlations that are greater than or equal to $\rho(X_1, Y_1)$ [47, 127]. For this approach, a minimum of $n = 20$ trials is needed if one wishes to possibly obtain a p-value of 0.05 or below.

However, the number of replicates is sometimes constrained by considerations of logistics, ethics, or cost [209], and single-digit replicate numbers are common in biomedical research [210]. For instance, two influential longitudinal human gut microbiome sampling studies relied on cohorts of two subjects each [163, 211]. Additionally, secondary analysis of public data from earlier studies can be a resource-efficient mode of research [212], and this approach is necessarily limited to the number of replicates within the existing data set.

A straightforward strategy in testing for dependence with smaller replicate numbers is to perform a single permutation test (Fig 49A) using the data from all replicates [213, 214]. As an example with $n = 4$ trials, the test procedure begins by computing the mean within-trial correlation (ρ denotes correlation):

$$\frac{1}{4}\left(\rho(X_1, Y_1) + \rho(X_2, Y_2) + \rho(X_3, Y_3) + \rho(X_4, Y_4)\right).$$

Next, as a null model, recompute the mean correlation after randomly permuting the Y time series (while

holding the X time series in the original order). For instance, one such permutation might be:

$$\frac{1}{4}\left(\rho(X_1, Y_3) + \rho(X_2, Y_2) + \rho(X_3, Y_4) + \rho(X_4, Y_1)\right).$$

A p-value can be calculated as the proportion of permutations (including the original ordering) that produce a mean correlation that is as strong as, or stronger than, the original ordering (Fig 49A). One may then detect a correlation at the significance level α if $p \leq \alpha$. We emphasize that these permutations are obtained by swapping trials, not by swapping time points. This test has been used to detect correlations between time series in neuroscience and psychology settings [215, 213, 216]. It has also been used to detect correlations between variables measured in brain images, which can have similar nonstationarity challenges as time series [126]. A noteworthy advantage of this approach is that it is valid (meaning that the false positive rate is guaranteed to not exceed α) under very mild assumptions. Namely, the test is valid if the X_i trials are exchangeable with one another (i.e. all permutations of the sequence $X_1, ..., X_n$ have the same joint probability distribution), or if the Y_i trials are exchangeable (see Corollary 18 in Appendix A1).

A substantial downside to this strategy is that any evidence against independence is mathematically limited by the number of trials, even when the average within-trial correlation is much greater than the average between-trial correlation. Specifically, the lowest p-value that can be obtained with n replicates is $1/n!$ since there are $n!$ possible permutations. For example, with $n = 3$, the lowest possible p-value is $1/3! \sim 0.17$ and with $n = 4$, the lowest is $1/4! \sim 0.04$. Thus, the level of evidence against the null hypothesis will always be modest at best for very small numbers of replicates.

In this article, we show that by adding one additional step to the permutation test, we may achieve p-values as low as $1/n^n$. For $n = 3$ or 4 respectively, the lowest possible p-value is then ~ 0.04 or ~ 0.004. This new result only requires that the replicates be iid. Thus, this modified permutation test allows the data analyst to detect dependence with stronger confidence when there is sufficient evidence to do so. As a demonstration of the method, we verify the observation that milling zebrafish swim faster when they are more aligned directionally, using a previously published dataset with only 3 replicate trials.

2 Results

The concept of an X-perfect or Y-perfect match.

Central to our approach is the concept of an X-perfect match or Y-perfect match. To motivate the ideas, suppose that a group of n graduate students $\boldsymbol{X} = (X_1, ..., X_n)$ has been paired with a group of n advisors $\boldsymbol{Y} = (Y_1, ..., Y_n)$ so that the ith student (X_i) is paired with the ith advisor (Y_i). Moreover, let $\rho(X_i, Y_j)$ be a

number that measures how well the ith student and the jth advisor get along. We say that the ith student is "happy" if they get along with their own advisor strictly better than they get along with any other advisor, meaning that $\rho(X_i, Y_i) > \rho(X_i, Y_j)$ for all $j \neq i$. Similarly, the ith advisor is "happy" if $\rho(X_i, Y_i) > \rho(X_j, Y_i)$. The arrangement of student-advisor pairs is X-perfect if all students are happy, and Y-perfect if all advisors are happy.

Analogously, if $\boldsymbol{X} = (X_1, ..., X_n)$ and $\boldsymbol{Y} = (Y_1, ..., Y_n)$ are collections of time series with n trials each, and if ρ is some correlation function, we say that an X-perfect match has occurred if (and only if) all X trials are "happy", meaning that:

$$\rho(X_i, Y_i) > \rho(X_i, Y_j) \text{ for all pairs } (i, j) \text{ such that } i \neq j.$$

Similarly, a Y-perfect match has occurred if and only if:

$$\rho(X_i, Y_i) > \rho(X_j, Y_i) \text{ for all pairs } (i, j) \text{ such that } i \neq j.$$

Throughout the article, we use bolded $\boldsymbol{X}$ (or $\boldsymbol{Y}$) to indicate a collection, X_i (or Y_i) to indicate a trial of the collection, and X (or Y) to refer to the variable in general.

A perfect match (either X-perfect or Y-perfect), provides especially strong evidence to support the hypothesis that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are dependent. In the following, we mostly write about the Y-perfect match, but it should be clear by symmetry that analogous results apply to the X-perfect match. Specifically, we consider the null hypothesis $H_0^{(Y)}$ wherein (1) the Y_i trials are mutually independent and follow identical distributions, and (2) $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent. Under this null hypothesis, the probability of observing a Y-perfect match cannot exceed $1/n^n$. In Appendix A2, we restate this as Lemma 21.

Like the definition of a Y-perfect match, its corresponding null hypothesis $H_0^{(Y)}$ is asymmetric: It requires the Y_i trials to iid, but it imposes no such requirement upon the X_i trials. It is possible to conceive of scenarios where one variable is approximately iid, but not the other, even when they are dependent. For example, in a yearly survey of a migratory animal species and a non-migratory animal species at different locations (trials), the population sizes of the non-migrator may be approximately iid across locations, whereas those of the migrator are likely dependent (due to migration) [217]. Alternatively, consider a microtiter plate whose wells (trials) contain two-species microbial communities in which species X does not affect species Y, but Y releases compounds that affect X, including a gas that diffuses to neighboring wells. Then, the population size of Y is iid across wells, and X is dependent on Y, yet the population size of X is non-iid because of the gas diffusion.

A modified permutation test for dependence

We now describe our modified permutation test for dependence between nonstationary time series. Suppose that an experiment produces time series $\boldsymbol{X} = (X_1, ..., X_n)$ and $\boldsymbol{Y} = (Y_1, ..., Y_n)$, where X_i and Y_i are time series obtained from the ith replicate, and we want to test whether $\boldsymbol{X}$ and $\boldsymbol{Y}$ are dependent. For example, we may have video recordings of n animals in separate but identically constructed enclosures; for the ith animal, we extract a time series of movement speed (X_i) and a time series of the distance from a light source (Y_i). Using these data, we may wish to investigate whether the animals under study tend to move at different speeds depending on their distance from the light source.

First, choose some function ρ to measure the correlation between a pair of time series. The Pearson correlation coefficient may be used, but ultimately we may choose any function that maps a pair of time series to a number. Next, perform a permutation test using the average correlation (Fig 49A). That is, let p_{perm} be the proportion of all possible permutations of $(Y_1, ..., Y_n)$ (including the original ordering) that produce an average correlation that is at least as large as the original ordering.

Next, check for a Y-perfect match using the same correlation function ρ (Figure 49B). Define a new variable p_{match}, and let $p_{match} = 1/n^n$ if a Y-perfect match is observed. Otherwise, let $p_{match} = 1$.

Finally, report a p-value as the lower of p_{perm} and p_{match}. That is,

$$p = \min(p_{perm}, p_{match}). \tag{15}$$

We will refer to this as the permute-match test (Fig 49C). The complete null hypothesis of this test is the same as that of the Y-perfect match test (i.e. the Y_i replicates are iid, and $\boldsymbol{X}$ is independent of $\boldsymbol{Y}$). Under this null hypothesis, the permute-match test is valid, meaning that for any significance level α, we have $P(p \leq \alpha) \leq \alpha$ where P denotes probability. This fact is stated and proven as Theorem 23 in Appendix A2.

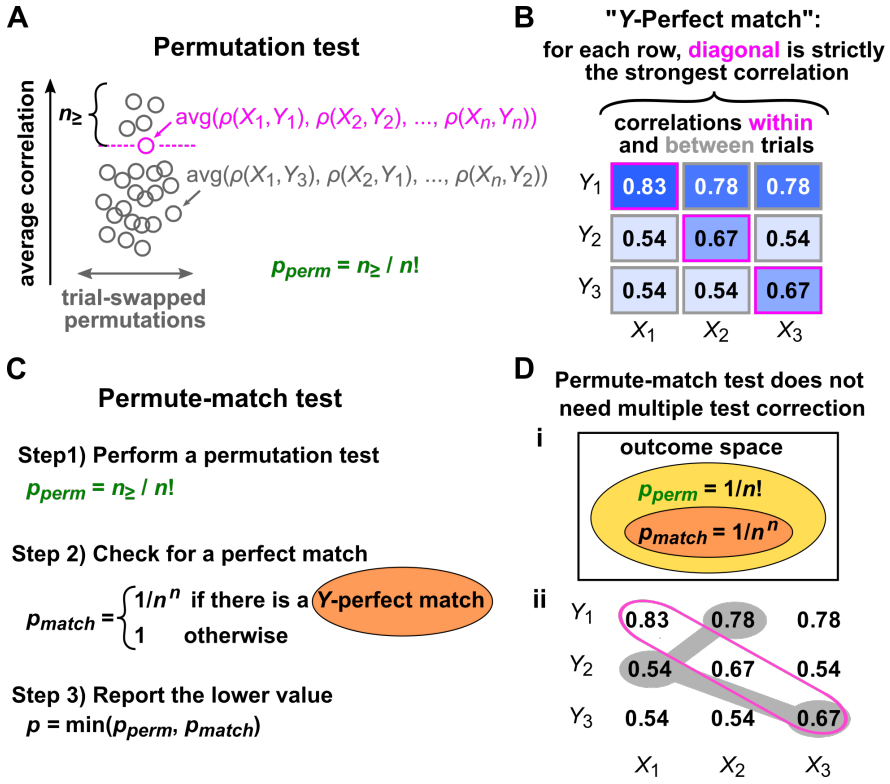


Figure 49: The permutation test, perfect match concept, and permute-match test. (**A**) The permutation test. Circles indicate the average correlation obtained with (grey) or without (pink) trial swapping. Here, $n_{\geq}$ is the number of average correlations (including the original one) that are equal to or larger than the original correlation. p_{perm} is then $n_{\geq}/n!$ because $n!$ is the total number of average correlations (with and without trial swapping). (**B**) The perfect match concept. For each row, the correlation on the diagonal (i.e. $\rho(X_i, Y_i)$) is strictly stronger than other correlations in the same row (i.e. $\rho(X_j, Y_i)$). This is a "Y-perfect match". (**C**) The permute-match test proceeds in three steps. First, perform a permutation test based on average correlations as in (A). Second, check for a Y-perfect match, and let p_{match} be $1/n^n$ (if a Y-perfect match occurs) or 1 (if not). Lastly, report the lower of the two intermediate p-values. (**D**) Although the permute-match test involves taking the lower of two p-values, it does not require a multiple test correction. (**i**) This is because a perfect match can only occur if the permutation test has achieved the lowest possible p-value ($1/n!$). (**ii**) To see why, note that the average original correlation (pink dot in A) is the average of the diagonal values from the correlation matrix (pink outline), whereas an average permuted correlation (grey dots in A) is an average of n terms from the correlation matrix with the condition that each row and each column contributes exactly one term (e.g. grey shade). If there is a Y-perfect match, then any off-diagonal grey-encircled value must be less than the on-diagonal value in its row. Thus, the average original correlation (pink) is greater than any average permuted correlation (grey), and so $p_{perm} = 1/(n!)$.

Note that the permute-match test (Fig 49C) is an "upgrade" to the permutation test (Fig 49A) since the permute-match test always reports the same or lower p-value. Also note that we do not need a multiple hypothesis test correction for Eq. 15 (see Appendix A2 for a formal justification). Intuitively, this is because the permute-match test consists of two tests that are *nested*. That is, a perfect match is possible only if p_{perm} is already at its lowest possible value of $1/n!$ (e.g. Fig 49D). Thus, unlike combining un-nested tests where each can separately contribute false positives, we can safely combine the perfect match test and the

permutation test using Eq. 15.

The permute-match test has a slightly more restrictive data requirement than the permutation test. The permutation test is valid either if the X_i trials are exchangeable, or if the Y_i trials are exchangeable (Corollary 18 in Appendix A1). The permute-match test requires that the Y_i trials are iid. If trials are iid, then they are necessarily exchangeable. However, exchangeable trials are not always iid. For example, the first five cards drawn from a shuffled deck are exchangeable because all possible orderings of cards are equally likely, but they are not iid because if the first card is an ace of hearts, the second card cannot be the ace of hearts. In practice, biological experimental designs typically use replicates that are both exchangeable and iid, so the more restrictive requirement of the permute-match test is often inconsequential.

Illustration of permute-match test with synthetic data

Here, using a simulated nonstationary process, we compare the behavior of the permute-match test and the permutation test. Consider time series of two variables (X and Y) from a small number of replicates (between 3 and 5). In this case, the time series are simply a linear time trend with coupled additive noise ($r_{X,Y}$ determines the strength of coupling; Fig 50A). We use the Pearson correlation coefficient as our correlation function ρ. Fig 50B shows how the power (i.e. "true positive rate"; proportion of simulations in which dependence is detected) of both tests varies with coupling strength, replicate number, and the significance level. For significance levels that are accessible to both tests, the detection power is the same (Fig 50B, e.g. the two white-shaded curves in column 2 line up). However, if coupling is sufficiently strong, the permute-match test can detect dependence at a more confident significance level than the permutation test (Fig 50B, more curves in bottom row than in top row).

A simple example was chosen here for ease of presentation. However, this example is so simple that it may be correctly treated by just fitting and removing the linear trend. In Appendix A3, we illustrate the same ideas using a parallel example with a more complicated nonstationary system (a logistic map with time-varying parameters) and a nonlinear correlation function.

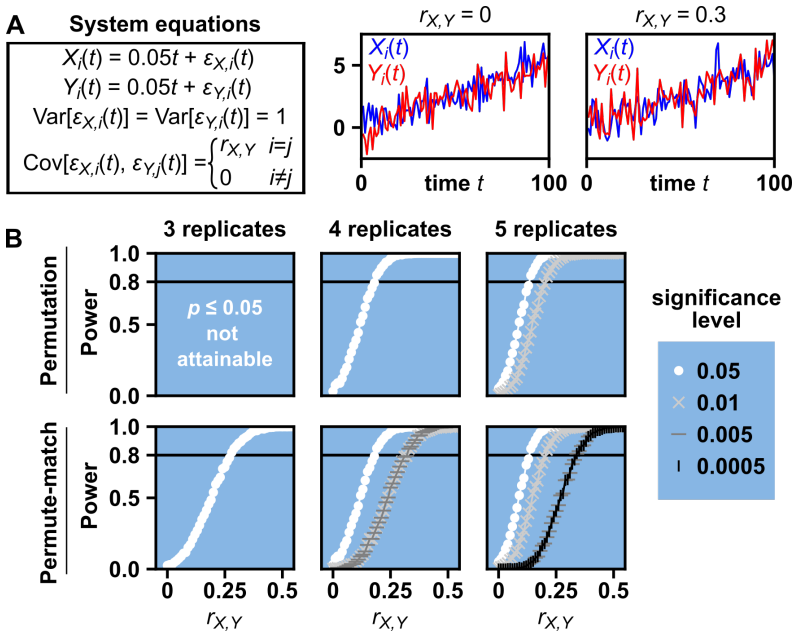


Figure 50: Statistical power of the permute-match test and permutation test in a simple nonstationary system. (**A**) System equations and example dynamics. The processes X and Y are given by a linear trend with additive noise. The noise terms $\epsilon_{X,i}(t)$ and $\epsilon_{Y,i}(t)$ are drawn from a bivariate normal distribution with a mean of 0, variance of 1, and covariance of $r_{X,Y}$. The charts show example time series where X and Y are either independent (middle; $r_{X,Y} = 0$) or dependent (right; $r_{X,Y} = 0.3$). (**B**) Statistical power of the permutation test and permute-match test as a function of the number of replicates, the significance level, and $r_{X,Y}$. Power was calculated from 2000 simulations at each value of $r_{X,Y}$. All curves that are shared between the permutation test and permute-match test are the same measurements, since the tests are nested (Fig 49B). For the permute-match test, we checked for a Y-perfect match, although this choice does not matter since both the ground-truth system and the correlation function are symmetric in this example.

The chances of observing an X-perfect or Y-perfect match may differ

If both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are iid across trials, one may test for dependence either by checking for an X-perfect match, or by checking for a Y-perfect match. In general, the probability of observing an X-perfect match may differ from the probability of observing a Y-perfect match, and it is useful to check for the more likely perfect match in order to maximize statistical power.

To obtain some intuition about which perfect match is more likely to occur (X-perfect or Y-perfect), we return to the analogy of students (X) and advisors (Y). If one particular advisor is far superior to the others, it may be that all students would prefer to be paired with this superstar advisor. But since only one student can be paired with the superstar advisor, obtaining an X-perfect match is essentially impossible. By symmetry, if just one student is far superior to the others, a Y-perfect match may be difficult to obtain.

To demonstrate this principle, consider a system that is similar to the trend-plus-noise system from Fig 50, but with one important change: Now, the slope of X_i is not fixed, but is instead determined by the parameter $S_{X,i}$ (Fig 51A). Conversely, the slopes of all Y trials are fixed. Note that the slope of an X time

series influences its correlation with a Y time series, both within and between trials (Fig 51B). In this way, an X time series with a high slope is like the "superstar" student in the previous paragraph. Indeed, in one example realization (Fig 51C), the X time series with the highest slope produces the largest correlations, both within and between trials. Since each of the Y time series prefers this high-slope X time series but only one Y time series can be paired to it, a Y-perfect match does not occur. However, the X-perfect match does occur. Strikingly, even if X's slope $S_{X,i}$ is always lower than the (fixed) slope value of Y, an X-perfect match is observed more frequently than a Y-perfect match (Fig 51D).

In general, if X (but not Y) has parameters that vary between trials in a way that universally increases or decreases inter-time-series correlations (like the slope in this example), we expect that an X-perfect match will be more likely than a Y-perfect match. Note that this asymmetry does not exist for the permutation test.

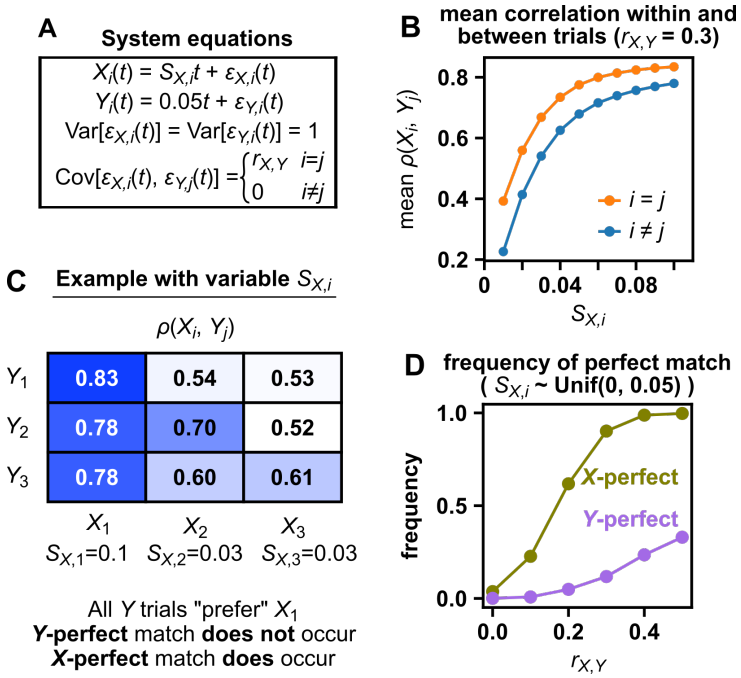


Figure 51: The probabilities of obtaining an X-perfect match and Y-perfect match may differ. (**A**) System equations. The processes X and Y are given by a linear trend with additive noise. The noise terms $\epsilon_X(t)$ and $\epsilon_Y(t)$ are drawn from a bivariate normal distribution with a mean of 0, variance of 1, and covariance of $r_{X,Y}$. The slope of the X trend is determined by $S_{X,i}$ and the slope of the Y trend is fixed. (**B**) The correlation between the two time series depends on the slope of the X time series. (**C**) Suppose that $S_{X,i}$ varies among three different trials, with $S_{X,1} = 0.1; S_{X,2} = 0.03; S_{X,3} = 0.03$. Then the column corresponding to $S_{X,1} = 0.1$ tends to dominate the table of correlations. Because of this, a Y-perfect match is unlikely to occur. In this example, we again used a coupling strength of $r_{X,Y} = 0.3$. (**D**) The frequency of an X-perfect or Y-perfect match when $S_{X,i}$ is drawn from a uniform distribution between 0 and 0.05. Values in D were calculated from 1000 simulations of checking for a perfect match with 3 trials. In Panels C-D, the Pearson correlation coefficient was used as the correlation function ρ, and the time series length was 100.

An example from animal behavior science

Living in groups is a common experience among animals. For example, at least half of fish species are thought to form groups at some stage of life [218, 176]. The properties of these groups can impart a variety of important fitness effects [219]. In groups of fish, coordinated swimming may facilitate foraging (e.g. by enabling fish to "communicate" the location of food to others) and predator escape (e.g. by confusing predators) [220, 176]. One study [176] used videos of small groups of zebrafish (8 fish per group) to observe that fish swam faster during segments when they were in a high-polarization state (i.e. when they were directionally aligned) than during segments (of the same video) when they were in a low-polarization state. A correlation between polarization and speed can be due to statistical dependence between the two variables, or due to temporal trends (e.g. polarization and swimming speed both happen to decrease with time).

Using a publicly available data set [175] containing fish trajectories from 3 replicate 10-minute videos of small groups of juvenile *Danio rerio* zebrafish (10 fish per group), we performed a permute-match test of dependence between polarization and fish swimming speed. As we will see, although these quantities are potentially nonstationary, the permute-match test nevertheless detects a statistical dependence between them.

To quantify polarization, we use "circular variance" (v_{circ}; see Methods), a common measure of angular dispersion [221]. Since high polarization means low dispersion, we quantify polarization as $1 - v_{circ}$, and call this quantity the *circular concentration*, since it indicates "how concentrated the [circular] data is toward the center" ([222] pg. 15). The circular concentration is bounded between 0 (low alignment) and 1 (perfect alignment). To quantify speed, we took the "average individual speed", which is a term we use to denote an average across individuals, not across time. More precisely, the average individual speed of a 10-fish group at time t is $(s_{1,t} + s_{2,t} + \cdots + s_{10,t})/10$ where $s_{i,t}$ is the speed of fish i at time t. Note that average individual speed is distinct from the speed of the group center, which has also been studied in *Danio* fish [219].

Neither average individual speed nor circular concentration is obviously stationary. By visual inspection, the average individual speed seems to decrease across time, and all time series are deemed nonstationary by a KPSS test (Fig 52A; $p < 0.01$), and so we may be on shaky ground to use methods that require stationarity. Additionally, since only 3 trials are available, the permutation test cannot detect dependence at the 0.05 level and, so we focus on checking for a perfect match.

Applying the perfect match test to these data, we found that the correlation between average individual speed and circular concentration is significant (Fig 52B; $p = 1/3^3 \approx 0.04$). Both a "speed"-perfect and a "circular concentration"-perfect match is obtained, and so in this case, significance does not depend on which match is tested. Additionally, the match test can detect dependence using as little as the first 20

seconds (= 640 frames at 32 frames per second) of the 10-minute data (Fig 52D). Note that although trials are independent of each other, all between-trial correlations from the full data set are positive (Fig 52B), emphasizing the danger in applying naive data analysis methods to data that are autocorrelated or even nonstationary.

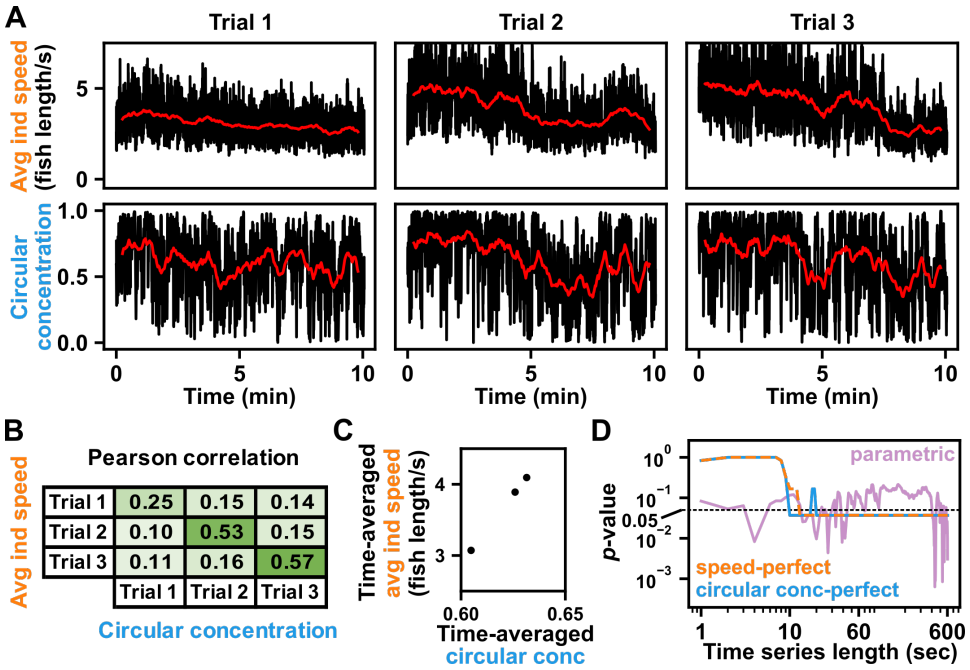


Figure 52: Dependence between average individual speed and circular concentration in small groups of zebrafish. (**A**) Time series of average individual speed and circular concentration for the three replicate videos. Black curves show original time series, and red curves show a 30-second moving average. Average individual speed appears to decrease with time in all trials. All 12 time series (2 variables $\times$ 3 trials $\times$ 2 smoothing conditions) were deemed nonstationary by a Kwiatkowski–Phillips–Schmidt–Shin (KPSS; [181]) test ($p < 0.01$). Note that the KPSS test seeks to reject a stationary null hypothesis in contrast to other common tests, whose null hypotheses are nonstationary. (**B**) Pearson correlation between circular concentration and average individual speed, both within trials and between trials, using the full 10 minutes of data. Table entries are shaded by correlation. We observe a speed-perfect match (and a circular concentration-perfect match), and thus detect dependence with $p = 1/3^3 \approx 0.04$. (**C**) To try a parametric alternative, for each trial we averaged values of speed and circular concentration over the full 10 minutes (e.g. the scatter plot shown, where each point is one trial). The sample Pearson correlation of the time-averaged variables is 0.99996, and a standard one-tailed test of significance (as implemented in the function stats.pearsonr from the Python package Scipy [223]) gives $p \approx 0.003$. The null hypothesis of this parametric test is that the points in (C) are drawn from a bivariate Gaussian distribution with zero covariance. This parametric test is comparable to the permute-match test because it is valid even when the time series are nonstationary. (**D**) The permute-match test can detect dependence with fewer data. We truncated time series to different lengths (starting from the first frame), and compared the parametric test with the two possible permute-match tests, i.e. either checking for a speed-perfect match or a circular-concentration-perfect match. As long as we use more than the first 20 seconds of data, the permute-match test detects the dependence at the 0.05 level between average individual speed and circular concentration. Conversely, the outcome of the parametric test is more sensitive to changes in the time series length. We used the Pearson correlation (not its magnitude) in the permute-match tests since the alternative hypothesis is that the correlation is positive, not merely nonzero.

3 Discussion

Outside the context of time series, permutation tests have been celebrated because they require only minimal assumptions [224]. To satisfy the requirements of permutation tests, it is sufficient (see Corollary 18 in Appendix A1) to collect data from exchangeable replicates. Data in each trial can be autocorrelated (e.g. temporally or spatially) and nonstationary. Such radical freedom from assumptions stands in stark contrast to the situation in the analysis of single-replicate time series, where the practitioner may be forced to make assumptions that can be difficult to verify. For example, most statistical tests of correlation between two time series require stationary data at a minimum [139]. Yet, it is difficult to verify that a single time series is stationary because stationarity is a property of an entire ensemble of time series [139], so that checking for stationarity in a single time series is philosophically analogous to checking for the Gaussianity of a single data point.

However, applying a trial-swapping permutation test to check for significant correlations requires that a sufficient number of trials be available to permute amongst one another. When only small numbers of trials are available, the permutation test is severely limited by mathematics. The minimum p-value that can possibly be achieved by a permutation test is $(n!)^{-1}$, which will provide only modest evidence against a null hypothesis when $n = 4$ and is essentially useless when $n \leq 3$, regardless of how strong the "true" correlation may be. Here, we have provided the permute-match test (Fig 49), a slightly modified permutation test. The permute-match test can reduce the minimum achievable p-value to $(n^n)^{-1}$ while imposing only a very minor additional data requirement (i.e. iid replicates in at least one variable). Even when a permutation test has detected dependence, the permute-match test can access lower p-values and thus provide greater confidence in an effect. The statistical power of the permute-match test can be further strengthened by appropriately choosing whether to check for an X-perfect or Y-perfect match (Fig 51).

Since permutation-based tests such as the permute-match test can be easy to misinterpret [225], we wish to alert readers to a possible point of confusion, namely the distinction between population-level and individual-level dependence results. A rejection of the permute-match test's null hypothesis indicates a dependence between $\boldsymbol{X}$ and $\boldsymbol{Y}$. It is important to note that this is a population-level result and does not describe individual-level dependencies. In other words, this population-level dependence condition does not necessarily imply that each X_i replicate is dependent on each Y_i replicate. In fact, it is theoretically possible for the set (X_1, X_2) to be dependent on the set (Y_1, Y_2) without either X_1 depending on Y_1 or X_2 depending on Y_2. (One such situation occurs when X_1, X_2, Y_1, and Y_2 are each "coin flips" taking on 0 or 1 with equal chance, but such that $X_1 = X_2$ if and only if $Y_1 = Y_2$). However, we find it difficult to envision a similar counterexample that is biologically realistic. Indeed, in many cases where population-level dependence is

observed, the scientist may reasonably also expect that most or all X_i replicates are dependent on their respective Y_i replicates, but this logical step should flow from scientific reasoning, not mathematical fact.

Although we have here used language and examples involving time series, the permute-match test can in principle be applied to other scenarios where iid replicates are available, but data within each replicate are dependent. One example is spatial data in brain imaging, where the permutation test has been performed using multiple scanned brains [126]. Beyond spatial processes, dependent data also appear in nucleotide sequences and natural language text. Overall, methods that take advantage of multiple replicates will facilitate statistical testing without requiring assumptions that are difficult to verify.

4 Methods

Data preprocessing

We obtained fish trajectory data from the web address at `https://drive.google.com/drive/folders/1UmzlX-yJhzQ5KX5rGry8wZgXvcz6HefD`. For the first trial we used the file at the path "10/1/trajectories_wo_gaps.npy" (where "10" indicates the number of fish and "1" indicates the first trial). For the second and third trials, the file path was the same except it began with "10/2/" and "10/3/" respectively. These files contain sequences of fish positions indexed by video frame, as well as constants such as frame rate and approximate fish body length.

We estimated the velocity of each fish as

$$\overrightarrow{v}_{i,t} = \frac{\overrightarrow{p}_{i,t+1} - \overrightarrow{p}_{i,t-1}}{2} R$$

where $\overrightarrow{p}_{i,t}$ is a 2-dimensional column vector specifying the position of the ith fish in units of body length at video frame t, $\overrightarrow{v}_{i,t}$ is a 2-dimensional column vector specifying the fish velocity in units of body length per second at video frame t, and R is the frame rate (32 frames per second). The reason that a 2 appears in the denominator is because $\overrightarrow{p}_{i,t+1}$ is 2 frames after $\overrightarrow{p}_{i,t-1}$.

A group's average individual speed at each time was then given by

$$\frac{1}{N} \sum_{i=1}^{N} |\overrightarrow{v}_{i,t}|$$

where N is the number of fish. The circular concentration C_t of a group at each time was given by the

formula

$$C_t = \frac{1}{N}\sqrt{\left(\sum_{i=1}^{N}\cos\left(\theta_{i,t}\right)\right)^2 + \left(\sum_{i=1}^{N}\sin\left(\theta_{i,t}\right)\right)^2}$$

where

$$\theta_{i,t} = \mathrm{atan2}(v_{x,i,t}, v_{y,i,t}).$$

The calculation was carried out using the function "stats.circstats.circvar" in the Python package Astropy (ver. 3.2.2) [226]. Here, atan2 is the 2-argument arctangent function (i.e. "arctan2" in the Python package Numpy, ver. 1.21.6). Also, $v_{x,i,t}$ and $v_{y,i,t}$ denote the x and y components of the velocity vector $\overrightarrow{v}_{i,t}$.

Statistical analysis

The permute-match test is described in Fig 49 and its validity is proved in Appendix A2. For the KPSS test, we used the function "statsmodels.tsa.stattools.kpss" in the Python package Statsmodels (ver. 0.10.1) [67]. We set the "regression" argument to "c" (which tests the null hypothesis of stationarity rather than trend-stationarity), and we set the "nlags" argument to "auto" (which means that the number of lags used for testing stationarity is automatically chosen by a data-dependent method).

Appendix to Part III

In Appendix A1, we review the basic permutation test of independence and a proof of its validity. This is because, although these results are known to the statistical community, they are often presented in an abstract form whose connection to independence testing may not be obvious to the nonspecialist, or parts of the argument are simply left as an exercise to the reader. In Appendix A2, we provide the theoretical justification for the permute-match test, which is the main topic of this work. Appendix A3 contains supplementary data.

A1 Justification of the permutation test of independence

In this section we provide a proof that the permutation test of independence is valid. No major novelty is claimed in this subsection. For instance, the results of Lemma 17 can be mostly reconstructed by piecing together various theorems, examples, and homework problems from section 15.2.1 of Lehmann and Romano [124]. Nevertheless, we present the complete argument here as a courtesy to the reader. Related proofs or proof sketches of various kinds of permutation tests can be found in [214, 227].

The following lemma describes a general rank-based method to test whether the distribution of a random variable changes upon applying a set of transformations. This lemma and proof are based on Theorem 15.2.1 in Lehmann and Romano [124]. We begin with this result because it comes first in the chain of logical steps that justify the permutation test, but it is somewhat abstract, and so the reader who prefers to start out with concrete concepts may wish to skip to Def 16 and return to this lemma when its necessity becomes apparent later. As a notational clarification, note that for the transformation h_i, to denote "h_i applied to Z" we write $h_i Z$ rather than the more traditional $h_i(Z)$.

Lemma 15

Let $Z \in \mathcal{Z}$ be a random variable and let $H = \{h_1, ..., h_m\}$ be a set of functions from $\mathcal{Z}$ to $\mathcal{Z}$ such that:

1. the probability distributions of Z and $h_i Z$ are the same for all functions h_i in H, and
2. the unordered sets $\{h_1 Z, ..., h_m Z\}$ and $\{h_1 h_i Z, ..., h_m h_i Z\}$ are equal for all h_i in H.

Let the statistic T be a function from $\mathcal{Z}$ to $\mathbb{R}$. Given some $Z = z$, let

$$T^{(1)}(z) \leq T^{(2)}(z) \leq \cdots \leq T^{(m)}(z)$$

be the ordered values of $\{T(h_1 z), ..., T(h_m z)\}$. Choose some significance level $\alpha \in [0, 1]$ and let the rank k be

$$k = m - \lfloor m\alpha \rfloor$$

where $\lfloor m\alpha \rfloor$ is the largest integer less than or equal to $m\alpha$. Then,

$$P(T(Z) > T^{(k)}(Z)) \leq \alpha.$$

Proof: Let $I(A)$ be the indicator function for the event A (meaning that $I(A) = 1$ if A occurs and $I(A) = 0$ otherwise). If there are no ties for $T^{(k)}(Z)$, then

$$\begin{aligned} \lfloor m\alpha \rfloor &= \sum_{i=1}^{m} I(T(h_i Z) > T^{(k)}(Z)) \\ &= \sum_{i=1}^{m} I(T(h_i Z) > T^{(k)}(h_i Z)). \end{aligned}$$

To arrive at the second equality, we used the fact that $T^{(k)}(Z) = T^{(k)}(h_i Z)$, which follows from the second requirement of the lemma (i.e. $\{h_1 Z, ..., h_m Z\} = \{h_1 h_i Z, ..., h_m h_i Z\}$).

If there is a tie for $T^{(k)}(Z)$ (meaning that there is some $j \neq k$ such that $T^{(j)}(Z) = T^{(k)}(Z)$), then we have a similar formula, but with an inequality instead of an equality:

$$\lfloor m\alpha \rfloor \geq \sum_{i=1}^{m} I(T(h_i Z) > T^{(k)}(h_i Z)).$$

This version with the inequality is of course general since it is true under both cases. Using this formula,

$$m\alpha \geq \lfloor m\alpha \rfloor = E[\lfloor m\alpha \rfloor] \geq \sum_{i=1}^{m} E\left[I(T(h_i Z) > T^{(k)}(h_i Z))\right].$$

Since Z and $h_i Z$ have the same distribution, we may remove the h_i in the above formula, and we have

$$m\alpha \geq \sum_{i=1}^{m} E\left[I(T(Z) > T^{(k)}(Z))\right] = mE\left[I(T(Z) > T^{(k)}(Z))\right].$$

Dividing through by m gives

$$\alpha \geq E\left[I(T(Z) > T^{(k)}(Z))\right] = P(T(Z) > T^{(k)}(Z))$$

as required.

Next we give a definition of the permutation test. The definition here is somewhat more general than in the main text. As with h_i, we will denote "g_i applied to Z" by writing $g_i Z$.

Definition 16 *The permutation test of independence*

Let $\boldsymbol{X} = (X_1, ..., X_n)$ and $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be sequences of random variables. Let $G = \{g_1, ..., g_{n!}\}$ be the complete set of functions that permute a sequence of length n. For example, if $\boldsymbol{Y}$ is $(2, 7, 4)$, then $g_i \boldsymbol{Y}$ might be $(2, 4, 7)$. Define the statistic $T(\boldsymbol{X}, \boldsymbol{Y})$ to be a function that returns a real number, typically interpreted as the overall correlation strength. Repeatedly compute the statistic after permuting the elements of $\boldsymbol{Y}$. That is, compute

$$\{T(\boldsymbol{X}, g_1\boldsymbol{Y}), T(\boldsymbol{X}, g_2\boldsymbol{Y}), ..., T(\boldsymbol{X}, g_{n!}\boldsymbol{Y})\}.$$

Let $n_{\geq}$ be the number of permuted statistic values that are at least as large as the original statistic value. That is,

$$n_{\geq} = \sum_{i=1}^{n!} I(T(\boldsymbol{X}, g_i\boldsymbol{Y}) \geq T(\boldsymbol{X}, \boldsymbol{Y}))$$

where $I(A)$ is the indicator function for the event A (meaning that $I(A) = 1$ if A occurs and $I(A) = 0$ otherwise). Then the p-value of the test is given by

$$p_{perm} = n_{\geq}/n!$$

We now prove the validity of the permutation test.

Lemma 17 *The permutation test of independence is valid when $\boldsymbol{Y}$ is exchangeable.*

Let $\boldsymbol{X} = (X_1, ..., X_n)$ and $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be sequences of random variables such that $\boldsymbol{Y}$ is exchangeable and where $\boldsymbol{X}$ is independent of $\boldsymbol{Y}$. Perform the permutation test of independence (Def. 16) to obtain p_{perm}. Then, $P(p_{perm} \leq \alpha) \leq \alpha$ for any $\alpha \in [0, 1]$.

Proof: Although the symbols g_i and T were not defined in the lemma statement, we use them here with their meanings from Def. 16.

We first set up the problem in a way that allows us to apply Lemma 15. To construct the random variable Z and transformation set H for Lemma 15, we choose $Z = (\boldsymbol{X}, \boldsymbol{Y})$ and $h_i Z = (\boldsymbol{X}, g_i \boldsymbol{Y})$. Lemma 15 has two requirements. First, Z and $h_i Z$ must have the same distribution. This requirement is satisfied: Since $\boldsymbol{Y}$ is exchangeable and $\boldsymbol{X}$ is independent of $\boldsymbol{Y}$, it follows that $(\boldsymbol{X}, \boldsymbol{Y})$ has the same distribution as $(\boldsymbol{X}, g_i \boldsymbol{Y})$. The second requirement is that $\{h_1 Z, ..., h_m Z\} = \{h_1 h_i Z, ..., h_m h_i Z\}$. This requirement is also satisfied: The set of permutations of $\boldsymbol{Y}$ is the same as the set of permutations of $g_i \boldsymbol{Y}$. Since both requirements are satisfied, we may apply Lemma 15.

We now prove the result directly. Below, $A \leftrightarrow B$ means "A if and only if B".

$$\begin{aligned}
p_{perm} \leq \alpha &\leftrightarrow \sum_{i=1}^{n!} I(T(\boldsymbol{X}, g_i \boldsymbol{Y}) \geq T(\boldsymbol{X}, \boldsymbol{Y})) \leq n! \alpha \\
&\leftrightarrow \sum_{i=1}^{n!} (1 - I(T(\boldsymbol{X}, g_i \boldsymbol{Y}) \geq T(\boldsymbol{X}, \boldsymbol{Y}))) \geq n! - n! \alpha \\
&\leftrightarrow \sum_{i=1}^{n!} (I(T(\boldsymbol{X}, g_i \boldsymbol{Y}) < T(\boldsymbol{X}, \boldsymbol{Y}))) \geq n! - n! \alpha \\
&\leftrightarrow \sum_{i=1}^{n!} (I(T(\boldsymbol{X}, g_i \boldsymbol{Y}) < T(\boldsymbol{X}, \boldsymbol{Y}))) \geq n! - \lfloor n! \alpha \rfloor .
\end{aligned}$$

In the language of Lemma 15, the last line is equivalent to $T(\boldsymbol{X}, \boldsymbol{Y}) > T^{(k)}(\boldsymbol{X}, \boldsymbol{Y})$ where $k = n! - \lfloor n! \alpha \rfloor$. Thus, we may directly apply Lemma 15, thereby obtaining

$$P(p_{perm} \leq \alpha) = P(T(\boldsymbol{X}, \boldsymbol{Y}) > T^{(k)}(\boldsymbol{X}, \boldsymbol{Y})) \leq \alpha$$

as required.

A close read of this result suggests that perhaps it is not as strong as it could be. Specifically, Lemma 17 requires that $\boldsymbol{Y}$ be exchangeable, but intuitively, permuting $\boldsymbol{Y}$ seems to be "the same" as permuting $\boldsymbol{X}$, so shouldn't it be equally sufficient for $\boldsymbol{X}$ to be exchangeable instead? As it turns out, this intuitive leap requires one additional requirement. Specifically, it is required that applying the same permutation to both $\boldsymbol{X}$ and $\boldsymbol{Y}$ must leave the value of $T(\boldsymbol{X}, \boldsymbol{Y})$ unchanged. That is, $T(\boldsymbol{X}, \boldsymbol{Y}) = T(g_i \boldsymbol{X}, g_i \boldsymbol{Y})$ for any permutation function g_i. Importantly, the form of T that is used for the permute-match test (i.e. $T(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, Y_i)$) satisfies this requirement because in this case, $T(g_i \boldsymbol{X}, g_i \boldsymbol{Y})$ simply rearranges the order of the terms in the summation, which clearly has no effect on the value of T.

Corollary 18 *If T is "nice", then the permutation test is valid when either $\boldsymbol{X}$ or $\boldsymbol{Y}$ is exchangeable.*

Let $\boldsymbol{X} = (X_1, ..., X_n)$ and $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be sequences of random variables such that at least one of $(\boldsymbol{X}, \boldsymbol{Y})$ is exchangeable and where $\boldsymbol{X}$ is independent of $\boldsymbol{Y}$. Perform the permutation test of independence (Def. 16) using a correlation function with the property that

$$T(\boldsymbol{X}, \boldsymbol{Y}) = T(g_i \boldsymbol{X}, g_i \boldsymbol{Y}) \tag{16}$$

for any permutation function g_i. Then, $P(p_{perm} \leq \alpha) \leq \alpha$ for any $\alpha \in [0, 1]$.

Proof: Lemma 17 already covers the case where $\boldsymbol{Y}$ is exchangeable. We now consider the case where $\boldsymbol{X}$ is exchangeable, but not $\boldsymbol{Y}$.

We proceed by showing that the permutation test where $\boldsymbol{Y}$ is permuted is equivalent to the permutation test where $\boldsymbol{X}$ is permuted. In other words,

$$\{T(\boldsymbol{X}, g_1 \boldsymbol{Y}), ..., T(\boldsymbol{X}, g_{n!} \boldsymbol{Y})\} = \{T(g_1 \boldsymbol{X}, \boldsymbol{Y}), ..., T(g_{n!} \boldsymbol{X}, \boldsymbol{Y})\}$$

where the left and right sides both denote unordered sets. It will be useful to refer to the inverse permutation g_i^{-1}, which is the permutation that "undoes" g_i. More precisely, g_i^{-1} is defined by the relation $g_i^{-1} g_i Z = Z$. Using Eq. 16, we have:

$$\begin{aligned} T(\boldsymbol{X}, g_i \boldsymbol{Y}) &= T(g_i^{-1} \boldsymbol{X}, g_i^{-1} g_i \boldsymbol{Y}) \\ &= T(g_i^{-1} \boldsymbol{X}, \boldsymbol{Y}). \end{aligned} \tag{17}$$

Note that the inverse permutation g_i^{-1} is guaranteed to exist because any permutation g_i can be "undone" by some g_i^{-1}. In the edge case where g_i is the trivial identity permutation, g_i^{-1} is also the identity permutation. Additionally, the set of permutation functions is the same as the set of inverse permutation functions, meaning that

$$\{g_1, ..., g_{n!}\} = \{g_1^{-1}, ..., g_{n!}^{-1}\} \tag{18}$$

because each permutation is an inverse permutation, and each inverse permutation is a permutation. It follows that

$$\begin{aligned} \{T(\boldsymbol{X}, g_1 \boldsymbol{Y}), ..., T(\boldsymbol{X}, g_{n!} \boldsymbol{Y})\} &= \{T(g_1^{-1} \boldsymbol{X}, \boldsymbol{Y}), ..., T(g_{n!}^{-1} \boldsymbol{X}, \boldsymbol{Y})\} \\ &= \{T(g_1 \boldsymbol{X}, \boldsymbol{Y}), ..., T(g_{n!} \boldsymbol{X}, \boldsymbol{Y})\} \end{aligned}$$

where the first equality follows from Eq. 17 and the second equality follows from Eq. 18. This shows that the permutation test where $\boldsymbol{Y}$ is permuted is equivalent to the permutation test where $\boldsymbol{X}$ is permuted.

It is clear from Lemma 17 by symmetry that this alternative view of the permutation test (where $\boldsymbol{X}$ is permuted instead of $\boldsymbol{Y}$) could be used if the requirement of an exchangeable $\boldsymbol{Y}$ in Lemma 17 were replaced by the requirement of an exchangeable $\boldsymbol{X}$. Applying this exchangeable-$\boldsymbol{X}$ version of Lemma 17 completes the proof.

A2 Validity of the permute-match test

We now justify the validity of the permute-match test, which is the primary contribution of the present work. We begin with a result about dice rolls, which will prove to be useful shortly. Although this lemma may seem unrelated to the task of detecting dependence, its later use may be foreshadowed by mentioning that this lemma describes an event in which a die rolls an i on the ith roll, so that the result of the roll "perfectly matches" the index of the roll.

Lemma 19

Let $(C_1, C_2, ..., C_n)$ be iid discrete random variables where each $C_i \in \{1, 2, ..., n\}$. That is, each C_i is the result from an independent roll of the same (possibly unfair) n-sided die. Then, the probability of

$$C_1 = 1, C_2 = 2, ..., C_n = n$$

is at most $1/n^n$.

Proof:

$$\begin{aligned} P(C_1 = 1, C_2 = 2, ..., C_n = n) &= P(C_1 = 1)P(C_2 = 2)...P(C_n = n) \\ &= P(C_1 = 1)P(C_1 = 2)...P(C_1 = n) \end{aligned}$$

The first equality comes from the fact that the C_i variables are independent and the second equality comes from the fact that they are identically distributed. To simplify notation, let us define $P_i = P(C_1 = i)$. Since the P_i terms sum to 1, we may substitute $P_n = 1 - \sum_1^{n-1} P_i$ so that the above probability may be rewritten as $f(P_1, ..., P_{n-1})$, defined by:

$$f(P_1, ..., P_{n-1}) = \left(\prod_1^{n-1} P_i\right)\left(1 - \sum_1^{n-1} P_i\right) = P_1 P_2 ... P_n.$$

Next, we find the maximum of f over its domain, which we take to be the space of valid probabilities:

$$\left\{ P_1, ..., P_{n-1} : \sum_{i=1}^{n-1} P_i \leq 1 \text{ and } P_i \geq 0 \text{ for } i = 1, ..., n-1 \right\}$$

Since this space is closed and bounded, and since f is continuous and differentiable, we know that f has extrema in its domain (i.e. the extreme value theorem), and that these extrema can only occur where $\nabla f = 0$ and/or on the boundary. Setting $\nabla f = 0$ gives the system of equations:

$$\begin{bmatrix} 1/P_1 \\ \vdots \\ 1/P_{n-1} \end{bmatrix} \left(\prod_1^{n-1} P_i \right) \left(1 - \sum_1^{n-1} P_i \right) - \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \left(\prod_1^{n-1} P_i \right) = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}.$$

There are two categories of solutions to $\nabla f = 0$. The first category occurs when $P_1, ..., P_{n-1}$ are all nonzero. In this case, we may divide by $\left(\prod_1^{n-1} P_i \right)$ to obtain

$$\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \left(1 - \sum_1^{n-1} P_i \right) = \begin{bmatrix} P_1 \\ \vdots \\ P_{n-1} \end{bmatrix}.$$

In this solution, $P_1, ..., P_n$ are all equal to one another. Since these variables sum to 1, this extremum is achieved when $1/n = P_1 = P_2 = ... = P_n$, and thus $f = 1/n^n$ at this point.

The second category of solutions occurs when at least one of $P_1, ..., P_{n-1}$ is equal to zero. For example, one such solution is obtained if we set each of $P_1, ..., P_{n-1}$ to zero. For all solutions in this category, we have $f = 0$.

Additionally, on the boundary of the domain (which occurs when at least one of $P_1, ..., P_{n-1}$ is zero or when $\sum_1^{n-1} P_i = 1$), we again have $f = 0$.

Since $f(P_1 = 1/n, ..., P_{n-1} = 1/n) = 1/n^n$ is the only positive value among candidate extrema (i.e. solutions to $\nabla f = 0$ and boundary points), we conclude that $1/n^n$ is the maximum of f on its domain. That is,

$$P(C_1 = 1, C_2 = 2, ..., C_n = n) = f(P_1, ..., P_{n-1}) \leq 1/n^n$$

which is what we set out to show.

Next we define the Y-perfect match test. Recall from the main text that the null hypothesis of this test is that $Y_1, ..., Y_n$ are iid and the sequence $\boldsymbol{Y}$ is independent of $\boldsymbol{X}$.

Definition 20 *The Y-perfect match test*

Let $\boldsymbol{X} = (X_1, ..., X_n)$ and $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be sequences of random variables. Let ρ (the "correlation function") be a function that maps a pair (X_i, Y_j) to a real number. Let the p-value p_{match} be $1/n^n$ if:

$$\rho(X_i, Y_i) > \rho(X_j, Y_i)$$

for all pairs (i, j) such that $j \neq i$. Otherwise, let p_{match} be 1.

Lemma 21 *The Y-perfect match test is valid*

Let $\boldsymbol{X} = (X_1, ..., X_n)$ be a sequence of random variables and let $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be a sequence of iid random variables. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be independent. Perform the match test defined immediately above. Let α be a real number between 0 and 1. Then $P(p_{match} \leq \alpha) \leq \alpha$.

Proof: As a preliminary, we point out that $(Y_1, ..., Y_n)$ are iid given $\boldsymbol{X}$. To see this, first note that $(Y_1|\boldsymbol{X}, ..., Y_n|\boldsymbol{X})$ are conditionally independent:

$$\begin{aligned} P(Y_1 \in B_Y^{(1)}, ..., Y_n \in B_Y^{(n)}|\boldsymbol{X}) &= P(Y_1 \in B_Y^{(1)}, ..., Y_n \in B_Y^{(n)}) \\ &= \prod_{i=1}^{n} P\left(Y_i \in B_Y^{(i)}\right) \\ &= \prod_{i=1}^{n} P(Y_i \in B_Y^{(i)}|\boldsymbol{X}). \end{aligned}$$

where $B_Y^{(1)}, ..., B_Y^{(n)}$ are Borel sets. Second, note that $(Y_1|\boldsymbol{X}, ..., Y_n|\boldsymbol{X})$ are identically distributed:

$$P(Y_i \in B_Y^{(i)}|\boldsymbol{X}) = P(Y_i \in B_Y^{(i)}) = P(Y_j \in B_Y^{(i)}) = P(Y_j \in B_Y^{(i)}|\boldsymbol{X})$$

With that preliminary out of the way, begin the main argument now. Let us define the random variables $(C_1, ..., C_n)$ as follows:

$$C_i = \underset{j}{\text{argmax}} \left(\rho(X_j, Y_i)\right)$$

That is $C_i \in \{1, ..., n\}$ is defined to be the choice of j that maximizes $\rho(X_j, Y_i)$. In the case of a multi-way tie, C_i is chosen uniformly at random from among the maximizing index values. These C_i variables are useful because if $p_{match} = 1/n^n$, then it must be the case that

$$C_1 = 1, ..., C_n = n.$$

If we could apply Lemma 19 to $(C_1, ..., C_n)$ we would nearly be done. However, Lemma 19 requires iid random variables. Unfortunately, $(C_1, ..., C_n)$ are not iid since they have a common dependence on $\boldsymbol{X}$. Fortunately however, $(C_1|\boldsymbol{X}, ..., C_n|\boldsymbol{X})$ are (conditionally) iid. More precisely, since $C_i|\boldsymbol{X}$ is a function of $Y_i|\boldsymbol{X}$, it follows that

$$P(C_1 = c_1, ..., C_n = c_n|\boldsymbol{X}) = \prod_{i=1}^{n} P(C_i = c_i|\boldsymbol{X})$$

and

$$P(C_i = c_i|\boldsymbol{X}) = P(C_j = c_i|\boldsymbol{X})$$

for $c_i \in \{1, ..., n\}$.

We may now apply Lemma 19, which tells us that

$$P(C_1 = 1, ..., C_n = n|\boldsymbol{X}) \leq 1/n^n. \tag{19}$$

Putting it all together, we have

$$P(p_{match} \leq 1/n^n) \overset{(a)}{\leq} P(C_1 = 1, ..., C_n = n) \overset{(b)}{=} E\left[P(C_1 = 1, ..., C_n = n|\boldsymbol{X})\right] \overset{(c)}{\leq} E\left[1/n^n\right] \overset{(d)}{=} 1/n^n$$

where we have indexed the equalities and inequalities so that we could easily identify them now and give explanations for each: The inequality $\overset{(a)}{\leq}$ occurs because $p_{match} \leq 1/n^n$ implies $p_{match} = 1/n^n$, which implies $C_1 = 1, ..., C_n = n$. The equality $\overset{(b)}{=}$ uses the definition of conditional probability (e.g. [195] pg. 152). The inequality $\overset{(c)}{\leq}$ simply applies Eq. 19. The equality $\overset{(d)}{=}$ is trivial.

To complete the proof, note we may have $\alpha < 1/n^n$, in which case $P(p_{match} \leq \alpha) = 0$, or alternatively we may have $\alpha \geq 1/n^n$, in which case $P(p_{match} \leq \alpha) \leq 1/n^n$ as shown immediately above. In either case, $P(p_{match} \leq \alpha) \leq \alpha$, which is what we set out to show.

Next, a brief proof is given of the fact that the permutation test (with an appropriate choice of T) and the perfect match test are nested. Fig 49D already described the same claim, but using visual arguments instead of written ones.

Lemma 22

Let $\boldsymbol{X} = (X_1, ..., X_n)$ and $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be sequences of random variables. Let ρ be a function that maps a pair (X_i, Y_j) to a real number. Define the average correlation

$$T(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, Y_i). \tag{20}$$

Perform the permutation test of independence (Def. 16) using T as the statistic, and obtain the p-value p_{perm}. Also perform the match test (Def. 20) using ρ as the correlation function, and obtain the p-value p_{match}. Then, $p_{match} = 1/n^n$ implies $p_{perm} = 1/n!$.

Proof: Suppose $p_{match} = 1/n^n$. Then,

$$\rho(X_i, Y_i) > \rho(X_j, Y_i) \; \forall \; (i \neq j).$$

In other words, for each Y_i, its correlation term $\rho(\cdot, Y_i)$ is strictly maximized when Y_i is paired with X_i. Applying a permutation to $\boldsymbol{Y}$ (other than the trivial identity permutation) means that for at least two choices of i, the $\rho(X_i, Y_i)$ term in Eq. 20 becomes replaced with $\rho(X_j, Y_i)$ for some $j \neq i$. But $\rho(X_i, Y_i) > \rho(X_j, Y_i)$. Therefore, $T(\boldsymbol{X}, \boldsymbol{Y})$ must be strictly greater than all permuted versions $T(\boldsymbol{X}, g_i\boldsymbol{Y})$ (other than when g_i is the identity permutation). It follows that $p_{perm} = 1/n!$.

Theorem 23 *The permute-match test is valid*

Let $\boldsymbol{X} = (X_1, ..., X_n)$ be a sequence of random variables and let $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be a sequence of iid random variables. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be independent. Let the correlation function ρ be a function that maps a pair (X_i, Y_j) to a real number. Define the average correlation

$$T(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, Y_i).$$

Perform the permutation test of independence (Def. 16) using T as the statistic, and obtain the p-value

p_{perm}. Also perform the match test (Def. 20) using ρ as the correlation function, and obtain the p-value p_{match}. Then, for any significance level $\alpha \in [0, 1]$,

$$P(\min(p_{perm}, p_{match}) \leq \alpha) \leq \alpha.$$

Proof: Begin by expanding the expression of interest using the addition rule of probability to more clearly see each component.

$$\begin{aligned} P(\min(p_{perm}, p_{match}) \leq \alpha) &= P(p_{perm} \leq \alpha \cup p_{match} \leq \alpha) \\ &= P(p_{perm} \leq \alpha) + P(p_{match} \leq \alpha) - P(p_{perm} \leq \alpha \cap p_{match} \leq \alpha) \end{aligned} \tag{21}$$

Next, we consider three cases that α can fall into and show that the theorem holds for each. Either $\alpha < 1/n!$ or $1/n! \leq \alpha < 1$ or $\alpha = 1$. If $\alpha < 1/n!$, then p_{perm} cannot be less than or equal to α because p_{perm} cannot go below $1/n!$. In this case, Eq. 21 reduces to

$$P(\min(p_{perm}, p_{match}) \leq \alpha) = P(p_{match} \leq \alpha)$$

which is at most α by Lemma 21.

If $1/n! \leq \alpha < 1$, then by Lemma 22,

$$\begin{aligned} P(p_{match} \leq \alpha) &= P(p_{match} = 1/n^n) \\ &= P(p_{perm} = 1/n^n \cap p_{match} = 1/n!) \\ &\leq P(p_{perm} \leq \alpha \cap p_{match} \leq \alpha) \end{aligned}$$

where we use an inequality on the third line because $\alpha \geq 1/n! \geq 1/n^n$. However, we may also point the inequality the other way:

$$P(p_{match} \leq \alpha) \geq P(p_{perm} \leq \alpha \cap p_{match} \leq \alpha)$$

because the event on the right side is a subset of the event on the left. Since the inequality points in both directions, we have

$$P(p_{match} \leq \alpha) = P(p_{perm} \leq \alpha \cap p_{match} \leq \alpha).$$

Therefore, Eq. 21 reduces to

$$P(\min(p_{perm}, p_{match}) \leq \alpha) = P(p_{perm} \leq \alpha)$$

which is at most α by Lemma 17.

Finally, if $\alpha = 1$, then the theorem holds trivially because probabilities are always at most 1. Since the theorem holds in all three cases, it holds in general.

A3 Illustration of permute-match test with logistic map system

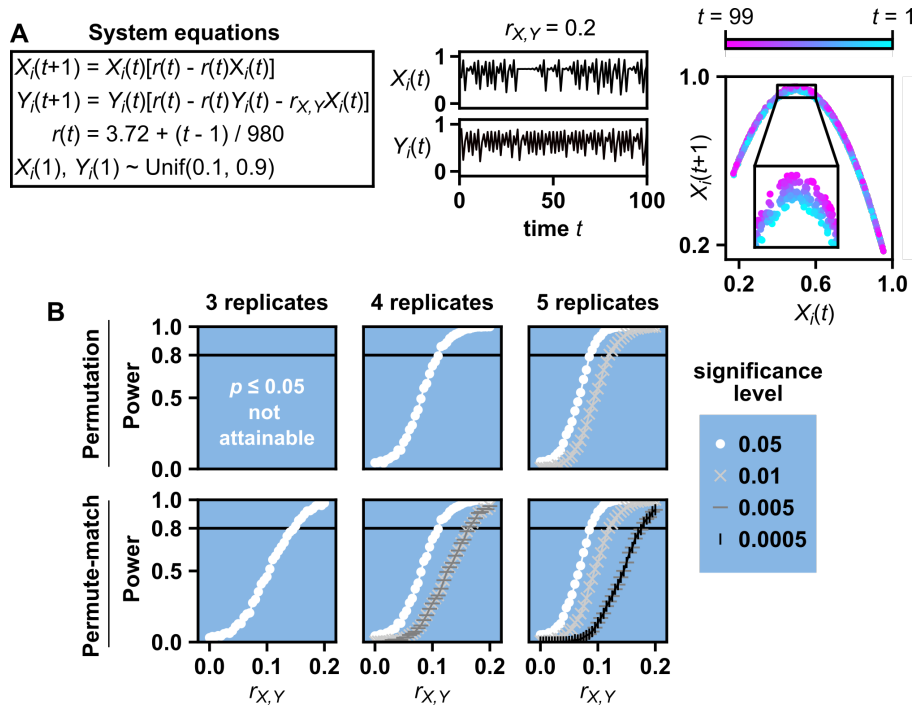


Figure A53: Statistical power of the permute-match test and permutation test in a nonlinear and nonstationary system. (**A**) System equations and example dynamics. The processes X and Y are given by a coupled logistic map. The system is nonstationary because the parameter $r(t)$ varies with time from 3.72 when $t = 1$ to 3.82 when $t = 99$. To see the nonstationarity clearly, the rightmost panel plots $X_i(t)$ against $X_i(t+1)$ and colors points by time, showing that as time goes on, there is an upward drift in the parabola that the points lie on. To better show the trend, 10 replicates are shown simultaneously in this chart. (B) Statistical power of the permutation test and permute-match test as a function of the number of replicates, the significance level, and $r_{X,Y}$. Power was calculated from 2000 simulations at each value of $r_{X,Y}$. For the permute-match test, we checked for a Y-perfect match. For both the permute-match test and the permutation test, the correlation statistic (i.e. the ρ) was cross-map skill, which is known to readily detect dependence between X and Y in this system [18]. For the cross-map skill calculation, we used Y to estimate X, which corresponds to a scenario where the data analyst hypothesizes that X influences Y. Cross-map skill requires two parameters, the embedding dimension and the embedding lag, and these were set to 2 and 1 respectively following prior works that used the logistic map for benchmarking [18, 96].

References

[1] P. Spirtes and K. Zhang, "Causal discovery and inference: concepts and recent methodological advances," in *Applied informatics*, vol. 3, pp. 1–28, SpringerOpen, 2016.

[2] A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. N. Balasubramanian, "Neural network attributions: A causal perspective," in *International Conference on Machine Learning*, pp. 981–990, PMLR, 2019.

[3] C. Glymour, K. Zhang, and P. Spirtes, "Review of causal discovery methods based on graphical models," *Frontiers in genetics*, vol. 10, p. 524, 2019.

[4] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science Advances*, vol. 5, no. 11, p. eaau4996, 2019.

[5] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, *et al.*, "Inferring causation from time series in earth system sciences," *Nature communications*, vol. 10, no. 1, pp. 1–13, 2019.

[6] R. Sanchez-Romero, J. D. Ramsey, K. Zhang, M. R. Glymour, B. Huang, and C. Glymour, "Estimating feedforward and feedback effective connections from fmri time series: Assessments of statistical methods," *Network Neuroscience*, vol. 3, no. 2, pp. 274–306, 2019.

[7] S. Leng, H. Ma, J. Kurths, Y.-C. Lai, W. Lin, K. Aihara, and L. Chen, "Partial cross mapping eliminates indirect causal influences," *Nature communications*, vol. 11, no. 1, pp. 1–9, 2020.

[8] B. C. Daniels and I. Nemenman, "Automated adaptive inference of phenomenological dynamical models," *Nature communications*, vol. 6, no. 1, pp. 1–8, 2015.

[9] N. M. Mangan, S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Inferring biological networks by sparse identification of nonlinear dynamics," *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, vol. 2, no. 1, pp. 52–63, 2016.

[10] R. R. Stein, V. Bucci, N. C. Toussaint, C. G. Buffie, G. Ratsch, E. G. Pamer, C. Sander, and J. B. Xavier, "Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota," *PLoS computational biology*, vol. 9, no. 12, p. e1003388, 2013.

[11] C. K. Fisher and P. Mehta, "Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression," *PLOS ONE*, vol. 9, pp. 1–10, 07 2014.

[12] V. Bucci, B. Tzen, N. Li, M. Simmons, T. Tanoue, E. Bogart, L. Deng, V. Yeliseyev, M. L. Delaney, Q. Liu, B. Olle, R. R. Stein, K. Honda, L. Bry, and G. K. Gerber, "Mdsine: Microbial dynamical systems inference engine for microbiome time-series analyses," *Genome Biology*, vol. 17, p. 121, Jun 2016.

[13] S. H. Levine, "Competitive interactions in ecosystems," *The American Naturalist*, vol. 110, no. 976, pp. 903–910, 1976.

[14] J. T. Wootton, "Indirect effects in complex ecosystems: recent progress and future challenges," *Journal of Sea Research*, vol. 48, no. 2, pp. 157–172, 2002.

[15] B. Momeni, L. Xie, and W. Shou, "Lotka-volterra pairwise modeling fails to capture diverse pairwise microbial interactions," *Elife*, vol. 6, p. e25051, 2017.

[16] A. R. Coenen, S. K. Hu, E. Luo, D. Muratore, and J. S. Weitz, "A primer for microbiome time-series analysis," *Frontiers in Genetics*, vol. 11, 2020.

[17] C. W. Granger, "Testing for causality: a personal viewpoint," *Journal of Economic Dynamics and control*, vol. 2, pp. 329–352, 1980.

[18] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems," *Science*, vol. 338, no. 6106, pp. 496–500, 2012.

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

[20] L. Barnett and A. K. Seth, "The mvgc multivariate granger causality toolbox: a new approach to granger-causal inference," *Journal of neuroscience methods*, vol. 223, pp. 50–68, 2014.

[21] J. Woodward, "Causation and manipulability," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, winter 2016 ed., 2016.

[22] J. Pearl, *Causality*. Cambridge university press, 2000.

[23] D. Harnack, E. Laminski, M. Schünemann, and K. R. Pawelzik, "Topological causality in dynamical systems," *Physical review letters*, vol. 119, no. 9, p. 098301, 2017.

[24] M. Luo, H. Kantz, N.-C. Lau, W. Huang, and Y. Zhou, "Questionable dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature," *Proceedings of the National Academy of Sciences*, vol. 112, no. 34, pp. E4638–E4639, 2015.

[25] E. B. Baskerville and S. Cobey, "Does influenza drive absolute humidity?," *Proceedings of the National Academy of Sciences*, vol. 114, no. 12, pp. E2270–E2271, 2017.

[26] L. Tiokhin and D. Hruschka, "No evidence that an ebola outbreak influenced voting preferences in the 2014 elections after controlling for time-series autocorrelation: A commentary on beall, hofer, and schaller (2016)," *Psychological science*, vol. 28, no. 9, pp. 1358–1360, 2017.

[27] M. Schaller, M. K. Hofer, and A. T. Beall, "Evidence that an ebola outbreak influenced voting preferences, even after controlling (mindfully) for autocorrelation: Reply to tiokhin and hruschka (2017)," *Psychological science*, vol. 28, no. 9, pp. 1361–1363, 2017.

[28] L. Barnett, A. B. Barrett, and A. K. Seth, "Misunderstandings regarding the application of granger causality in neuroscience," *Proceedings of the National Academy of Sciences*, vol. 115, no. 29, pp. E6676–E6677, 2018.

[29] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

[30] J. Runge, "Causal network reconstruction from time series: From theoretical assumptions to practical estimation," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 7, p. 075310, 2018.

[31] C. Hitchcock, "Causal Models," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, summer 2020 ed., 2020.

[32] D. M. Hausman and J. Woodward, "Independence, invariance and the causal markov condition," *The British journal for the philosophy of science*, vol. 50, no. 4, pp. 521–583, 1999.

[33] A. R. Coenen and J. S. Weitz, "Limitations of correlation-based inference in complex virus-microbe communities," *mSystems*, vol. 3, no. 4, pp. e00084–18, 2018.

[34] A. Carr, C. Diener, N. S. Baliga, and S. M. Gibbons, "Use and abuse of correlation analyses in microbial ecology," *The ISME journal*, p. 1, 2019.

[35] K. Mainali, S. Bewick, B. Vecchio-Pagan, D. Karig, and W. F. Fagan, "Detecting interaction networks in the human microbiome with conditional granger causality," *PLoS computational biology*, vol. 15, no. 5, p. e1007037, 2019.

[36] S. Weiss, W. Van Treuren, C. Lozupone, K. Faust, J. Friedman, Y. Deng, L. C. Xia, Z. Z. Xu, L. Ursell, E. J. Alm, *et al.*, "Correlation detection strategies in microbial data sets vary widely in sensitivity and precision," *The ISME journal*, vol. 10, no. 7, pp. 1669–1681, 2016.

[37] Q. Ruan, D. Dutta, M. S. Schwalbach, J. A. Steele, J. A. Fuhrman, and F. Sun, "Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors," *Bioinformatics*, vol. 22, no. 20, pp. 2532–2538, 2006.

[38] C. Hitchcock and M. Rédei, "Reichenbach's common cause principle," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, spring 2020 ed., 2020.

[39] R. G. Moulder, S. M. Boker, F. Ramseyer, and W. Tschacher, "Determining synchrony between behavioral time series: An application of surrogate data generation for establishing falsifiable null-hypotheses.," *Psychological methods*, vol. 23, no. 4, p. 757, 2018.

[40] S. Afyouni, S. M. Smith, and T. E. Nichols, "Effective degrees of freedom of the pearson's correlation coefficient under autocorrelation," *NeuroImage*, vol. 199, pp. 609–625, 2019.

[41] B. J. Pyper and R. M. Peterman, "Comparison of methods to account for autocorrelation in correlation analyses of fish data," *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 55, no. 9, pp. 2127–2140, 1998.

[42] W. Ebisuzaki, "A method to estimate the statistical significance of a correlation when the data are serially correlated," *Journal of Climate*, vol. 10, no. 9, pp. 2147–2153, 1997.

[43] G. Lancaster, D. Iatsenko, A. Pidde, V. Ticcinelli, and A. Stefanovska, "Surrogate data for hypothesis testing of physical systems," *Physics Reports*, vol. 748, pp. 1–60, 2018.

[44] A. Eiler, F. Heinrich, and S. Bertilsson, "Coherent dynamics and association networks among lake bacterioplankton taxa," *The ISME journal*, vol. 6, no. 2, pp. 330–342, 2012.

[45] A. Shade, P. S. McManus, and J. Handelsman, "Unexpected diversity during community succession in the apple flower microbiome," *MBio*, vol. 4, no. 2, 2013.

[46] V. Cyriaque, A. Géron, G. Billon, J. Nesme, J. Werner, D. C. Gillan, S. J. Sørensen, and R. Wattiez, "Metal-induced bacterial interactions promote diversity in river-sediment microbiomes," *FEMS Microbiology Ecology*, vol. 96, no. 6, p. fiaa076, 2020.

[47] T. Schreiber and A. Schmitz, "Surrogate time series," *Physica D: Nonlinear Phenomena*, vol. 142, no. 3-4, pp. 346–382, 2000.

[48] R. G. Andrzejak, A. Kraskov, H. Stögbauer, F. Mormann, and T. Kreuz, "Bivariate surrogate techniques: necessity, strengths, and caveats," *Physical review E*, vol. 68, no. 6, p. 066202, 2003.

[49] K.-S. Chan, "On the validity of the method of surrogate data," *Fields Inst. Commun*, vol. 11, pp. 77–97, 1997.

[50] A. Papana, C. Kyrtsou, D. Kugiumtzis, and C. Diks, "Assessment of resampling methods for causality testing: A note on the us inflation behavior," *PloS one*, vol. 12, no. 7, p. e0180852, 2017.

[51] M. Thiel, M. C. Romano, J. Kurths, M. Rolfs, and R. Kliegl, "Twin surrogates to test for complex synchronisation," *EPL (Europhysics Letters)*, vol. 75, no. 4, p. 535, 2006.

[52] C. Heinze-Deml, M. H. Maathuis, and N. Meinshausen, "Causal structure learning," 2017.

[53] S. M. Gibbons, S. M. Kearney, C. S. Smillie, and E. J. Alm, "Two dynamic regimes in the human gut microbiome," *PLoS computational biology*, vol. 13, no. 2, p. e1005364, 2017.

[54] D. Ai, X. Li, G. Liu, X. Liang, and L. C. Xia, "Constructing the microbial association network from large-scale time series data using granger causality," *Genes*, vol. 10, no. 3, p. 216, 2019.

[55] F. Barraquand, C. Picoche, M. Detto, and F. Hartig, "Inferring species interactions using granger causality and convergent cross mapping," *Theoretical Ecology*, vol. 14, no. 1, pp. 87–105, 2021.

[56] C. Diks and V. Panchenko, "A new statistic and practical guidelines for nonparametric granger causality testing," *Journal of Economic Dynamics and Control*, vol. 30, no. 9-10, pp. 1647–1669, 2006.

[57] S. D. Bekiros and C. G. Diks, "The nonlinear dynamic relationship of exchange rates: Parametric and nonparametric causality testing," *Journal of macroeconomics*, vol. 30, no. 4, pp. 1641–1650, 2008.

[58] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, "Transfer entropy—a model-free measure of effective connectivity for the neurosciences," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 45–67, 2011.

[59] F. Roux, M. Wibral, W. Singer, J. Aru, and P. J. Uhlhaas, "The phase of thalamic alpha activity modulates cortical gamma-band activity: evidence from resting-state meg recordings," *Journal of Neuroscience*, vol. 33, no. 45, pp. 17827–17835, 2013.

[60] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, B. Schölkopf, *et al.*, "Quantifying causal influences," *The Annals of Statistics*, vol. 41, no. 5, pp. 2324–2358, 2013.

[61] M. Gong, K. Zhang, B. Schoelkopf, D. Tao, and P. Geiger, "Discovering temporal causal relations from subsampled data," in *International Conference on Machine Learning*, pp. 1898–1906, PMLR, 2015.

[62] A. Hyttinen, S. Plis, M. Järvisalo, F. Eberhardt, and D. Danks, "Causal discovery from subsampled time series data by constraint optimization," in *Conference on Probabilistic Graphical Models*, pp. 216–227, PMLR, 2016.

[63] M. Gong, K. Zhang, B. Schölkopf, C. Glymour, and D. Tao, "Causal discovery from temporally aggregated time series," in *Uncertainty in artificial intelligence: proceedings of the... conference. Conference on Uncertainty in Artificial Intelligence*, vol. 2017, NIH Public Access, 2017.

[64] P. Newbold, "Feedback induced by measurement errors," *International Economic Review*, pp. 787–791, 1978.

[65] H. Nalatore, M. Ding, and G. Rangarajan, "Mitigating the effects of measurement noise on granger causality," *Physical Review E*, vol. 75, no. 3, p. 031123, 2007.

[66] N. Ay and D. Polani, "Information flows in causal networks," *Advances in complex systems*, vol. 11, no. 01, pp. 17–41, 2008.

[67] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[68] H. Y. Toda and P. C. Phillips, "The spurious effect of unit roots on vector autoregressions: an analytical study," *Journal of Econometrics*, vol. 59, no. 3, pp. 229–255, 1993.

[69] L. E. Ohanian, "The spurious effects of unit roots on vector autoregressions: A monte carlo study," *Journal of Econometrics*, vol. 39, no. 3, pp. 251–266, 1988.

[70] Z. He and K. Maekawa, "On spurious granger causality," *Economics Letters*, vol. 73, no. 3, pp. 307–313, 2001.

[71] S. Li, Y. Xiao, D. Zhou, and D. Cai, "Causal inference in nonlinear systems: Granger causality versus time-delayed mutual information," *Physical Review E*, vol. 97, no. 5, p. 052216, 2018.

[72] E. L. Feige and D. K. Pearce, "The casual causal relationship between money and income: Some caveats for time series analysis," *The Review of Economics and Statistics*, pp. 521–533, 1979.

[73] A. Papana, D. Kugiumtzis, and P. G. Larsson, "Detection of direct causal effects and application to epileptic electroencephalogram analysis," *International Journal of Bifurcation and Chaos*, vol. 22, no. 09, p. 1250222, 2012.

[74] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.

[75] T. Cover and J. Thomas, *Elements of Information Theory*. Elements of Information Theory, Wiley, 2006.

[76] A. Montalto, L. Faes, and D. Marinazzo, "Mute: a matlab toolbox to compare established and novel estimators of the multivariate transfer entropy," *PloS one*, vol. 9, no. 10, p. e109462, 2014.

[77] D. P. Shorten, R. E. Spinney, and J. T. Lizier, "Estimating transfer entropy in continuous time between neural spike trains or other event-based data," *PLoS computational biology*, vol. 17, no. 4, p. e1008054, 2021.

[78] J. Runge, "Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information," in *International Conference on Artificial Intelligence and Statistics*, pp. 938–947, PMLR, 2018.

[79] S. Behrendt, T. Dimpfl, F. J. Peter, and D. J. Zimmermann, "Rtransferentropy—quantifying information flow between different time series using effective transfer entropy," *SoftwareX*, vol. 10, p. 100265, 2019.

[80] P. Wollstadt, J. Lizier, R. Vicente, C. Finn, M. Martinez-Zarzuela, P. Mediano, L. Novelli, and M. Wibral, "Idtxl: The information dynamics toolkit xl: a python package for the efficient analysis of multivariate information dynamics in networks," *Journal of Open Source Software*, vol. 4, no. 34, p. 1081, 2019.

[81] A. Hyvärinen, K. Zhang, S. Shimizu, and P. O. Hoyer, "Estimation of a structural vector autoregression model using non-gaussianity.," *Journal of Machine Learning Research*, vol. 11, no. 5, 2010.

[82] M. Casdagli, S. Eubank, J. D. Farmer, and J. Gibson, "State space reconstruction in the presence of noise," *Physica D: Nonlinear Phenomena*, vol. 51, no. 1-3, pp. 52–98, 1991.

[83] D. Kugiumtzis, B. Lillekjendlie, and N. D. Christophersen, "Chaotic time series: Part 1: estimation of some invariant properties in state space," *Modeling, identification and control*, vol. 15, no. 4, pp. 205–224, 1994.

[84] T. Asefa, M. Kemblowski, U. Lall, and G. Urroz, "Support vector machines for nonlinear state space reconstruction: Application to the great salt lake time series," *Water resources research*, vol. 41, no. 12, 2005.

[85] B. Cummins, T. Gedeon, and K. Spendlove, "On the efficacy of state space reconstruction methods in determining causality," *SIAM Journal on Applied Dynamical Systems*, vol. 14, no. 1, pp. 335–381, 2015.

[86] E. Brookshire and T. Weaver, "Long-term decline in grassland productivity driven by increasing dryness," *Nature communications*, vol. 6, no. 1, pp. 1–7, 2015.

[87] K. L. Cramer, A. O'Dea, T. R. Clark, J.-x. Zhao, and R. D. Norris, "Prehistorical and historical declines in caribbean coral reef accretion rates driven by loss of parrotfish," *Nature communications*, vol. 8, no. 1, pp. 1–8, 2017.

[88] B. Hannisdal, K. A. Haaga, T. Reitan, D. Diego, and L. H. Liow, "Common species link global ecosystems to climate change: dynamical evidence in the planktonic fossil record," *Proceedings of the Royal Society B: Biological Sciences*, vol. 284, no. 1858, p. 20170722, 2017.

[89] S.-i. S. Matsuzaki, K. Suzuki, T. Kadoya, M. Nakagawa, and N. Takamura, "Bottom-up linkages between primary production, zooplankton, and fish in a shallow, hypereutrophic lake," *Ecology*, vol. 99, no. 9, pp. 2025–2036, 2018.

[90] M. Wang, C. Yoshimura, A. Allam, F. Kimura, and T. Honma, "Causality analysis and prediction of 2-methylisoborneol production in a reservoir using empirical dynamic modeling," *Water research*, vol. 163, p. 114864, 2019.

[91] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical systems and turbulence, Warwick 1980*, pp. 366–381, Springer, 1981.

[92] S. Cobey and E. B. Baskerville, "Limits to causal inference with state-space reconstruction for infectious disease," *PloS one*, vol. 11, no. 12, p. e0169050, 2016.

[93] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of statistical physics*, vol. 65, no. 3-4, pp. 579–616, 1991.

[94] H. Ma, K. Aihara, and L. Chen, "Detecting causality from nonlinear dynamics with short-term time series," *Scientific reports*, vol. 4, p. 7464, 2014.

[95] A. T. Clark, H. Ye, F. Isbell, E. R. Deyle, J. Cowles, G. D. Tilman, and G. Sugihara, "Spatial convergent cross mapping to detect causal relationships from short time series," *Ecology*, vol. 96, no. 5, pp. 1174–1181, 2015.

[96] H. Ye, E. R. Deyle, L. J. Gilarranz, and G. Sugihara, "Distinguishing time-delayed causal interactions using convergent cross mapping," *Scientific reports*, vol. 5, p. 14750, 2015.

[97] D. Mønster, R. Fusaroli, K. Tylén, A. Roepstorff, and J. F. Sherson, "Causal inference from noisy time-series data-testing the convergent cross-mapping algorithm in the presence of noise and external influence," *Future Generation Computer Systems*, vol. 73, pp. 52–62, 2017.

[98] E. R. Deyle and G. Sugihara, "Generalized theorems for nonlinear state space reconstruction," *PLOS ONE*, vol. 6, pp. 1–8, 03 2011.

[99] Y. Wang, J. Yang, Y. Chen, P. De Maeyer, Z. Li, and W. Duan, "Detecting the causal effect of soil moisture on precipitation using convergent cross mapping," *Scientific reports*, vol. 8, no. 1, pp. 1–8, 2018.

[100] A. Hastings, C. L. Hom, S. Ellner, P. Turchin, and H. C. J. Godfray, "Chaos in ecology: is mother nature a strange attractor?," *Annual review of ecology and systematics*, vol. 24, no. 1, pp. 1–33, 1993.

[101] G. Sugihara and R. M. May, "Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series," *Nature*, vol. 344, no. 6268, pp. 734–741, 1990.

[102] B. Lusch, P. D. Maia, and J. N. Kutz, "Inferring connectivity in networked dynamical systems: Challenges using granger causality," *Physical Review E*, vol. 94, no. 3, p. 032220, 2016.

[103] C.-W. Chang, M. Ushio, and C.-h. Hsieh, "Empirical dynamic modeling for beginners," *Ecological Research*, vol. 32, no. 6, pp. 785–796, 2017.

[104] P. A. Stokes and P. L. Purdon, "A study of problems encountered in granger causality analysis from a neuroscience perspective," *Proceedings of the National Academy of Sciences*, vol. 114, no. 34, pp. E7063–E7072, 2017.

[105] S. B. Munch, A. Brias, G. Sugihara, and T. L. Rogers, "Frequently asked questions about nonlinear dynamics and empirical dynamic modelling," *ICES Journal of Marine Science*, vol. 77, no. 4, pp. 1463–1479, 2020.

[106] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.

[107] D. Eaton and K. Murphy, "Exact bayesian structure learning from uncertain interventions," in *Artificial intelligence and statistics*, pp. 107–114, PMLR, 2007.

[108] D. Rothenhäusler, C. Heinze, J. Peters, and N. Meinshausen, "Backshift: Learning causal cyclic graphs from unknown shift interventions," *arXiv preprint arXiv:1506.02494*, 2015.

[109] S. L. Brunton, B. W. Brunton, J. L. Proctor, E. Kaiser, and J. N. Kutz, "Chaos as an intermittently forced linear system," *Nature communications*, vol. 8, no. 1, pp. 1–9, 2017.

[110] L. Xie and W. Shou, "Steering ecological-evolutionary dynamics to improve artificial selection of microbial communities," *Nature communications*, vol. 12, no. 1, pp. 1–15, 2021.

[111] A. Koplenig and C. Müller-Spitzer, "Population size predicts lexical diversity, but so does the mean sea level–why it is important to correctly account for the structure of temporal data," *PloS one*, vol. 11, no. 3, p. e0150771, 2016.

[112] C. S. Rosenfeld, "Sex-dependent differences in voluntary physical activity," *Journal of neuroscience research*, vol. 95, no. 1-2, pp. 279–290, 2017.

[113] J. Zhang and P. Spirtes, "Detection of unfaithfulness and robust causal inference," *Minds and Machines*, vol. 18, no. 2, pp. 239–271, 2008.

[114] P. O. Hoyer, D. Janzing, J. M. Mooij, J. Peters, B. Schölkopf, *et al.*, "Nonlinear causal discovery with additive noise models.," in *NIPS*, vol. 21, pp. 689–696, Citeseer, 2008.

[115] J. Peters, J. Mooij, D. Janzing, and B. Schölkopf, "Identifiability of causal graphs using functional models," *arXiv preprint arXiv:1202.3757*, 2012.

[116] S. F. M. Hart, J. M. B. Pineda, C.-C. Chen, R. Green, and W. Shou, "Disentangling strictly self-serving mutations from win-win mutations in a mutualistic microbial community," *Elife*, vol. 8, p. e44812, 2019.

[117] S. F. M. Hart, C.-C. Chen, and W. Shou, "Pleiotropic mutations can rapidly evolve to directly benefit self and cooperative partner despite unfavorable conditions," *Elife*, vol. 10, p. e57838, 2021.

[118] Z. Jia, Y. Lin, Y. Liu, Z. Jiao, and J. Wang, "Refined nonuniform embedding for coupling detection in multivariate time series," *Physical Review E*, vol. 101, no. 6, p. 062113, 2020.

[119] C. T. Perretti, S. B. Munch, and G. Sugihara, "Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data," *Proceedings of the National Academy of Sciences*, vol. 110, no. 13, pp. 5253–5257, 2013.

[120] J. Huke, "Embedding nonlinear dynamical systems: A guide to takens' theorem," *MIMS EPrint*, 2006.

[121] G. U. Yule, "Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time-series," *Journal of the royal statistical society*, vol. 89, no. 1, pp. 1–63, 1926.

[122] C. Granger and P. Newbold, "Spurious regressions in econometrics," *Journal of Econometrics*, vol. 2, no. 2, pp. 111–120, 1974.

[123] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Application*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 1997.

[124] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*, vol. 3. Springer, 2008.

[125] W. Conover, "Distribution-free methods in statistics," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 1, no. 2, pp. 199–207, 2009.

[126] S. M. Weinstein, S. N. Vandekar, A. Adebimpe, T. M. Tapera, T. Robert-Fitzgerald, R. C. Gur, R. E. Gur, A. Raznahan, T. D. Satterthwaite, A. F. Alexander-Bloch, *et al.*, "A simple permutation-based test of intermodal correspondence," *Human brain mapping*, vol. 42, no. 16, pp. 5175–5187, 2021.

[127] A. E. Yuan and W. Shou, "Data-driven causal analysis of observational biological time series," *Elife*, vol. 11, p. e72518, 2022.

[128] P. C. Molenaar, "A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever," *Measurement*, vol. 2, no. 4, pp. 201–218, 2004.

[129] A. A. Tsonis, E. R. Deyle, R. M. May, G. Sugihara, K. Swanson, J. D. Verbeten, and G. Wang, "Dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature," *Proceedings of the National Academy of Sciences*, vol. 112, no. 11, pp. 3253–3256, 2015.

[130] E. H. Van Nes, M. Scheffer, V. Brovkin, T. M. Lenton, H. Ye, E. Deyle, and G. Sugihara, "Causal feedbacks in climate change," *Nature Climate Change*, vol. 5, no. 5, pp. 445–448, 2015.

[131] R. M. Warner, *Spectral analysis of time-series data*. Guilford Press, 1998.

[132] K. D. Harris, "A shift test for independence in generic time series," 2020.

[133] C. Diks and J. DeGoede, "A general nonparametric bootstrap test for granger causality," *Global analysis of dynamical systems*, pp. 391–403, 2001.

[134] M. C. Romano, M. Thiel, J. Kurths, K. Mergenthaler, and R. Engbert, "Hypothesis test for synchronization: twin surrogates revisited," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 19, no. 1, p. 015108, 2009.

[135] T. I. Netoff and S. J. Schiff, "Decreased neuronal synchronization during experimental seizures," *Journal of Neuroscience*, vol. 22, no. 16, pp. 7297–7307, 2002.

[136] L. Faes, A. Porta, and G. Nollo, "Mutual nonlinear prediction as a tool to evaluate coupling strength and directionality in bivariate time series: comparison among different strategies based on k nearest neighbors," *Physical Review E*, vol. 78, no. 2, p. 026201, 2008.

[137] I. Vlachos and D. Kugiumtzis, "Nonuniform state-space reconstruction and coupling detection," *Physical Review E*, vol. 82, no. 1, p. 016207, 2010.

[138] M. Bartlett, "Some aspects of the time-correlation problem in regard to tests of significance," *Journal of the Royal Statistical Society*, vol. 98, no. 3, pp. 536–543, 1935.

[139] A. E. Yuan and W. Shou, "An exactly valid and distribution-free statistical significance test for correlations between time series," *bioRxiv*, 2022.

[140] W. F. Stout, *Almost Sure Convergence*. Probability and mathematical statistics, Academic Press, 1974.

[141] G. Lindgren, *Stationary stochastic processes: theory and applications*. CRC Press, 2012.

[142] W. Greene, *Econometric Analysis*. Pearson, 2012.

[143] A. K. Ramdas, R. F. Barber, M. J. Wainwright, and M. I. Jordan, "A unified treatment of multiple testing with prior knowledge using the p-filter," *The Annals of Statistics*, vol. 47, no. 5, pp. 2790–2821, 2019.

[144] X. Chen, R. W. Doerge, and S. K. Sarkar, "A weighted fdr procedure under discrete and heterogeneous null distributions," *Biometrical Journal*, vol. 62, no. 6, pp. 1544–1563, 2020.

[145] R. FitzHugh, "Impulses and physiological states in theoretical models of nerve membrane," *Biophysical journal*, vol. 1, no. 6, pp. 445–466, 1961.

[146] J. Vano, J. Wildenberg, M. Anderson, J. Noel, and J. Sprott, "Chaos in low-dimensional lotka–volterra models of competition," *Nonlinearity*, vol. 19, no. 10, p. 2391, 2006.

[147] D. Tjøstheim, "Non-linear time series and markov chains," *Advances in applied probability*, vol. 22, no. 3, pp. 587–611, 1990.

[148] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical review E*, vol. 69, no. 6, p. 066138, 2004.

[149] D. N. Politis and J. P. Romano, "The stationary bootstrap," *Journal of the American Statistical association*, vol. 89, no. 428, pp. 1303–1313, 1994.

[150] P. Clifford, S. Richardson, and D. Hemon, "Assessing the significance of the correlation between two spatial processes," *Biometrics*, vol. 45, no. 1, pp. 123–134, 1989.

[151] O. M. Cliff, L. Novelli, B. D. Fulcher, J. M. Shine, and J. T. Lizier, "Assessing the significance of directed and multivariate measures of linear dependence between time series," *Phys. Rev. Research*, vol. 3, p. 013145, Feb 2021.

[152] R. M. May, "Simple mathematical models with very complicated dynamics," *The Theory of Chaotic Attractors*, pp. 85–93, 2004.

[153] E. W. Weisstein, "Sawtooth wave. From MathWorld—A Wolfram Web Resource." Last visited on Mar 16, 2022.

[154] J. Lucio, R. Valdés, and L. Rodríguez, "Improvements to surrogate data methods for nonstationary time series," *Physical Review E*, vol. 85, no. 5, p. 056202, 2012.

[155] J. Laskar, P. Robutel, F. Joutel, M. Gastineau, A. Correia, and B. Levrard, "A long-term numerical solution for the insolation quantities of the earth," *Astronomy & Astrophysics*, vol. 428, no. 1, pp. 261–285, 2004.

[156] M. Maslin, "Forty years of linking orbits to ice ages," *Nature*, vol. 540, no. 7632, pp. 208–209, 2016.

[157] J. D. Hays, J. Imbrie, N. J. Shackleton, *et al.*, "Variations in the earth's orbit: pacemaker of the ice ages," *science*, vol. 194, no. 4270, pp. 1121–1132, 1976.

[158] C. Lorius, J. Jouzel, C. Ritz, L. Merlivat, N. Barkov, Y. S. Korotkevich, and V. Kotlyakov, "A 150,000-year climatic record from antarctic ice," *Nature*, vol. 316, no. 6029, pp. 591–596, 1985.

[159] P. Huybers and C. Wunsch, "Obliquity pacing of the late pleistocene glacial terminations," *Nature*, vol. 434, no. 7032, pp. 491–494, 2005.

[160] P. Huybers, "Glacial variability over the last two million years: an extended depth-derived agemodel, continuous obliquity pacing, and the pleistocene progression," *Quaternary Science Reviews*, vol. 26, no. 1-2, pp. 37–55, 2007.

[161] D. C. Howell, *Statistical methods for psychology*. Cengage Learning, seventh ed., 2010.

[162] C. Huttenhower, D. Gevers, R. Knight, S. Abubucker, J. H. Badger, A. T. Chinwalla, H. H. Creasy, A. M. Earl, M. G. FitzGerald, R. S. Fulton, *et al.*, "Structure, function and diversity of the healthy human microbiome," *nature*, vol. 486, no. 7402, p. 207, 2012.

[163] J. G. Caporaso, C. L. Lauber, E. K. Costello, D. Berg-Lyons, A. Gonzalez, J. Stombaugh, D. Knights, P. Gajer, J. Ravel, N. Fierer, *et al.*, "Moving pictures of the human microbiome," *Genome biology*, vol. 12, no. 5, pp. 1–8, 2011.

[164] G. B. Gloor, J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue, "Microbiome datasets are compositional: and this is not optional," *Frontiers in microbiology*, vol. 8, p. 2224, 2017.

[165] A. Gonzalez, J. A. Navas-Molina, T. Kosciolek, D. McDonald, Y. Vázquez-Baeza, G. Ackermann, J. DeReus, S. Janssen, A. D. Swafford, S. B. Orchanian, *et al.*, "Qiita: rapid, web-enabled microbiome meta-analysis," *Nature methods*, vol. 15, no. 10, pp. 796–798, 2018.

[166] L. C. Xia, J. A. Steele, J. A. Cram, Z. G. Cardon, S. L. Simmons, J. J. Vallino, J. A. Fuhrman, and F. Sun, "Extended local similarity analysis (elsa) of microbial community and other time series data with replicates," in *BMC systems biology*, vol. 5, p. S15, BioMed Central, 2011.

[167] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal statistical society: series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[168] Y. Benjamini and D. Yekutieli, "The control of the false discovery rate in multiple testing under dependency," *Annals of statistics*, pp. 1165–1188, 2001.

[169] M. B. Orger and G. G. de Polavieja, "Zebrafish behavior: opportunities and challenges," *Annual review of neuroscience*, vol. 40, pp. 125–147, 2017.

[170] A. Attanasi, A. Cavagna, L. Del Castello, I. Giardina, S. Melillo, L. Parisi, O. Pohl, B. Rossaro, E. Shen, E. Silvestri, *et al.*, "Collective behaviour without collective order in wild swarms of midges," *PLoS computational biology*, vol. 10, no. 7, p. e1003697, 2014.

[171] K. van der Vaart, M. Sinhuber, A. M. Reynolds, and N. T. Ouellette, "Mechanical spectroscopy of insect swarms," *Science advances*, vol. 5, no. 7, p. eaaw9305, 2019.

[172] A. Reynolds, "Langevin dynamics encapsulate the microscopic and emergent macroscopic properties of midge swarms," *Journal of The Royal Society Interface*, vol. 15, no. 138, p. 20170806, 2018.

[173] A. K. Zienkiewicz, F. Ladu, D. A. Barton, M. Porfiri, and M. Di Bernardo, "Data-driven modelling of social forces and collective behaviour in zebrafish," *Journal of Theoretical Biology*, vol. 443, pp. 39–51, 2018.

[174] F. J. Heras, F. Romero-Ferrero, R. C. Hinz, and G. G. de Polavieja, "Deep attention networks reveal the rules of collective motion in zebrafish," *PLoS computational biology*, vol. 15, no. 9, p. e1007354, 2019.

[175] F. Romero-Ferrero, M. G. Bergomi, R. C. Hinz, F. J. Heras, and G. G. de Polavieja, "Idtracker. ai: tracking all individuals in small or large collectives of unmarked animals," *Nature methods*, vol. 16, no. 2, pp. 179–182, 2019.

[176] N. Miller and R. Gerlai, "From schooling to shoaling: patterns of collective motion in zebrafish (danio rerio)," *PloS one*, vol. 7, no. 11, p. e48865, 2012.

[177] R. Q. Quiroga, A. Kraskov, T. Kreuz, and P. Grassberger, "Performance of different synchronization measures in real data: a case study on electroencephalographic signals," *Physical Review E*, vol. 65, no. 4, p. 041903, 2002.

[178] M. Ushio, C.-h. Hsieh, R. Masuda, E. R. Deyle, H. Ye, C.-W. Chang, G. Sugihara, and M. Kondoh, "Fluctuating interaction network and time-varying stability of a natural fish community," *Nature*, vol. 554, no. 7692, pp. 360–363, 2018.

[179] D. A. Dickey and W. A. Fuller, "Distribution of the estimators for autoregressive time series with a unit root," *Journal of the American statistical association*, vol. 74, no. 366a, pp. 427–431, 1979.

[180] A. Witt, J. Kurths, and A. Pikovsky, "Testing stationarity in time series," *physical Review E*, vol. 58, no. 2, p. 1800, 1998.

[181] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin, "Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?," *Journal of econometrics*, vol. 54, no. 1-3, pp. 159–178, 1992.

[182] R. Davidson, J. G. MacKinnon, *et al.*, *Econometric theory and methods*, vol. 5. Oxford University Press New York, 2004.

[183] N. M. Odogwu, C. A. Onebunne, J. Chen, F. A. Ayeni, M. R. Walther-Antonio, O. O. Olayemi, N. Chia, and A. O. Omigbodun, "Lactobacillus crispatus thrives in pregnancy hormonal milieu in a nigerian patient cohort," *Scientific reports*, vol. 11, no. 1, pp. 1–19, 2021.

[184] D. A. Jones and D. R. Cox, "Nonlinear autoregressive processes," *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, vol. 360, no. 1700, pp. 71–95, 1978.

[185] R. D. Shah and J. Peters, "The hardness of conditional independence testing and the generalised covariance measure," *The Annals of Statistics*, vol. 48, no. 3, pp. 1514–1538, 2020.

[186] E. Candes, Y. Fan, L. Janson, and J. Lv, "Panning for gold:?model-x?knockoffs for high dimensional controlled variable selection," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 80, no. 3, pp. 551–577, 2018.

[187] M. Sesia, C. Sabatti, and E. J. Candès, "Gene hunting with hidden markov model knockoffs," *Biometrika*, vol. 106, no. 1, pp. 1–18, 2019.

[188] J. F. Donges, J. Heitzig, B. Beronov, M. Wiedermann, J. Runge, Q. Y. Feng, L. Tupikina, V. Stolbova, R. V. Donner, N. Marwan, *et al.*, "Unified functional network and nonlinear time series analysis for complex systems science: The pyunicorn package," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 25, no. 11, p. 113101, 2015.

[189] G. Tauzin, U. Lupo, L. Tunstall, J. B. Pérez, M. Caorsi, A. Medina-Mardones, A. Dassatti, and K. Hess, "giotto-tda: A topological data analysis toolkit for machine learning and data exploration," 2020.

[190] M. B. Kennel, R. Brown, and H. D. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Physical review A*, vol. 45, no. 6, p. 3403, 1992.

[191] R. G. Kope and L. W. Botsford, "Determination of factors affecting recruitment of chinook salmon oncorhynchus tshawytscha in central california," *Fishery Bulletin*, vol. 88, no. 2, p. I990.

[192] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[193] S. K. Lam, A. Pitrou, and S. Seibert, "Numba: A llvm-based python jit compiler," in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pp. 1–6, 2015.

[194] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, "Measuring and testing dependence by correlation of distances," *The annals of statistics*, vol. 35, no. 6, pp. 2769–2794, 2007.

[195] J. S. Rosenthal, *First Look At Rigorous Probability Theory, A*. World Scientific Publishing Company, 2006.

[196] H. White, *Asymptotic theory for econometricians*. Academic press, 1984.

[197] M. Haenggi, *Stochastic Geometry for Wireless Networks*. Cambridge University Press, 2012.

[198] D. L. Cohn, *Measure theory*. Springer, 2013.

[199] A. M. Petrock, D. L. Donnelly, and M. L. Rosenberg, "Quantifying cardio-pulmonary correlations using the cross-wavelet transform: Validating a correlative method," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 2940–2943, IEEE, 2008.

[200] A. Lasota and M. C. Mackey, *Chaos, fractals, and noise: stochastic aspects of dynamics*, vol. 97. Springer Science & Business Media, 2013.

[201] B. Gärtner, "Fast and robust smallest enclosing balls," in *European symposium on algorithms*, pp. 325–338, Springer, 1999.

[202] E. W. Weisstein, "Rotation matrix. From MathWorld—A Wolfram Web Resource." Last visited on May 26, 2022.

[203] S. M. Ross, *A first course in probability*. Pearson, 2014.

[204] K. H. Chan, J. C. Hayya, and J. K. Ord, "A note on trend removal methods: The case of polynomial regression versus variate differencing," *Econometrica: Journal of the Econometric Society*, pp. 737–744, 1977.

[205] H. Isliker and J. Kurths, "A test for stationarity: finding parts in time series apt for correlation dimension estimates," *International Journal of Bifurcation and Chaos*, vol. 3, no. 06, pp. 1573–1579, 1993.

[206] D. Guarin, A. Orozco, and E. Delgado, "A new surrogate data method for nonstationary time series," 2010.

[207] W. Greene, *Econometric Analysis*. Pearson, 2012.

[208] D. Iatsenko, A. Bernjak, T. Stankovski, Y. Shiogai, P. J. Owen-Lynch, P. Clarkson, P. V. McClintock, and A. Stefanovska, "Evolution of cardiorespiratory interactions with age," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1997, p. 20110622, 2013.

[209] P. Bacchetti, S. G. Deeks, and J. M. McCune, "Breaking free of sample size dogma to perform innovative translational research," *Science translational medicine*, vol. 3, no. 87, pp. 87ps24–87ps24, 2011.

[210] H. V. Vesterinen, K. Egan, A. Deister, P. Schlattmann, M. R. Macleod, and U. Dirnagl, "Systematic survey of the design, statistical analysis, and reporting of studies published in the 2008 volume of the journal of cerebral blood flow and metabolism," *Journal of Cerebral Blood Flow & Metabolism*, vol. 31, no. 4, pp. 1064–1072, 2011.

[211] L. A. David, A. C. Materna, J. Friedman, M. I. Campos-Baptista, M. C. Blackburn, A. Perrotta, S. E. Erdman, and E. J. Alm, "Host lifestyle affects human microbiota on daily timescales," *Genome biology*, vol. 15, no. 7, p. R89, 2014.

[212] C. S. Greene, L. X. Garmire, J. A. Gilbert, M. D. Ritchie, and L. E. Hunter, "Celebrating parasites," *Nature genetics*, vol. 49, no. 4, pp. 483–484, 2017.

[213] D. C. Martin, V. Buffington, and J. Becker, "Randomization test of paired data: Application to evoked responses," *Psychophysiology*, vol. 18, no. 5, pp. 524–528, 1981.

[214] M. D. Ernst, "Permutation methods: a basis for exact inference," *Statistical Science*, pp. 676–685, 2004.

[215] V. Buffington, D. C. Martin, and J. Becker, "Ver similarity between alcoholic probands and their first-degree relatives," *Psychophysiology*, vol. 18, no. 5, pp. 529–533, 1981.

[216] M. Albert, Y. Bouret, M. Fromont, and P. Reynaud-Bouret, "Surrogate data methods based on a shuffling of the trials for synchrony detection: the centering issue," *Neural Computation*, vol. 28, no. 11, pp. 2352–2392, 2016.

[217] W. D. Koenig, "Spatial autocorrelation in california land birds," *Conservation Biology*, vol. 12, no. 3, pp. 612–620, 1998.

[218] E. Shaw, "Schooling fishes," *American Scientist*, vol. 66, no. 2, pp. 166–175, 1978.

[219] S. V. Viscido, J. K. Parrish, and D. Grünbaum, "Individual behavior and emergent properties of fish schools: a comparison of observation and theory," *Marine Ecology Progress Series*, vol. 273, pp. 239–249, 2004.

[220] U. Lopez, J. Gautrais, I. D. Couzin, and G. Theraulaz, "From behavioural analyses to models of collective motion in fish schools," *Interface focus*, vol. 2, no. 6, pp. 693–707, 2012.

[221] P. Berens, "Circstat: a matlab toolbox for circular statistics," *Journal of statistical software*, vol. 31, pp. 1–21, 2009.

[222] S. R. Jammalamadaka and A. Sengupta, *Topics in circular statistics*, vol. 5. world scientific, 2001.

[223] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[224] R. U. Kulkarni, C. L. Wang, and C. R. Bertozzi, "Analyzing nested experimental designs-a user-friendly resampling method to determine experimental significance," *PLoS computational biology*, vol. 18, no. 5, p. e1010061, 2022.

[225] A. F. Hayes, "Permutation test is not distribution-free: Testing h_0: ρ= 0.," *Psychological Methods*, vol. 1, no. 2, p. 184, 1996.

[226] T. P. Robitaille, E. J. Tollerud, P. Greenfield, M. Droettboom, E. Bray, T. Aldcroft, M. Davis, A. Ginsburg, A. M. Price-Whelan, W. E. Kerzendorf, *et al.*, "Astropy: A community python package for astronomy," *Astronomy & Astrophysics*, vol. 558, p. A33, 2013.

[227] J. Hemerik and J. Goeman, "Exact testing with random permutations," *Test*, vol. 27, no. 4, pp. 811–825, 2018.