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

Alex E. Yuan[1,2] and Wenying Shou[3]

March 13, 2023

1 Molecular and Cellular Biology PhD program, University of Washington, Seattle, Washington, United States; 2 Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States; 3 Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

alexericyuan@gmail.com; wenying.shou@gmail.com

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.

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 [1]. 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 [2, 3, 4]. 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 [5, 6].) 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) [7, 8]. 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 [9, 10]. 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 [9]; 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 [11]. 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 [12].

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" [13, 14].

The inter-subject surrogate approach can be used to test for correlations in each trial separately [15]. 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 $\rho(X_1, Y_1)$ [16, 17]. 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 [18], and single-digit replicate numbers are common in biomedical research [19]. For instance, two influential longitudinal human gut microbiome sampling studies relied on cohorts of two subjects each [20, 21]. Additionally, secondary analysis of public data from earlier studies can be a resource-efficient mode of research [22], 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 1A) using the data from all replicates [23, 24]. 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 1A). 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 [25, 23, 26]. It has also been used to detect correlations between variables measured in brain images, which can have similar nonstationarity challenges as time series [27]. 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 4 in Appendix 1).

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.

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 2, we restate this as Lemma 7.

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) [28]. 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 1A). 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 1B). 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,

$$p = \min(p_{perm}, p_{match}). \tag{1}$$

We will refer to this as the permute-match test (Fig 1C). 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 9 in Appendix 2.

Interestingly, $1/n^n$ is the lowest p-value that one may possibly achieve with a valid test under this null hypothesis. That is, if some new test with exactly the same null hypothesis as the permute-match test (with no additional assumptions) ever manages to produce a p-value below $1/n^n$, we would know immediately that this new test is not valid. This fact is proved as Theorem 10 in Appendix 2.

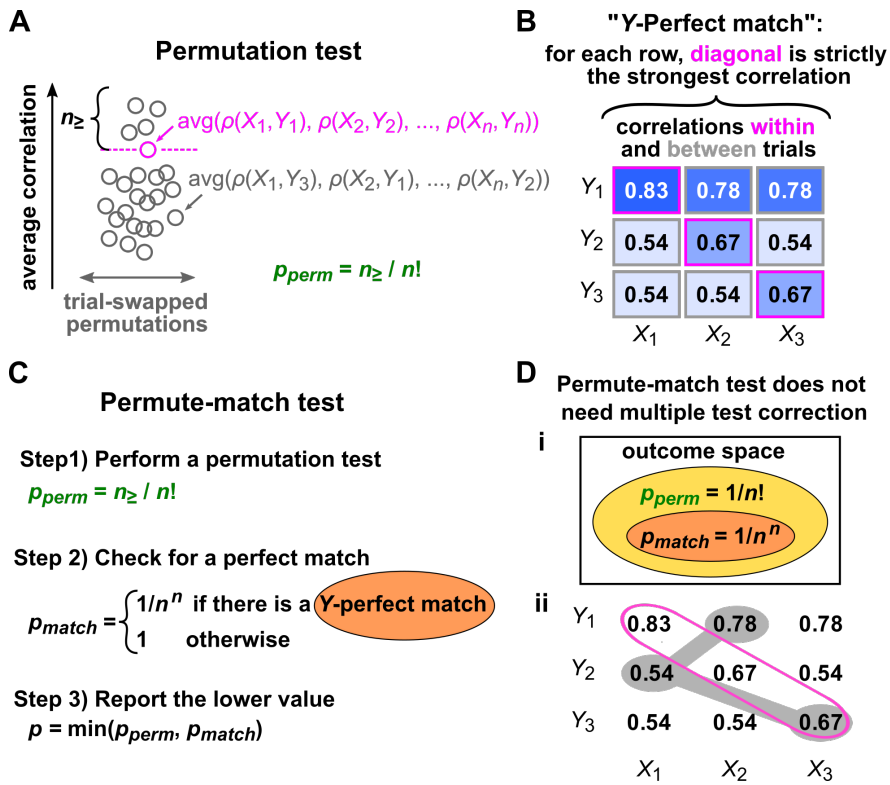

Figure 1: The permutation test, perfect match concept, and permute-match test. (**A**) The permutation test. Circles indicate the average correlation obtained with (grey) or without (pink) trial swapping. Here, $n_{\geq}$ is the number of average correlations (including the original one) that are equal 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 1C) is an "upgrade" to the permutation test (Fig 1A) 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. 1 (see Appendix 2 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 1D). 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. 1.

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 4 in Appendix 1). 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 2A). We use the Pearson correlation coefficient as our correlation function ρ. Fig 2B 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 2B, 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 2B, 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 3, 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.

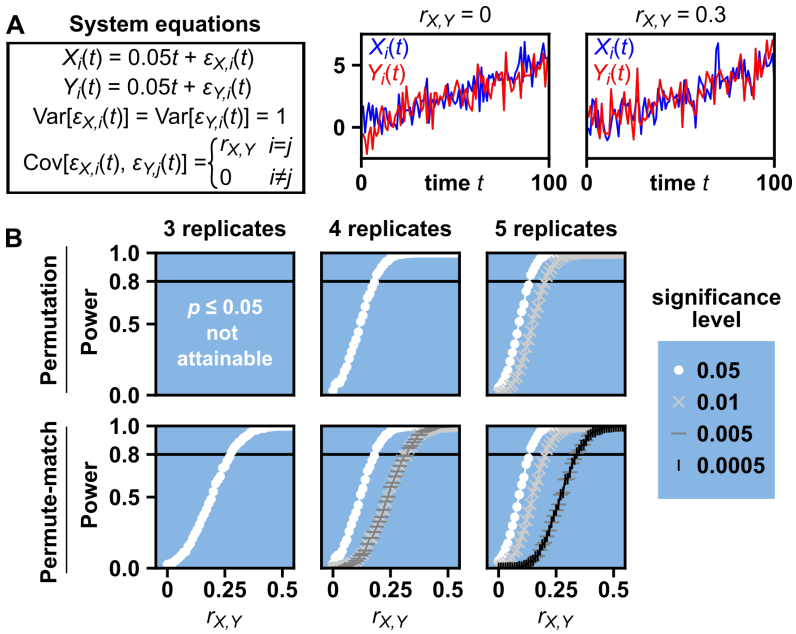

Figure 2: Statistical power of the permute-match test and permutation test in a simple nonstationary system. (**A**) System equations and example dynamics. The processes X and Y are given by a linear trend with additive noise. The noise terms $\epsilon_{X,i}(t)$ and $\epsilon_{Y,i}(t)$ are drawn from 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 1B). 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 2, 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 3A). 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 3B). 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 3C), 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 3D).

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.

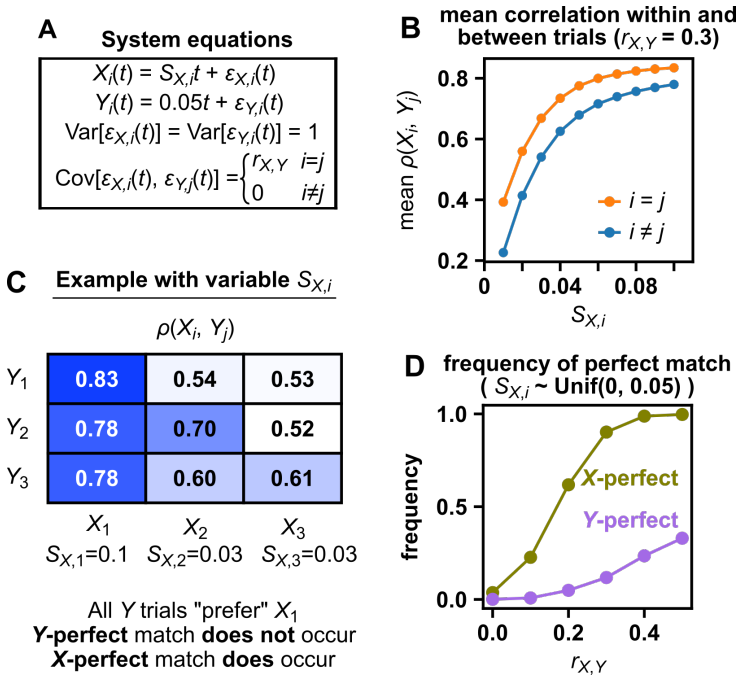

Figure 3: The probabilities of obtaining an X-perfect match and Y-perfect match may differ. (**A**) System equations. The processes X and Y are given by a linear trend with additive noise. The noise terms $\epsilon_X(t)$ and $\epsilon_Y(t)$ are drawn from a bivariate normal distribution with a mean of 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 [29, 30]. The properties of these groups can impart a variety of important fitness effects [31]. 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) [32, 30]. One study [30] 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 [33] 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 [34]. 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" ([35] 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 [31].

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 4A; $p < 0.01$), 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 4B; $p = 1/3^3 \approx 0.04$). 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 4D). Note that although trials are independent of each other, all between-trial correlations from the full data set are positive (Fig 4B), emphasizing the danger in applying naive data analysis methods to data that are autocorrelated or even nonstationary.

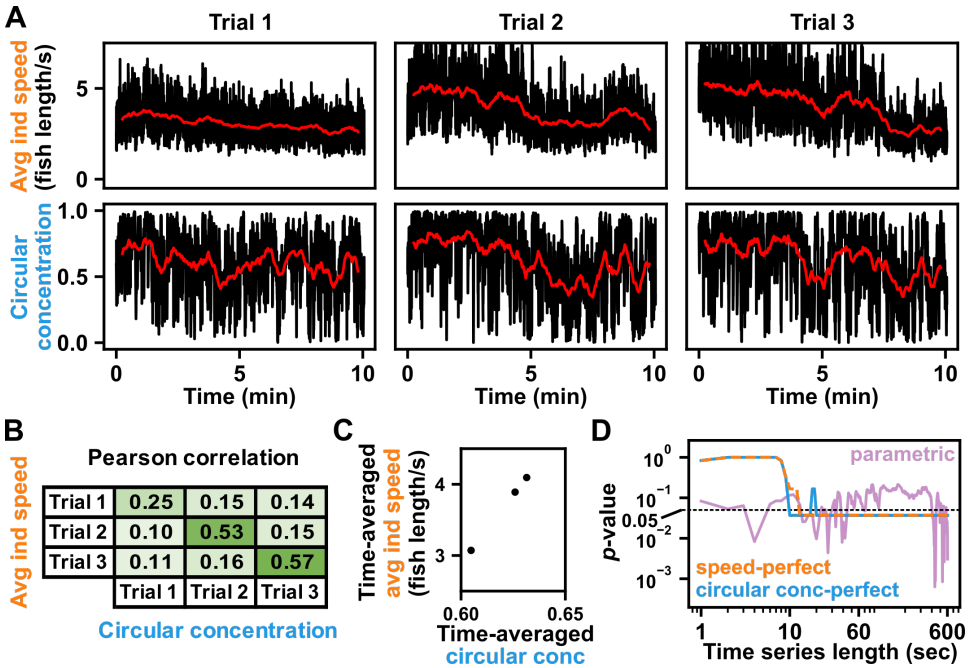

Figure 4: Dependence between average individual speed and circular concentration in small groups of zebrafish. (**A**) Time series of average individual speed and circular concentration for the three replicate videos. Black curves show original time series, and red curves show a 30-second moving average. 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; [36]) 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 [37]) 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.

Discussion

Outside the context of time series, permutation tests have been celebrated because they require only minimal assumptions [38]. To satisfy the requirements of permutation tests, it is sufficient (see Corollary 4 in Appendix 1) 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 [39]. 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 [39], 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 $n = 3$ regardless of how strong the "true" correlation may be. Here, we have provided the permute-match test (Fig. 1), 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. 3).

Since permutation-based tests such as the permute-match test can be easy to misinterpret [40], 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 [27]. 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.

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). 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) [41]. 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 1 and its validity is proved in Appendix 2. For the KPSS test, we used the function "statsmodels.tsa.stattools.kpss" in the Python package Statsmodels (ver. 0.10.1) [42]. 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).

References

[1] C. W. Granger and P. Newbold, "Spurious regressions in econometrics," *Journal of econometrics*, vol. 2, no. 2, pp. 111–120, 1974.

[2] 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.

[3] 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.

[4] 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.

[5] J. S. Rosenthal, *First Look At Rigorous Probability Theory, A*. World Scientific Publishing Company, 2006.

[6] S. M. Ross, *A first course in probability*. Pearson, 2014.

[7] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

[8] 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.

[9] 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.

[10] 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.

[11] D. Guarin, A. Orozco, and E. Delgado, "A new surrogate data method for nonstationary time series," 2010.

[12] W. Greene, *Econometric Analysis*. Pearson, 2012.

[13] 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.

[14] 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.

[15] 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.

[16] T. Schreiber and A. Schmitz, "Surrogate time series," *Physica D: Nonlinear Phenomena*, vol. 142, no. 3-4, pp. 346–382, 2000.

[17] A. E. Yuan and W. Shou, "Data-driven causal analysis of observational biological time series," *Elife*, vol. 11, p. e72518, 2022.

[18] 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.

[19] 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.

[20] 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.

[21] 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, pp. 1–15, 2014.

[22] 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.

[23] 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.

[24] M. D. Ernst, "Permutation methods: a basis for exact inference," *Statistical Science*, pp. 676–685, 2004.

[25] 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.

[26] 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.

[27] 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.

[28] W. D. Koenig, "Spatial autocorrelation in california land birds," *Conservation Biology*, vol. 12, no. 3, pp. 612–620, 1998.

[29] E. Shaw, "Schooling fishes," *American Scientist*, vol. 66, no. 2, pp. 166–175, 1978.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] P. Berens, "Circstat: a matlab toolbox for circular statistics," *Journal of statistical software*, vol. 31, pp. 1–21, 2009.

[35] S. R. Jammalamadaka and A. Sengupta, *Topics in circular statistics*, vol. 5. world scientific, 2001.

[36] 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.

[37] 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.

[38] 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.

[39] A. E. Yuan and W. Shou, "An exactly valid and distribution-free statistical significance test for correlations between time series," *bioRxiv*, 2022.

[40] A. F. Hayes, "Permutation test is not distribution-free: Testing h_0: ρ= 0.," *Psychological Methods*, vol. 1, no. 2, p. 184, 1996.

[41] 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.

[42] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," in *Proceedings of the 9th Python in Science Conference*, vol. 57, pp. 10–25080, Austin, TX, 2010.

[43] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses*, vol. 3. Springer, 2005.

[44] J. Hemerik and J. Goeman, "Exact testing with random permutations," *Test*, vol. 27, no. 4, pp. 811–825, 2018.

[45] 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.

[46] 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.

Appendix

In Appendix 1, 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 2, we provide the theoretical justification for the permute-match test, which is the main topic of this work. Appendix 3 contains supplementary data.

1 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 3 can be mostly reconstructed by piecing together various theorems, examples, and homework problems from section 15.2.1 of Lehmann and Romano [43]. 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 [24, 44].

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 [43]. 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 2 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 1

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 2 *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 3 *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. 2) 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. 2.

We first set up the problem in a way that allows us to apply Lemma 1. To construct the random variable Z and transformation set H for Lemma 1, we choose $Z = (\boldsymbol{X}, \boldsymbol{Y})$ and $h_i Z = (\boldsymbol{X}, g_i \boldsymbol{Y})$. Lemma 1 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 1.

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 1, 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 1, 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 3 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 4 *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. 2) using a correlation function with the property that

$$T(\boldsymbol{X}, \boldsymbol{Y}) = T(g_i \boldsymbol{X}, g_i \boldsymbol{Y}) \tag{2}$$

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

Proof: Lemma 3 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. 2, we have:

$$\begin{aligned} T(\boldsymbol{X}, g_i \boldsymbol{Y}) &= T(g_i^{-1}\boldsymbol{X}, g_i^{-1} g_i \boldsymbol{Y}) \\ &= T(g_i^{-1}\boldsymbol{X}, \boldsymbol{Y}). \end{aligned} \tag{3}$$

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

$$\{g_1, ..., g_{n!}\} = \{g_1^{-1}, ..., g_{n!}^{-1}\} \tag{4}$$

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. 3 and the second equality follows from Eq. 4. 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 3 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 3 were replaced by the requirement of an exchangeable $\boldsymbol{X}$. Applying this exchangeable-$\boldsymbol{X}$ version of Lemma 3 completes the proof.

2 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 5

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 6 *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 7 *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:

$$C_i = \operatorname*{argmax}_j \left(\rho(X_j, Y_i)\right)$$

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 5 to $(C_1, ..., C_n)$ we would nearly be done. However, Lemma 5 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 5, which tells us that

$$P(C_1 = 1, ..., C_n = n|\boldsymbol{X}) \leq 1/n^n. \tag{5}$$

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. [5] pg. 152). The inequality $\overset{(c)}{\leq}$ simply applies Eq. 5. 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 1D already described the same claim, but using visual arguments instead of written ones.

Lemma 8

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

$$T(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, Y_i). \tag{6}$$

Perform the permutation test of independence (Def. 2) using T as the statistic, and obtain the p-value p_{perm}. Also perform the match test (Def. 6) 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. 6 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 9 *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. 2) using T as the statistic, and obtain the p-value p_{perm}. Also perform the match test (Def. 6) 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.

$$\begin{aligned} P(\min(p_{perm}, p_{match}) \leq \alpha) &= P(p_{perm} \leq \alpha \cup p_{match} \leq \alpha) \\ &= P(p_{perm} \leq \alpha) + P(p_{match} \leq \alpha) - P(p_{perm} \leq \alpha \cap p_{match} \leq \alpha) \end{aligned} \tag{7}$$

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. 7 reduces to

$$P(\min(p_{perm}, p_{match}) \leq \alpha) = P(p_{match} \leq \alpha)$$

which is at most α by Lemma 7.

If $1/n! \leq \alpha < 1$, then by Lemma 8,

$$\begin{aligned} P(p_{match} \leq \alpha) &= P(p_{match} = 1/n^n) \\ &= P(p_{perm} = 1/n^n \cap p_{match} = 1/n!) \\ &\leq P(p_{perm} \leq \alpha \cap p_{match} \leq \alpha) \end{aligned}$$

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. 7 reduces to

$$P(\min(p_{perm}, p_{match}) \leq \alpha) = P(p_{perm} \leq \alpha)$$

which is at most α by Lemma 3.

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.

The above result showed that the permute-match test, which can produce a p-value as low as $1/n^n$, is valid under its null hypothesis (i.e. that $\boldsymbol{X}$ is independent of $\boldsymbol{Y}$, and the elements of $\boldsymbol{Y}$ are iid). The next result shows that any test that can produce a p-value *lower* than $1/n^n$ must not be universally valid under the same null hypothesis.

Before proceeding, we clarify the notation that we will use (although much of it is the same as before): For symbols using the letters "X" and "Y", an upper-case symbol denotes a random variable, a lower-case symbol denote a possible realization of a random variable, a bolded symbol denotes a sequence, and an unbolded symbol denotes an element of the sequence (i.e. an individual trial). In detail, X_i is a random variable in $\mathbb{R}^{d_X}$, and x_i is a point in the same space as X_i (i.e. $\mathbb{R}^{d_X}$), and a bold symbol is a sequence of length n, meaning that $\boldsymbol{X} = (X_1, ..., X_n)$ and $\boldsymbol{x} = (x_1, ..., x_n)$. We define Y_i, y_i, $\boldsymbol{Y}$, and $\boldsymbol{y}$ analogously.

Since we seek results that hold for a general class of hypothesis tests, we need some very general language to discuss testing procedures. To set this up, note that a "p-value" is, ultimately, a number that depends on the data. That is, here we will think of p as simply a function that maps the data $(\boldsymbol{x}, \boldsymbol{y})$ to a number.

Theorem 10

Let p ("the p-value") be a measurable function that maps $(\boldsymbol{x}, \boldsymbol{y})$ to a real number. Let α ("the significance level") be a number in the image of p, meaning that there is at least some input $(\boldsymbol{x}, \boldsymbol{y})$ for which $p(\boldsymbol{x}, \boldsymbol{y}) = \alpha$. Then, there exist random variables $(\boldsymbol{X}, \boldsymbol{Y})$ such that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, the elements of $\boldsymbol{Y}$ are iid, and

$$P(p(\boldsymbol{X}, \boldsymbol{Y}) = \alpha) \geq 1/n^n.$$

Proof: Let $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = ((\tilde{x}_1, ..., \tilde{x}_n), (\tilde{y}_1, ..., \tilde{y}_n))$ be a sequence of values such that $p(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = \alpha$. In other words, $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ is an element of the preimage of α under p. We know that $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}})$ is guaranteed to exist because the theorem requires that α be in the image of p.

The proof proceeds by example. That is, we will construct an $\boldsymbol{X}$ and $\boldsymbol{Y}$ that satisfy the requirements of the theorem. For $i = 1, ..., n$, let $\boldsymbol{X} = \tilde{\boldsymbol{x}}$ with probability of 1. We may do this because the only requirement imposed on $\boldsymbol{X}$ is that it needs to be independent of $\boldsymbol{Y}$. Next, let $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be a sequence of iid random variables where each Y_i has the probability distribution:

$$P(Y_i = \tilde{y}_j) = 1/n$$

for $j = 1, ..., n$. In other words, to generate each Y_i, we independently roll an n-sided die and we set Y_i equal to $\tilde{y}_j$, where j is the result of the die roll. From this construction, $\boldsymbol{X}$ and $\boldsymbol{Y}$ readily satisfy the null hypothesis (that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, and the elements of $\boldsymbol{Y}$ are iid).

Having constructed the example, we now show that it satisfies the required property of $P(p(\boldsymbol{X}, \boldsymbol{Y}) = \alpha) \geq 1/n^n$. We begin by noting that

$$\prod_{i=1}^{n} P(Y_i = \tilde{y}_i) \geq 1/n^n. \tag{8}$$

Here, we write $\geq$ because the left side is strictly equal to $1/n^n$ if all of the $\tilde{y}_i$ terms are unique, whereas the left side is greater than $1/n^n$ if the $\tilde{y}_i$ terms are not unique.

Next, we have:

$$\begin{aligned}\prod_{i=1}^{n} P(Y_i = \tilde{y}_i) &= P(\boldsymbol{Y} = \tilde{\boldsymbol{y}}) \\ &= P(\boldsymbol{X} = \tilde{\boldsymbol{x}}, \boldsymbol{Y} = \tilde{\boldsymbol{y}}) \leq P(p(\boldsymbol{X}, \boldsymbol{Y}) = \alpha) \end{aligned} \tag{9}$$

where the first line uses the fact that the Y_i terms are iid, the equality on the second line uses the fact that $\boldsymbol{X}$ is always equal to $\tilde{\boldsymbol{x}}$, and the inequality on the second line uses the fact that $p(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) = \alpha$.

Combining Eq. 8 and Eq. 9 completes the proof.

Let us briefly clarify the importance of this result. Suppose that there is some test (with the same null hypothesis as the permute-match test) that can return $p = \alpha$ for some significance level α that is below $1/n^n$. Then, Theorem 10 says that there is some pathological set of random variables $(\boldsymbol{X}, \boldsymbol{Y})$ for which the null hypothesis is satisfied, and where

$$P(p(\boldsymbol{X}, \boldsymbol{Y}) = \alpha) \geq 1/n^n.$$

But we just said that $1/n^n > \alpha$, so

$$P(p(\boldsymbol{X}, \boldsymbol{Y}) = \alpha) > \alpha$$

which means that such a test would not be valid.

3 Illustration of permute-match test with logistic map system

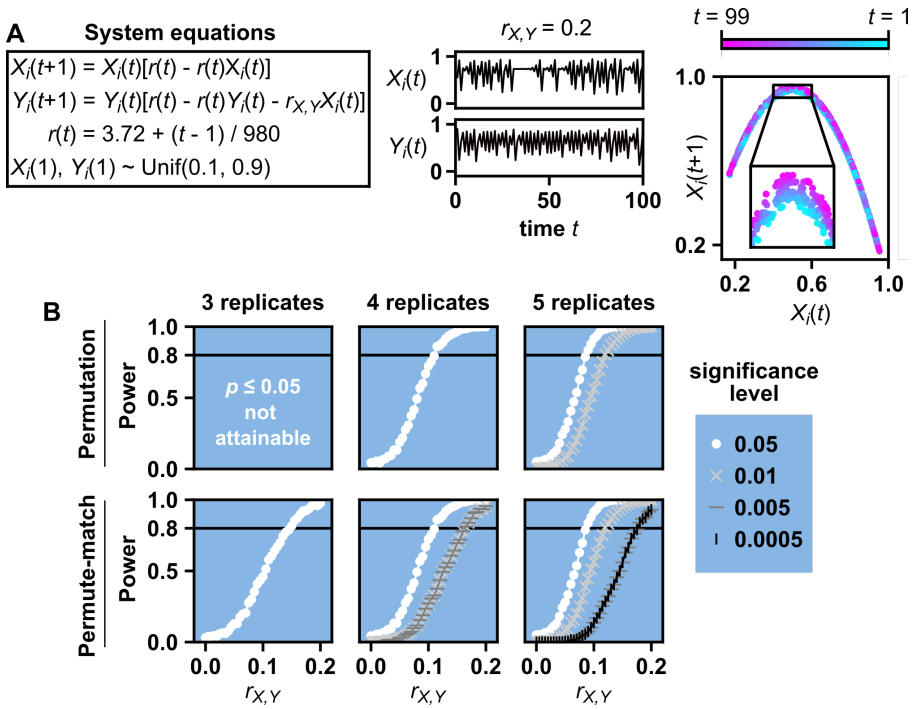

Figure 5: Statistical power of the permute-match test and permutation test in a nonlinear and nonstationary system. (**A**) System equations and example dynamics. The processes X and Y are given by a coupled logistic map. The system is nonstationary because the parameter $r(t)$ varies with time from 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 [45]. 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 [45, 46].