Improved permutation tests enable detection of dependence between nonstationary time series despite limited numbers of replicates

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

1 Molecular and Cellular Biology PhD program, University of Washington, Seattle, United States; 2 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

In disciplines from ecology to neuroscience, researchers analyze correlations between pairs of nonstationary time series to infer relationships among variables. This often involves a statistical test to determine whether an observed correlation is stronger than expected under the null hypothesis of independence. Testing for dependence between nonstationary time series with only one experimental replicate is exceedingly challenging. However, with many replicates, a nonparametric trial-swapping permutation test can be employed, comparing within-replicate correlations to between-replicate correlations. Although largely assumption-free, this test is severely limited by the number of replicates because its minimum achievable p-value is $1/n!$ where n is the number of replicates. This curtails its applicability to many biomedical studies, where n is frequently as low as 3, which would render significance thresholds like 0.05 unattainable. To address this, we propose modified permutation tests that can report lower p-values of $2/n^n$ or $1/n^n$ when there is strong evidence of dependence. We prove that the tests guarantee a false positive rate at or below the significance level, as long as replicates come from independent and identical experiments. We demonstrate this approach by confirming the observation that groups of zebrafish swim faster when directionally aligned, using an existing dataset with 3 biological replicates.

Data availability

All code used, along with execution instructions, is within S1 Code. The time series data analyzed in this paper can be obtained via the following link: https://drive.google.com/drive/folders/1UmzlX-yJhzQ5KX5rGry8wZgXvcz6HefD

Introduction

Scientists frequently look for correlations between variables to identify potentially important relationships, or to support conceptual models. In disciplines with a focus on dynamics such as ecology and physiology, 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 (such as mean or variance) 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 nonstationary.

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". 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.

In contrast to correlation, "dependence" is a hypothesis about the relationship between variables, and it has clearer scientific 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 be replaced by probability densities [5, 6].) The importance of dependence relationships can be attributed to Reichenbach's common cause principle, which states that if two variables are dependent, then either they share a common causal influence, or one variable causally influences the other (possibly indirectly) [7, 8]. Thus, it is useful to to first test whether an observed correlation indicates dependence before pursuing specific mechanistic 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, meaning that its 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 with each step. In a second approach, 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].

Alternatively, an approach based on permutations is possible when there are multiple identically distributed and independent (iid) replicates (or equivalently, 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 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. Since this procedure checks for dependence in trial 1 specifically, checking for dependence in other trials via this method would require analogous tests, and potentially a multiple testing correction.

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, 20]. For instance, two influential longitudinal human gut microbiome sampling studies relied on cohorts of two subjects each [21, 22]. Additionally, secondary analysis of public data from earlier studies can be a resource-efficient mode of research [23], and this approach is necessarily limited to the number of replicates within the existing dataset.

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 [24, 25]. As an example with $n = 4$ trials, the

test procedure begins by computing the mean within-trial correlation (ρ denotes correlation):

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

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. This is the same as permuting Xs and maintaining the order of the Ys. For instance, one such permutation might be:

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

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, ρ_{orig} (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 [26, 24, 27]. It has also been used to detect correlations between variables measured in brain images, which can have similar nonstationarity challenges as time series [28]. 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).

Yet, even the permutation test remains considerably limited by the number of trials. 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 $1/3! \approx 0.17$ and with $n = 4$, the lowest is $1/4! \approx 0.04$. These minima apply even if the average within-trial correlation is much greater than the average between-trial correlation. 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. We illustrate this approach using simulations as well as a real world example involving recordings of zebrafish swimming behavior with only 3 replicate trials.

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 $\rho(X_i, Y_i) > \rho(X_j, Y_i)$ for all $j \neq i$. 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) for all X trials,

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

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 boldface $\boldsymbol{X}$ or $\boldsymbol{Y}$ to indicate a collection of trials, X_i or Y_i to indicate an individual trial, 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. With a perfect match, the original correlation ρ_{orig} is strictly greater than any of the shuffled correlations, as illustrated in Fig 1C. Thus, a perfect match guarantees $p_{perm} = 1/n!$, the lowest p-value achievable with the permutation test (Fig 1D). Furthermore, letting P denote probability, we can prove that

$$P(X\text{−perfect match}) \leq 1/n^n$$

under the null hypothesis H_0 wherein (1) the Y_i trials are iid, (2) the X_i trials are iid, and (3) $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent (Lemma 6 in Appendix 2). By an analogous argument (Lemma 7 in Appendix 2), under H_0 we

have $P(Y-\text{perfect match}) \leq 1/n^n$ and thus

$$P(X - \text{perfect match or } Y-\text{perfect match}) \leq 2/n^n. \tag{1}$$

Permute-match tests: modified permutation tests of dependence

We now describe tests for dependence between nonstationary time series based on the ideas just laid out, which we will call permute-match tests. 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. We introduce two tests: simultaneous permute-match and sequential permute-match. As we will see, the former can be more reproducible while the latter can be more powerful.

In the simultaneous permute-match test (Fig 1E), we check for both an $X-$perfect and a $Y-$ perfect match. If either the X- or Y-perfect match occurs, we write down a p-value to be $p = 2/n^n$ per equation 1. If neither X nor Y achieves a perfect match, then we fall back to the permutation test: $p = p_{perm} = N_{\geq}/n!$ where $N_{\geq}$ is the number of all $n!$ possible permutations (including the original ordering) that produce an average correlation that is at least as large as the original ordering (Fig 1A).

In the sequential permute-match test, we begin by choosing a variable (either X or Y) for the first perfect match test. For example, suppose we choose X first (in which case we refer to the procedure as the "permute-match$_{(X \blacktriangleright Y)}$" test). If an X-perfect match is observed, then $p = 1/n^n$ (Fig 1F). If X-perfect match is not observed, we then check for a Y-perfect match, and write $p = 2/n^n$ if a Y-perfect match is observed. If neither an X-perfect nor Y-perfect match occurs, then $p = p_{perm}$. A permute-match$_{(Y \blacktriangleright X)}$ test is defined analogously.

Under H_0 (i.e. the Y_i replicates are iid, the X_i replicates are iid, and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent), all the above tests are valid, meaning that for any significance level α, we have $P(p \leq \alpha) \leq \alpha$. This fact is proven as Theorems 10-12 in Appendix 2.

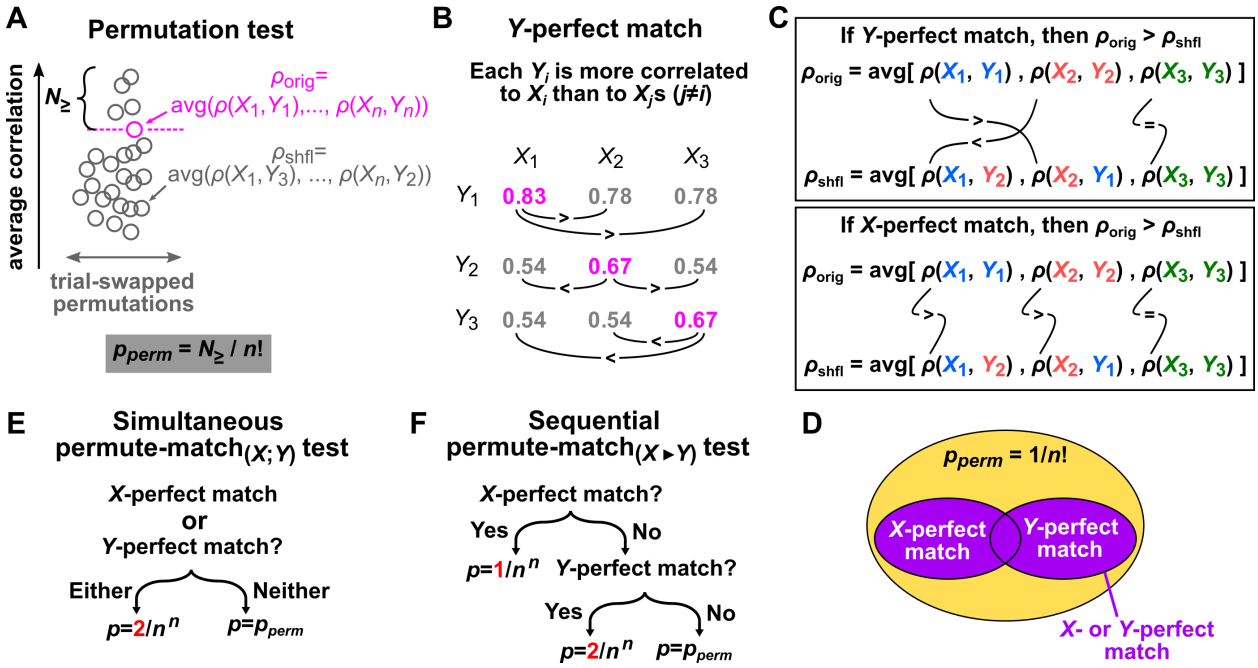


Figure 1: The permutation test, perfect match concept, and permute-match tests. (**A**) The permutation test. Circles indicate the average correlation obtained with (grey) or without (pink) trial swapping. Here, $N_{\geq}$ is the number of average correlations (including the original correlation) that 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, including the original. (**B**) Illustration of a Y-perfect match, represented as a table of correlations wherein each diagonal entry is the greatest in its own row. Each line with a "$>$" or "$<$" symbol denotes a required ordering relationship between the two numbers at either end of the line. Note that this example has a Y-perfect match, but not an X-perfect match. For an X-perfect match, each diagonal entry would need to be the greatest in its own column. (**C**) If an X- or Y-perfect match occurs, then the original correlation, ρ_{orig} is always greater than any shuffled correlation ρ_{shfl}. Top: If a Y-perfect match occurs, then each Y_i gives the greatest correlation when it s paired with the X_i from the same trial. A shuffled correlation ρ_{shfl} pairs Y_is from some trials with X_js from different trials, thereby reducing the correlation. Lines with symbols "$>$", "$<$", and "$=$" denote comparisons between terms. Bottom: If an X-perfect match occurs, a similar argument can be applied. (D) Since a perfect match ensures that the original correlation is greater than any shuffled correlation, a perfect match also ensures that p_{perm} takes on its lowest possible value of $1/n!$. (E) The simultaneous permute-match$_{(X;Y)}$ test. (F) The sequential permute-match$_{(X \blacktriangleright Y)}$ test; note that the permute-match$_{(Y \blacktriangleright X)}$ test is defined analogously.

The permute-match tests (Fig 1E-F) are "upgrades" to the permutation test (Fig 1A) because permute-match tests always report the same or a lower p-value (since $1/n^n \leq 2/n^n < 1/n!$ for all integers $n \geq 2$; see Lemma 9 in Appendix 1). In exchange for lower p-values, the permute-match tests have 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 tests require that all X_i trials are iid and that 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.

The different tests - permutation test, simultaneous permute-match test, and sequential permute-match tests - are valid tests of dependence given iid replicates (Theorems 10-12 in in Appendix 1) . In other words, they all guarantee false positive rates at or below the significance level. Two subtle aspects of the permute-match tests are worth mentioning here. First, no multiple testing correction is applied to account for the fact that both match tests and a permutation test are used. In other words, although the simultaneous permute-match$_{(X;Y)}$ test assigns $p = 2/n^n$ instead of $p = 1/n^n$ when an X- or Y-perfect match occurs as a Bonferroni-type correction, no such correction is needed when combining the match test and the permutation test (Fig 1E). This is because the permutation test and perfect match events are *nested*. That is, an X- or Y-perfect match is possible only if p_{perm} is already at its lowest possible value of $1/n!$ (Fig 1C-D). Thus, unlike combining un-nested tests where each can separately contribute false positives, we can safely combine the checks for perfect matches with the permutation test. The second subtlety refers specifically to the sequential permute-match tests. These tests use the lower p-value of $1/n^n$ if the first perfect match occurs, but use the more conservative "Bonferroni-corrected" p-value of $2/n^n$ if only the second match occurs (Fig 1F). This superficially resembles the practice of conducting additional tests merely because earlier tests did not result in a detection - a form of p-hacking that generally results in an invalid final p-value. However, our sequential permute-match procedure is actually valid because it combines binary events instead of continuous variables. See Appendix 3 for a detailed discussion of this point.

Permute-match test variants may differ in statistical power

Although the different tests - permutation test, simultaneous permute-match test, and sequential permute-match tests - are all valid in the sense of correctly controlling false positive rates, how do they vary in power - the probability of detecting true dependence? The answer depends on the number of replicates n and the desired significance level α. There are four regimes to consider (Fig 2A).

Regime 1: when $\alpha \geq 1/n!$, all tests have the same power. If the permutation test detects dependence, permute-match tests will too since they fall back to a permutation test in the "worst case" (Fig 2A, column 1). Conversely, if the permutation test did not detect dependence (i.e. $p_{perm} > \alpha$), then it follows that $p_{perm} > 1/n!$, which implies that neither an X-perfect nor Y-perfect match occurred (Fig 1D), and so the permute-match tests cannot have detected dependence either.

Regime 2: when $1/n! > \alpha \geq 2/n^n$, all permute-match tests are tied for highest power (Fig 2A, column

2). In this regime, the permutation test has zero power, since its p-value is limited at $1/n!$. Conversely, if either an X-perfect or a Y-perfect match occurs, then, the sequential permute-match tests as well as the simultaneous permute-match$_{(X;Y)}$ test will all detect dependence.

Regime 3: when $2/n^n > \alpha \geq 1/n^n$, the simultaneous permute-match$_{(X;Y)}$ test will have no power as the desired p-value is below the lowest possible p-value achievable by the test. Only the permute-match$_{(X \blacktriangleright Y)}$ or $_{(Y \blacktriangleright X)}$ tests now have power (Fig 2A, column 3). Note that permute-match$_{(X \blacktriangleright Y)}$ and permute-match$_{(Y \blacktriangleright X)}$ test may differ in power, as can be seen in the upper left panel of Fig 2D. We do not currently know how to pre-diagnose which of these two tests will be most powerful based on a given dataset, so the choice of which test to run in this case is arbitrary.

Regime 4: when $\alpha < 1/n^n$, no test has any power even when a perfect match occurs.

Figure 2B-D illustrates these different scenarios using a simulated nonstationary system and some common significance cutoffs. We simulated time series of two variables (X and Y) from a small number of replicates (between 3 and 5). Here, the time series were produced from a pair of linear time trends with coupled additive noise ($r_{X,Y}$ determines the strength of coupling; Fig 2B). We chose the Pearson correlation coefficient as our correlation function ρ. The various sub-panels of Fig 2D illustrate the different aforementioned regimes.

The above example was chosen for ease of presentation - in fact, the example is so simple that it may be correctly treated by just fitting and removing the linear trend. In Appendix 4, we show that the same ideas hold in an example with a more complex nonstationary data-generating process (a logistic map with time-varying parameters) and a nonlinear correlation function (cross-map skill).

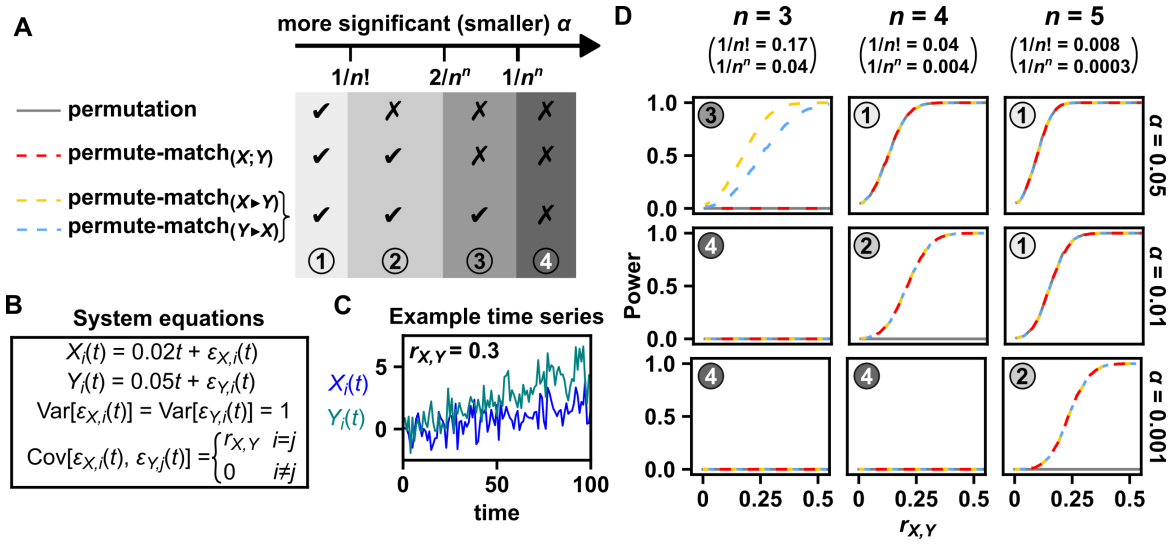


Figure 2: Statistical power of various permutation test variants as illustrated using a simple nonstationary system. (**A**) Summary of test power as a function of the significance level α and the number of replicates n. (**B**, **C**) System equations and example dynamics. The processes X and Y are given by a linear trend with additive noise on a time grid $t = 1, 2, ..., 100$. 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 chart shows an example pair of time series where X and Y are dependent ($r_{X,Y} = 0.3$). (**C**) Statistical power of the permutation test and various permute-match tests as a function of the replicate number n, significance level α, and strength of dependence $r_{X,Y}$. Power was calculated from 5000 simulations at each value of $r_{X,Y}$ between $r_{X,Y} = 0$ and 0.54 in steps of size 0.01. We chose the Pearson correlation coefficient as our correlation function ρ.

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 video segments when they were in a high-polarization state (i.e. when they were directionally aligned) than during segments when they were in a low-polarization state. A correlation between polarization and speed might indicate statistical dependence between the two variables, or it might just be a consequence of temporal trends (e.g. polarization and swimming speed both happen to decrease with time).

Using a publicly available dataset [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], page 15). 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].

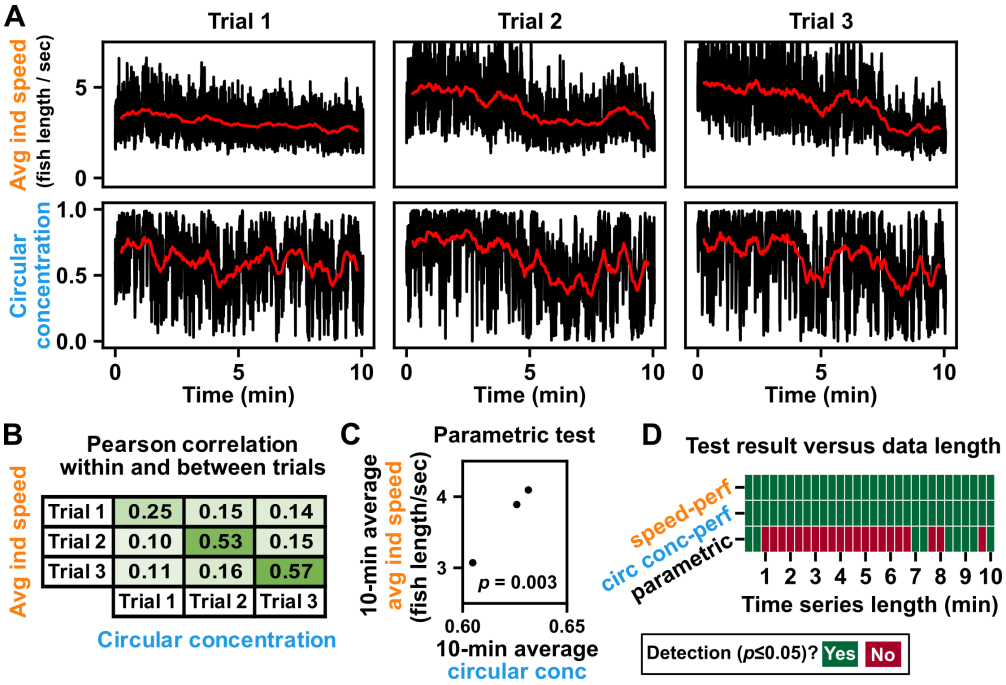
Neither average individual speed nor circular concentration is obviously stationary (Fig 3A). By visual inspection, the average individual speed seems to decrease across time, and all time series are deemed nonstationary by a Kwiatkowski–Phillips–Schmidt–Shin test ($p < 0.01$), and so we may be on shaky ground to use methods that require stationarity. Additionally, since only 3 trials are available, the permutation test and the simultaneous permute-match test cannot detect dependence at the 0.05 level, so we perform a sequential permute-match test.

We arbitrarily started with average individual speed and identified a "speed"-perfect match, indicating a significant correlation between average individual speed and circular concentration (Fig 3B; $p = 1/3^3 \approx 0.04$). Significance does not depend on the choice of which match is tested, since both a "speed"-perfect and a "circular concentration"-perfect match is obtained. Note that although trials are independent of each other, all between-trial correlations from the full dataset are positive (Fig 3B), illustrating the danger in applying naive data analysis methods to data that are autocorrelated or even nonstationary.

To try a parametric alternative, for each trial we averaged values of average individual speed and circular concentration over the full 10 minutes (Fig 3C; where each point is one trial). The sample Pearson correlation of the time-averaged variables is 0.99996, and a one-tailed test of significance (using the function stats.pearsonr from the Python package Scipy [36]) also detects dependence ($p \approx 0.003$), consistent with the permute-match tests. This test relies on a null distribution derived under the assumption that the three points in Fig 3C are drawn from a bivariate Gaussian distribution with zero covariance. We view this test as comparable to the permute-match tests because although it requires a parametric assumption, it can be valid even when the time series are nonstationary.

Since data are limited in many scientific applications, we asked whether the results of the sequential permute-match tests and the parametric Pearson correlation test were consistent across different time series lengths. We truncated time series to various lengths (from 20 seconds to 10 minutes), and checked whether each test reported a significant ($p \leq 0.05$) correlation (Fig 3D). Both sequential permute-match tests reported a significant correlation at all lengths tested, whereas the parametric test was inconsistent, finding a

significant correlation for about one third of lengths.



	Trial 1	Trial 2	Trial 3
Trial 1	0.25	0.15	0.14
Trial 2	0.10	0.53	0.15
Trial 3	0.11	0.16	0.57

Figure 3: A sequential permute-match test detects 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; [37]) 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**) A parametric test also detects dependence. As 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 one-tailed test of significance gives $p \approx 0.003$. (**D**) Permute-match tests detected a significant correlation between speed and circular concentration more consistently than the parametric test. We truncated time series to different lengths starting from the first frame, and compared the parametric test with the two possible permute-match tests. Length was varied between 20 seconds and 10 minutes in 20-second increments, with each increment representing 640 frames since the frame rate is 32 frames per second. 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. No distributional assumptions are required, and the correlation function can even be asymmetric so that $\rho(X, Y) \neq \rho(Y, X)$ as occurs when correlations are based on techniques that use one time series to predict values of another (e.g. Appendix 4). 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 $1/n!$, 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 dependence may be. This limitation is particularly severe for researchers testing several related hypotheses, as this ideally requires a test that can produce p-values sufficiently low to maintain significance even after a multiple-testing correction.

Here we have described modified permutation tests - the sequential and simultaneous permute-match tests - that have access to lower p-values. In the sequential permute-match test, if the first variable does not display a perfect match, this test falls back to the simultaneous permute-match test. Thus, the sequential permute-match test is more powerful than the simultaneous test. However, the sequential test requires the arbitrary choice of checking for an X- or Y-perfect match first, making it potentially less reproducible between studies than the simultaneous test, which does not involve an arbitrary choice. If neither an X- nor Y-perfect match is obtained, both tests default to a permutation test. This step is valid because the tests are nested in the sense that a perfect match can only occur when the permutation p-value takes on its lowest possible value (Lemma 8 in Appendix 1).

It seems likely that the tests could be further modified to report an even lower p-value when both an X- and Y-perfect match occur, as in Fig 3B. However, we have not yet determined a sharp upper bound on the probability of this two-way match event. This problem is left for future efforts.

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 [28]. Beyond spatial processes, dependent data also appear in nucleotide

sequences and natural language text. Overall, advances in methods that take advantage of multiple replicates may facilitate statistical testing without requiring assumptions that are difficult to verify.

Methods

Data preprocessing

We obtained fish trajectory data [33] 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

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

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 $\overrightarrow{v}_{i,t}$. The calculation was carried out using the function "stats.circstats.circvar" in the Python package Astropy (ver. 3.2.2) [40].

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) [41]. 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] Granger CW, Newbold P. Spurious regressions in econometrics. Journal of econometrics. 1974;2(2):111-20.

[2] Yule GU. 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. 1926;89(1):1-63.

[3] Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. The ISME journal. 2016;10(7):1669-81.

[4] Coenen AR, Weitz JS. Limitations of correlation-based inference in complex virus-microbe communities. MSystems. 2018;3(4):e00084-18.

[5] Rosenthal JS. First Look At Rigorous Probability Theory, A. World Scientific Publishing Company; 2006.

[6] Ross SM. A first course in probability. Pearson; 2014.

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

[8] Hitchcock C, Rédei M. Reichenbach's Common Cause Principle. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy. spring 2020 ed. Metaphysics Research Lab, Stanford University; 2020. .

[9] Chan KH, Hayya JC, Ord JK. A note on trend removal methods: The case of polynomial regression versus variate differencing. Econometrica: Journal of the Econometric Society. 1977:737-44.

[10] Isliker H, Kurths J. A test for stationarity: finding parts in time series apt for correlation dimension estimates. International Journal of Bifurcation and Chaos. 1993;3(06):1573-9.

[11] Guarin D, Orozco A, Delgado E. A new surrogate data method for nonstationary time series. arXiv; 2010. Available from: `https://arxiv.org/abs/1008.1804`.

[12] Greene WH. Econometric Analysis. Pearson; 2012.

[13] Iatsenko D, Bernjak A, Stankovski T, Shiogai Y, Owen-Lynch PJ, Clarkson P, et al. Evolution of cardiorespiratory interactions with age. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2013;371(1997):20110622.

[14] Lancaster G, Iatsenko D, Pidde A, Ticcinelli V, Stefanovska A. Surrogate data for hypothesis testing of physical systems. Physics Reports. 2018;748:1-60.

[15] Moulder RG, Boker SM, Ramseyer F, Tschacher W. Determining synchrony between behavioral time series: An application of surrogate data generation for establishing falsifiable null-hypotheses. Psychological methods. 2018;23(4):757.

[16] Schreiber T, Schmitz A. Surrogate time series. Physica D: Nonlinear Phenomena. 2000;142(3-4):346-82.

[17] Yuan AE, Shou W. Data-driven causal analysis of observational biological time series. Elife. 2022;11:e72518.

[18] Bacchetti P, Deeks SG, McCune JM. Breaking free of sample size dogma to perform innovative translational research. Science translational medicine. 2011;3(87):87ps24-4.

[19] Vesterinen HV, Egan K, Deister A, Schlattmann P, Macleod MR, Dirnagl U. 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. 2011;31(4):1064-72.

[20] Ristić-Djurović JL, Ćirković S, Mladenović P, Romčević N, Trbovich AM. Analysis of methods commonly used in biomedicine for treatment versus control comparison of very small samples. Computer Methods and Programs in Biomedicine. 2018;157:153-62. Available from: `https://www.sciencedirect.com/science/article/pii/S0169260717307344`.

[21] Caporaso JG, Lauber CL, Costello EK, Berg-Lyons D, Gonzalez A, Stombaugh J, et al. Moving pictures of the human microbiome. Genome biology. 2011;12(5):1-8.

[22] David LA, Materna AC, Friedman J, Campos-Baptista MI, Blackburn MC, Perrotta A, et al. Host lifestyle affects human microbiota on daily timescales. Genome biology. 2014;15(7):1-15.

[23] Greene CS, Garmire LX, Gilbert JA, Ritchie MD, Hunter LE. Celebrating parasites. Nature genetics. 2017;49(4):483-4.

[24] Martin DC, Buffington V, Becker J. Randomization test of paired data: Application to evoked responses. Psychophysiology. 1981;18(5):524-8.

[25] Ernst MD. Permutation methods: a basis for exact inference. Statistical Science. 2004:676-85.

[26] Buffington V, Martin DC, Becker J. VER similarity between alcoholic probands and their first-degree relatives. Psychophysiology. 1981;18(5):529-33.

[27] Albert M, Bouret Y, Fromont M, Reynaud-Bouret P. Surrogate data methods based on a shuffling of the trials for synchrony detection: the centering issue. Neural Computation. 2016;28(11):2352-92.

[28] Weinstein SM, Vandekar SN, Adebimpe A, Tapera TM, Robert-Fitzgerald T, Gur RC, et al. A simple permutation-based test of intermodal correspondence. Human brain mapping. 2021;42(16):5175-87.

[29] Shaw E. Schooling fishes. American Scientist. 1978;66(2):166-75.

[30] Miller N, Gerlai R. From schooling to shoaling: patterns of collective motion in zebrafish (Danio rerio). PloS one. 2012;7(11):e48865.

[31] Viscido SV, Parrish JK, Grünbaum D. Individual behavior and emergent properties of fish schools: a comparison of observation and theory. Marine Ecology Progress Series. 2004;273:239-49.

[32] Lopez U, Gautrais J, Couzin ID, Theraulaz G. From behavioural analyses to models of collective motion in fish schools. Interface focus. 2012;2(6):693-707.

[33] Romero-Ferrero F, Bergomi MG, Hinz RC, Heras FJ, De Polavieja GG. Idtracker. ai: tracking all individuals in small or large collectives of unmarked animals. Nature methods. 2019;16(2):179-82.

[34] Berens P. CircStat: a MATLAB toolbox for circular statistics. Journal of statistical software. 2009;31:1-21.

[35] Jammalamadaka SR, Sengupta A. Topics in circular statistics. vol. 5. world scientific; 2001.

[36] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods. 2020;17:261-72.

[37] Kwiatkowski D, Phillips PC, Schmidt P, Shin Y. 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. 1992;54(1-3):159-78.

[38] Kulkarni RU, Wang CL, Bertozzi CR. Analyzing nested experimental designs-A user-friendly resampling method to determine experimental significance. PLoS computational biology. 2022;18(5):e1010061.

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

[40] Robitaille TP, Tollerud EJ, Greenfield P, Droettboom M, Bray E, Aldcroft T, et al. Astropy: A community Python package for astronomy. Astronomy & Astrophysics. 2013;558:A33.

[41] Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. In: Proceedings of the 9th Python in Science Conference. vol. 57. Austin, TX; 2010. p. 10-25080.

[42] Lehmann EL, Romano JP. Testing statistical hypotheses. vol. 3. Springer; 2005.

[43] Hemerik J, Goeman J. Exact testing with random permutations. Test. 2018;27(4):811-25.

[44] Steele JM. The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities. MAA problem books series. Cambridge University Press; 2004. Available from: `https://books.google.com/books?id=7GDyRMrlgDsC`.

[45] Sugihara G, May R, Ye H, Hsieh Ch, Deyle E, Fogarty M, et al. Detecting causality in complex ecosystems. Science. 2012;338(6106):496-500.

[46] Ye H, Deyle ER, Gilarranz LJ, Sugihara G. Distinguishing time-delayed causal interactions using convergent cross mapping. Scientific reports. 2015;5:14750.

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 expands on a brief claim made in the main text about p-values of sequential statistical tests, and Appendix 4 contains supplementary data.

Contents

1 Justification of the permutation test of independence **19**

2 Validity of the permute-match test **23**

3 The sequential permute-match test versus a similar flawed procedure **27**

4 Illustration of permute-match test with logistic map system **30**

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 [42]. 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 [25, 43].

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 [42]. 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. 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 $z \in \mathcal{Z}$, let the ordered values of $\{T(h_1 z), ..., T(h_m z)\}$ be

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

Choose some significance level $\alpha \in [0, 1]$ and let the rank k be

$$k = m - \lfloor m\alpha \rfloor \tag{A-1}$$

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

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

where $P(\cdot)$ denotes the probability of an event.

Remark: Before proceeding to the proof, let us walk through the lemma, using a concrete example to help with the most abstract aspects. Suppose that random variable Z is the result of a fair coin flip. That is, $Z = 0$ with 50% chance and $Z = 1$ with 50% chance. Then $\mathcal{Z}$ (the set of valid Z values) is the the set $\{0, 1\}$, and z can take on the value of 0 or 1. Next the lemma introduces H, which is a set of functions that need to satisfy certain requirements with respect to Z. Let's choose $H = \{h_1, h_2\}$ where h_1 "turns the coin over" and h_2 leaves the coin as it is. That is, $h_1 z = 1 - z$ and $h_2 z = z$. To check that H satisfies requirement (1), note that flipping the fair coin over does not change the probability distribution of its outcome, and neither does leaving it alone. To see that H satisfies requirement (2), we need to check that

$$\{h_1 Z, h_2 Z\} = \{h_1 h_1 Z, h_2 h_1 Z\} = \{h_1 h_2 Z, h_2 h_2 Z\}$$

If $Z = 0$, then the above expression becomes

$$\{1, 0\} = \{0, 1\} = \{1, 0\}$$

and if $Z = 1$, then it is

$$\{0, 1\} = \{1, 0\} = \{0, 1\}$$

and since these are unordered sets, they are all equal, so we have verified that H satisfies requirement (2). Next, we need a function T. For example, choose $T(z) = z + 10$. Thus, for $z = 0$ we have $\{T(h_1 z), T(h_2 z)\} = \{11, 10\}$. The lemma then introduces a ranking notation where $T^{(i)}(Z)$ is the ith smallest term in the

sequence. This notation is useful as we will be dealing with the set $\{T(h_1Z), ..., T(h_mZ)\}$ for potentially large values of m. Returning to our example, if $z = 0$, then $T^{(1)}(0) = 10$ and $T^{(2)}(0) = 11$. Similarly for $z = 1$, $T^{(1)}(1) = 10$ and $T^{(2)}(1) = 11$. Also, in the statement of the lemma, $T^{(i)}$ is defined as a function of z, not Z. This is to emphasize that the ordering depends on the particular outcome of the random variable and is not fixed. Finally, the lemma makes a claim about the probability that $T(Z)$ exceeds $T^{(k)}(Z)$, which is the kth lowest statistic in the set. In the example, if $\alpha = 0.05$, then $k = 2 - \lfloor 0.1 \rfloor = 2$. $T^{(2)}(Z)$ is always going to be the second smallest, i.e. the largest, element of $T(h_1z), T(h_2z)$, which is 11. Of course, the probability of $T(Z)$ being *even greater* than the largest number is 0, and since 0 is less than $\alpha = 0.05$ we have verified the lemma in this case. As another example, suppose $\alpha = 0.5$. Then $k = 2 - \lfloor 1 \rfloor = 1$. $T^{(k)}(Z)$ is then $T^{(1)}(Z)$, the smallest element of $\{T(h_1z), T(h_2z)\}$. The probability of $T(Z)$ being greater than the smallest number is $0.5 \leq \alpha = 0.5$, which verifies the lemma for $\alpha = 0.5$.

Proof of Lemma 1: Let $I(A)$ be the indicator function for the event A (meaning that $I(A) = 1$ if A occurs and $I(A) = 0$ otherwise).

First suppose that $T^{(k)}(Z)$ is not tied with any other $T^{(i)}(Z)$. Then Eq. A-1 implies that there are $\lfloor m\alpha \rfloor$ values of $T^{(i)}(Z)$ that are greater than $T^{(k)}(Z)$. That is,

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

To arrive at the second equality, we used the fact that $T^{(k)}(Z) = T^{(k)}(h_iZ)$, which follows from the second requirement of the lemma (i.e. $\{h_1Z, ..., h_mZ\} = \{h_1h_iZ, ..., h_mh_iZ\}$).

Alternatively, if there is possibly a tie for $T^{(k)}(Z)$ (meaning that there may be 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:

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

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

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

Since Z and h_iZ 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_iZ.

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: Here we use the symbols g_i and T 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 $h_iZ = (\boldsymbol{X}, g_i\boldsymbol{Y})$. Lemma 1 has two requirements. First, Z and h_iZ 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 $\{h_1Z, ..., h_mZ\} = \{h_1h_iZ, ..., h_mh_iZ\}$. 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".

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

As in Eq. A-1, let us define $k = m - \lfloor m\alpha \rfloor$, where $n! = m$. Then, the final line above says "the number of permuted $\boldsymbol{Y}$s that produce a smaller T than the original $\boldsymbol{Y}$ is greater than or equal to k." Said another way, "when we rank the Ts from least to greatest, the T from the original $\boldsymbol{Y}$ will have a higher rank than k". This in turn is equivalent to $T(\boldsymbol{X}, \boldsymbol{Y}) > T^{(k)}(\boldsymbol{X}, \boldsymbol{Y})$. 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.

Lemma 3 is "asymmetric" in that it requires $\boldsymbol{Y}$ to be exchangeable and says nothing about $\boldsymbol{X}$. It is possible to write down a variant of this lemma where either an exchangeable $\boldsymbol{Y}$ or an exchangeable $\boldsymbol{X}$ is sufficient, but this requires an extra condition on T. 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. We establish this formally in the corollary below.

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 "nice" property that

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

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. Notice that to show this equivalence it is sufficient to establish that

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

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 $g_i^{-1}g_iZ = Z$. Using Eq. A-2, we have:

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

Note that the inverse permutation g_i^{-1} is guaranteed to exist because any permutation g_i can be "undone" by some other permutation (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

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

because each permutation is an inverse permutation (because g_i inverts g_i^{-1}), 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. A-3 and the second equality follows from Eq. A-4. This shows that the permutation test where $\boldsymbol{Y}$ is permuted is equivalent to the permutation test where $\boldsymbol{X}$ is permuted.

From Def. 2 and Lemma 3 it is clear that a permutation test where $\boldsymbol{X}$ is permuted instead of $\boldsymbol{Y}$ would be valid (i.e. $P(p_{perm} \leq \alpha) \leq \alpha$) if $\boldsymbol{X}$ is exchangeable. Since we have just now shown that the present procedure is equivalent to a permutation test where $\boldsymbol{X}$ is permuted instead of $\boldsymbol{Y}$, it follows that this procedure is valid when $\boldsymbol{X}$ is exchangeable, which 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. Recall from the main text that the null hypothesis of this test is that $X_1, ..., X_n$ are iid, $Y_1, ..., Y_n$ are iid and $\boldsymbol{Y}$ is independent of $\boldsymbol{X}$.

Definition 5 *The perfect match events (M_X and M_Y)*

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. We say that the event M_Y (also called a "Y-perfect match") occurs if and only if:

$$\rho(X_i, Y_i) > \rho(X_j, Y_i)$$

for all pairs (i, j) such that $i \neq j$.

Similarly, the event M_X (also called an "X-perfect match") occurs if and only if

$$\rho(X_i, Y_i) > \rho(X_i, Y_j)$$

for all pairs (i, j) such that $i \neq j$.

Lemma 6

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. Then, the probability of a Y-perfect match is at most $1/n^n$.

Proof: Let us define the random variables $(C_1, ..., C_n)$ as follows:

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

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 the lowest from among the maximizing index values. (For example, if $j = 3$ and $j = 7$ both produced the greatest value of $\rho(X_j, Y_i)$, then C_i would be 3 since $3 < 7$.) These C_i variables are useful because if a Y-perfect match occurs, then it must be the case that

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

Notice that $C_1, ..., C_n$ are iid given $\boldsymbol{X} = \boldsymbol{x}$ for any particular $\boldsymbol{x}$ that $\boldsymbol{X}$ may take. This is because C_i is a function of only Y_i and $\boldsymbol{X}$, so once we condition on $X = \boldsymbol{x}$, each C_i is then a function of only Y_i. Moreover, the $Y_1, ..., Y_n$ are of course iid given $\boldsymbol{X} = \boldsymbol{x}$ because $\boldsymbol{Y}$ and $\boldsymbol{X}$ are independent.

Since the C_i terms are identically distributed given $\boldsymbol{X} = \boldsymbol{x}$, we have

$$P(C_1 = i | \boldsymbol{X} = \boldsymbol{x}) = P(C_i = i | \boldsymbol{X} = \boldsymbol{x})$$

for all i. Therefore we have

$$1 = \sum_{i=1}^{n} P(C_1 = i | \boldsymbol{X} = \boldsymbol{x}) = \sum_{i=1}^{n} P(C_i = i | \boldsymbol{X} = \boldsymbol{x})$$

By dividing the above equation through by n, we may write

$$\frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} P(C_i = i | \boldsymbol{X} = \boldsymbol{x}) \geq \left(\prod_{i=1}^{n} P(C_i = i | \boldsymbol{X} = \boldsymbol{x}) \right)^{1/n} \tag{A-5}$$

where the inequality comes from the AM-GM inequality (e.g. [44] pg. 20).

Since the C_i terms are independent given $\boldsymbol{X} = \boldsymbol{x}$, we have

$$P(C_1 = 1, ..., C_n = n | \boldsymbol{X} = \boldsymbol{x}) = \prod_{i=1}^{n} P(C_i = i | \boldsymbol{X} = \boldsymbol{x}) \tag{A-6}$$

Substituting Eq. A-6 into the right side of inequality A-5 and raising both sides to the nth power gives us

$$\frac{1}{n^n} \geq P(C_1 = 1, ..., C_n = n | \boldsymbol{X} = \boldsymbol{x})$$

Since $\boldsymbol{X}$ must take on some value $\boldsymbol{x}$, it is clear that $P(C_1 = 1, ..., C_n = n) \leq 1/n^n$. More formally, applying the law of total probability we find

$$P(C_1 = 1, ..., C_n = n) = E\left[P(C_1 = 1, ..., C_n = n | \boldsymbol{X} = \boldsymbol{x})\right] \leq E\left[1/n^n\right] = 1/n^n \tag{A-7}$$

Since a Y-perfect match can only occur when $(C_1 = 1, ..., C_n = n)$, we may then write

$$P(M_Y) \leq P(C_1 = 1, ..., C_n = n) \tag{A-8}$$

Combining inequalities A-7 and A-8 completes the proof.

Lemma 7

Let $\boldsymbol{X} = (X_1, ..., X_n)$ be a sequence of iid random variables and let $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be a sequence of random variables. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be independent. Then, the probability of an X-perfect match is at most $1/n^n$.

Proof: The argument is analogous to that of Lemma 6. The main differences are that in this case we define $C_i = \underset{j}{\operatorname{argmax}}\left(\rho(X_i, Y_j)\right)$ instead of $C_i = \underset{j}{\operatorname{argmax}}\left(\rho(X_j, Y_i)\right)$, and here we condition on $\boldsymbol{Y} = \boldsymbol{y}$ instead of $\boldsymbol{X} = \boldsymbol{x}$.

Lemma 8 *A perfect match implies $p_{perm} = 1/n!$*

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

$$T(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, Y_i). \tag{A-9}$$

Perform the permutation test of independence (Def. 2) using T as the statistic, and obtain the p-value p_{perm}. Also check for an X-perfect and Y-perfect match test (Def. 5) using ρ as the correlation function. Then, the occurrence of either an X-perfect match or a Y-perfect match implies that $p_{perm} = 1/n!$.

Proof: This claim is shown by a visual argument in Fig 1C. Here we walk through the same logic in prose.

Suppose a Y-perfect match occurs, meaning that

$$\rho(X_i, Y_i) > \rho(X_j, Y_i) \ \forall \ (i \neq j).$$

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 one choice of i, the $\rho(X_i, Y_i)$ term in Eq. A-9 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 $p_{perm} = 1/n!$.

Suppose now that an X-perfect match occurs. It follows that for each X_i its correlation term $\rho(X_i, \cdot)$ is strictly maximized when X_i is paired with Y_i. Applying a permutation to $\boldsymbol{X}$ (other than the identity permutation) means that for at least one choice of i, the $\rho(X_i, Y_i)$ term in Eq. A-9 becomes replaced with $\rho(X_i, Y_j)$ for some $j \neq i$. Therefore, $T(\boldsymbol{X}, \boldsymbol{Y})$ is be strictly greater than all permuted versions $T(g_i\boldsymbol{X}, \boldsymbol{Y})$ whenever g_i is not the identity permutation. So $p_{perm} = 1/n!$.

Lemma 9

$1/n! \geq 2/n^n$ for all integers $n > 1$, with equality only when $n = 2$.

Proof: When $n = 2$ it is clear that $1/n! = 2/n^n = 1/2$.

For $n > 2$, we establish $1/n! > 2/n^n$ by an inductive argument. For the base case, suppose $n = 3$. Then, $1/n! > 2/n^n$ becomes $1/6 > 2/27$, which is true.

For the inductive step, suppose the target inequality $1/n! > 2/n^n$ holds. Multiplying the target inequality on both sides by $1/(n+1)$, we obtain

$$\frac{1}{(n+1)!} > \frac{2}{n^n}\left(\frac{1}{n+1}\right) > \frac{2}{(n+1)^n}\left(\frac{1}{n+1}\right) = \frac{2}{(n+1)^{n+1}}$$

This shows that $1/n! > 2/n^n$ implies $1/(n+1)! > 2/(n+1)^{n+1}$. We conclude by the inductive principle that the target inequality is true for $n > 2$.

Theorem 10 *The permute-match$_{X \blacktriangleright Y}$ test is valid.*

Let $\boldsymbol{X} = (X_1, ..., X_n)$ and $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be sequences 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 check for an X-perfect match test and a Y-perfect match test (Def. 5) using ρ as the correlation function. Let

$$p = \begin{cases} 1/n^n & \text{if an } X-\text{perfect match occurs} \\ 2/n^n & \text{if a } Y-\text{perfect match occurs, but not an } X-\text{perfect match} \\ p_{perm} & \text{if neither occurs} \end{cases}$$

Then $P(p \leq \alpha)$ for $\alpha \in [0, 1]$ and $n > 1$.

Proof: For notational convenience, we denote the events of an X-perfect match and Y-perfect match as M_X and M_Y.

There are four cases to consider. Either (case i) $\alpha \geq 1/n!$, or (case ii) $1/n! > \alpha \geq 2/n^n$, or (case iii) $2/n^n > \alpha \geq 1/n^n$, or (case iv) $\alpha < 1/n^n$. Note that Lemma 9 tells us that $1/n! \geq 2/n^n$.

Case i: Suppose $\alpha \geq 1/n!$. In this case, $p \leq \alpha$ if M_X occurs or M_Y occurs or $p_{perm} \leq \alpha$.

$$\begin{aligned} P(p \leq \alpha) &= P(p_{perm} \leq \alpha \text{ or } M_X \text{ or } M_Y) \\ &= P(p_{perm} \leq \alpha \text{ or } (M_X \text{ or } M_Y)) \end{aligned}$$

This may be split into 3 terms via the identity $P(A \text{ or } B) = P(A \text{ and } B) + P(A \text{ and not } B) + P((\text{not } A) \text{ and } B)$, as follows:

$$
\begin{aligned}
P(p \leq \alpha) = & P(p_{perm} \leq \alpha \text{ and } (M_X \text{ or } M_Y)) \\
& + P(p_{perm} \leq \alpha \text{ and not } (M_X \text{ or } M_Y)) \\
& + P(p_{perm} > \alpha \text{ and } (M_X \text{ or } M_Y))
\end{aligned}
$$

However, by Lemma 8, if either M_X or M_Y occurs, then $p_{perm} = 1/n!$, which would in turn imply $p_{perm} \leq \alpha$ under case i. That is, the term $P(p_{perm} > \alpha \text{ and } (M_X \text{ or } M_Y))$ in the above summation is zero. Removing this term, we have

$$
\begin{aligned}
P(p \leq \alpha) &= P(p_{perm} \leq \alpha \text{ and } (M_X \text{ or } M_Y)) + P(p_{perm} \leq \alpha \text{ and not } (M_X \text{ or } M_Y)) \\
&= P(p_{perm} \leq \alpha) \leq \alpha
\end{aligned}
$$

where the final inequality is due to Lemma 3. This establishes the claim for case i.

Case ii: Suppose $1/n! > \alpha \geq 2/n^n$. In this case, $p_{perm} > \alpha$ since p_{perm} cannot be less than $1/n!$, so we have

$$
\begin{aligned}
P(p \leq \alpha) &= P(M_X \text{ or } M_Y) \\
&\leq P(M_X) + P(M_Y)
\end{aligned}
$$

From Lemmas 6-7 we know $P(M_X) \leq 1/n^n$ and $P(M_Y) \leq 1/n^n$. Thus the above inequality becomes

$$
P(p \leq \alpha) \leq P(M_X) + P(M_Y) \leq 1/n^n + 1/n^n \leq \alpha
$$

which establishes the claim in case ii.

Case iii: Suppose $2/n^n > \alpha \geq 1/n^n$. In this case, only M_X will give us $p \leq \alpha$. So we have

$$
P(p \leq \alpha) = P(M_X) \leq 1/n^n \leq \alpha
$$

which verifies the claim in case iii.

Case iv: If $\alpha < 1/n^n$, then, $p \leq \alpha$ is impossible so $P(p \leq \alpha) = 0 \leq \alpha$, so the claim is trivially satisfied in case iv.

Having been shown in all four cases, the claim has been proven.

Theorem 11 *The permute-match$_{Y \blacktriangleright X}$ test is valid.*

Consider the same setting as Theorem 10, except that now

$$
p = \begin{cases} 1/n^n & \text{if a } Y\text{−perfect match occurs} \\ 2/n^n & \text{if an } X\text{−perfect match occurs, but not an } Y\text{−perfect match} \\ p_{perm} & \text{if neither occurs} \end{cases}
$$

Then, $P(p \leq \alpha)$ for $\alpha \in [0, 1]$ and $n > 1$.

Proof: The proof is analogous to that of Theorem 11. In fact, the only thing we need to change here is that in case iii, we use the fact that $P(M_Y) \leq 1/n^n$ instead of that $P(M_X) \leq 1/n^n$.

Theorem 12 *The simultaneous permute-match$_{(X;Y)}$ test is valid.*

Consider the same setting as Theorem 10, except that now

$$
q = \begin{cases} 2/n^n & \text{if an } X\text{−perfect match or } Y\text{−perfect match occurs} \\ p_{perm} & \text{if neither occurs} \end{cases}
$$

Then, $P(q \leq \alpha)$ for $\alpha \in [0, 1]$ and $n > 1$.

Proof: We first show by cases that p of Theorem 10 is always less than or equal to q. First, if an X-perfect match occurs, then $p = 1/n^n < 2/n^n = q$. Second, if there is a Y-perfect match but not an X-perfect match, then $p = 1/n^n = q$, Third, if neither an X-perfect nor a Y-perfect match occurs, then $p = p_{perm} = q$. Overall, $p \leq q$. It follows that $q \leq \alpha$ implies $p \leq \alpha$.

Then it follows from Theorem 10 that

$$P(q \leq \alpha) \leq P(p \leq \alpha) \leq \alpha$$

so $P(q \leq \alpha) \leq \alpha$ as required.

3 The sequential permute-match test versus a similar flawed procedure

In the main text we briefly noted that the sequential permute-match test resembles the flawed practice of conducting additional tests merely because earlier tests did not result in a detection. Here we provide a detailed explanation of why this practice generally results in an invalid p-value, and point out that the sequential permute-match test does not have this problem.

To spell out the flawed practice, suppose a researcher runs one test (test X) and obtains a p-value, p_X. If p_X is below their desired significance level α, they stop there and report p_X. Conversely, if $p_X > \alpha$, this researcher performs another test (test Y) that assesses the same biological hypothesis in a slightly different way, obtaining a second p-value, p_Y. However, the researcher has a vague understanding that if one conducts multiple tests, a correction is required, so at this point they perform a Bonferroni correction to account for two tests and report $2p_Y$. Overall, the p-value obtained from this series of steps is:

$$p = \begin{cases} p_X & \text{if } p_X \leq \alpha \\ 2p_Y & \text{otherwise} \end{cases} \tag{A-10}$$

One may ask whether this overall procedure has the desired property (known as 'validity') that it prevents the false positive rate from exceeding the significance level. Unfortunately, this procedure is generally not valid, meaning that it commonly leads to $P(p \leq \alpha) > \alpha$ under the null hypothesis.

To demonstrate this fact mathematically, we make three assumptions. First, we assume that the null hypothesis is true for both test X and Y so that $P(p \leq \alpha)$ is the false positive rate. Second, we assume that p_X and p_Y are both random variables following a uniform distribution between 0 and 1. Finally, we assume that the significance level α is less than 1/2.

The false positive rate of the procedure in Eq A-10 depends on the relationship between the two tests. If tests X and Y tend to report the same result, they will behave more like a single test and the false positive rate of Eq A-10 will be lower; if the tests frequently report different results, they will behave like two genuinely distinct tests, and the false positive rate will be greater. This relationship is quantified in Table A-1, which describes the joint and marginal distributions of the X and Y test outcomes. We write the probability that test X is significant (or not) along the bottom row, and the probability that test Y is significant (or not) along the rightmost column. We then define a parameter (P') which is the probability that test Y is significant while test X is not. We then fill out the rest of the table, corresponding to the probabilities of the other possible test results, by ensuring that the rows and columns of the 2×2 "Joint probability" table sum to the entries in the "Total probability" row and column.

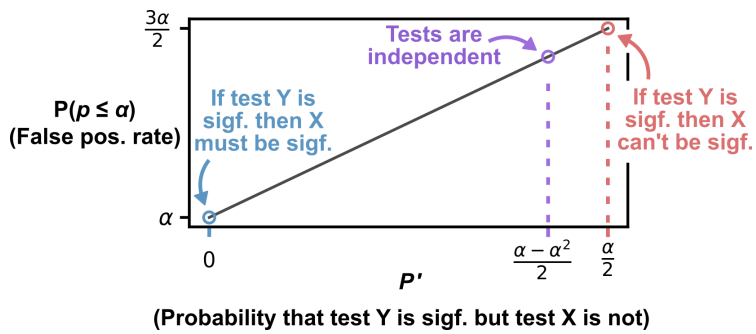


Figure A-1: The procedure of Eq A-10 produces a false positive rate that generally exceeds α, and depends on the relationship between tests X and Y. This relationship X and Y can be quantified by P', the probability that test Y is significant (i.e. $2p_Y \leq \alpha$) but test X is not significant (i.e. $p_X > \alpha$). The false positive range is shown over the full range of P', from 0 to $\alpha/2$. The only way for the overall procedure of Eq A-10 to be valid is when $P' = 0$, where tests X and Y are so tightly coupled that a significant result in test Y deterministically ensures that test X is also significant. Other than this edge case, the false positive rate of the overall procedure will exceed α.

		Joint probability		Total probability
		Test X is significant ($p_X \leq \alpha$)	Test X not significant ($p_X > \alpha$)	
Joint probability	Test Y not significant ($2p_Y > \alpha$)	$\alpha/2 + P'$	$1 - \alpha - P'$	$1 - \alpha/2$
	Test Y is significant ($2p_Y \leq \alpha$)	$\alpha/2 - P'$	P'	$\alpha/2$
Total probability		α	$1 - \alpha$	

Table A-1: Joint and marginal probabilities of the outcomes of tests X and Y, assuming that p_X and p_Y both follow $\text{Unif}(0, 1)$. The parameter P' is the probability that test Y is significant but test X is not.

Each joint probability is of course bounded between 0 and 1. Thus, moving clockwise from the upper left of the 2×2 inner table we have the following constraints on P':

$$
\begin{aligned}
&0 \leq \alpha/2 + P' \leq 1 \\
&0 \leq 1 - \alpha - P' \leq 1 \\
&0 \leq P' \leq 1 \\
&0 \leq \alpha/2 - P' \leq 1
\end{aligned}
$$

Among these constraints, the tightest bounds on P' are $0 \leq P' \leq \alpha/2$. Notice that the false positive rate of the overall procedure is given by:

$$
\begin{aligned}
P(p \leq \alpha) &= 1 - P(p_X > \alpha \text{ and } 2p_Y > \alpha) \\
&= \alpha + P'
\end{aligned}
\tag{A-11}
$$

Fig A-1 plots Eq A-11 within the allowed bounds of P', noting the extreme cases of $P' = 0$ and $P' = \alpha/2$, as well as the case of independence in which $P' = P(2p_Y \leq \alpha) \times (p_X > \alpha)$. Other than the extreme case of $P' = 0$, in which a significant Y test result deterministically ensures that the X test is also significant, the false positive rate of the overall procedure, $P(p \leq \alpha)$, will be greater than α. Thus the procedure is not valid.

The approach for the sequential permute-match test (Fig 1F) is different because it is based on events, not continuous random variables. Perhaps counterintuitively, we can actually get away with a similar recipe while producing a valid p-value. To best compare our approach to the flawed procedure above, we presently leave out the permutation part and focus only on the matching test component in the following proposition:

Proposition 13

Let M_X and M_Y be events with $P(M_X) \leq 1/n^n$ and $P(M_Y) \leq 1/n^n$. Let

$$p = \begin{cases} 1/n^n & \text{if } M_X \text{ occurs} \\ 2/n^n & \text{if } M_Y \text{ occurs, but not } M_X \\ 1 & \text{neither occurs} \end{cases} \tag{A-12}$$

Then, $P(p \leq \alpha) \leq \alpha$ for $0 \leq \alpha \leq 1$ and for all integers $n \geq 2$.

Proof: We consider four cases, each corresponding to possible values of α.

First, if $\alpha = 1$, then $P(p \leq \alpha) \leq \alpha$ trivially because probability does not exceed 1.

Second is the case $1 > \alpha \geq 2/n^n$. In this case, either M_X or M_Y is sufficient to achieve $p \leq \alpha$. So we have:

$$\begin{aligned} P(p \leq \alpha) &= P(M_X \text{ or } M_Y) \\ &\leq P(M_X) + P(M_Y) \leq 2/n^n \leq \alpha \end{aligned}$$

as required.

Third, consider the case $2/n^n > \alpha \geq 1/n^n$. In this case, only M_X will give us $p \leq \alpha$. So we have

$$P(p \leq \alpha) = P(M_X) \leq 1/n^n \leq \alpha$$

as required.

Finally consider the case $\alpha < 1/n^n$. In this case $p \leq \alpha$ is impossible so $P(p \leq \alpha) = 0 \leq \alpha$, as required. Since the claim holds under our (exhaustive) set of cases, the proof is complete.

The same argument holds if $1/n^n$ and $2/n^n$ is replaced q and $2q$ for an arbitrary q between 0 and 1, but we use $1/n^n$ and $2/n^n$ to highlight the connection to the permute-match test. Our complete proof full (Theorem 10) followed largely the same steps as shown here. Overall, we hope that readers of this section now have a deeper understanding of why the sequential permute-match test is valid, and an appreciation of why similar approaches are invalid in the more standard context where statistical tests are not simply binary event checks.

4 Illustration of permute-match test with logistic map system

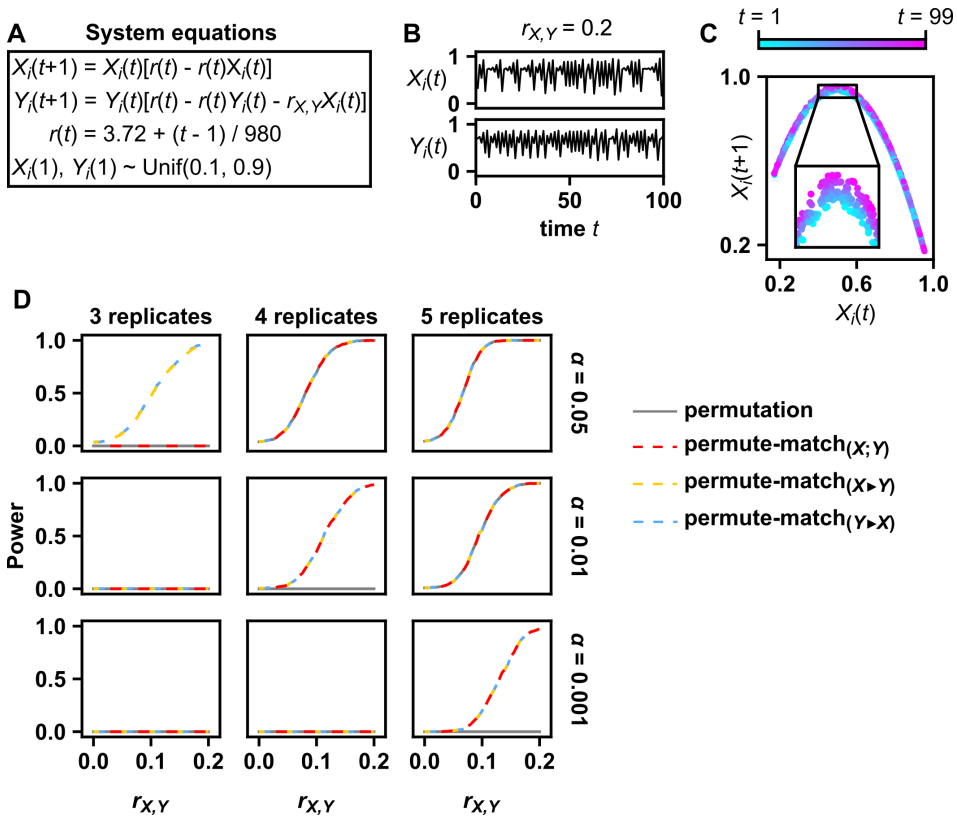


Figure A-2: Statistical power of permute-match tests and the permutation test in a nonlinear and nonstationary system. (**A**) System equations. (**B**) Example dynamics. The processes X and Y are given by a coupled logistic map on a time grid $t = 1, 2, ..., 100$. The system is nonstationary because the parameter $r(t)$ varies with time from 3.72 when $t = 1$ to 3.82 when $t = 99$. To see the nonstationarity clearly, we plot $X_i(t)$ against $X_i(t+1)$ and color points by time, showing that as time progresses, 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. (**D**) Statistical power of the permutation test and permute-match tests as a function of the number of replicates, the significance level, and $r_{X,Y}$. Power was calculated from 5000 simulations at each value of $r_{X,Y}$ between $r_{X,Y} = 0$ and $r_{X,Y} = 0.2$ in steps of size 0.005. The correlation statistic (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].