Statistical resampling of time series by permutation

San To Chan [*] and Eliot Fried [†]
Okinawa Institute of Science and Technology Graduate University, Onna, Okinawa 904-0495, Japan

(Received 18 May 2026; accepted 14 October 2026; published 17 November 2026)

We propose a statistical resampling scheme that generates an ensemble of permuted time series by repeatedly swapping randomly selected pairs of adjacent data points. In the infinite-ensemble limit, we derive the exact ensemble-average operator, which describes how the ensemble average evolves with the number of swaps and satisfies a discrete diffusion equation. Under a periodic idealization, the corresponding transfer function reduces asymptotically to the Gaussian response in the low-frequency limit when the number of swaps is on the order of the data length. The ensemble-average operator also affords a dimensionless relation involving the residue between the original and ensemble-averaged time series, the magnitude of the first difference, the number of swaps, and the data length. When the time series is dominated by white noise, this relation reduces to a similarity function that depends solely on the normalized number of swaps. This result motivates a procedure based on the Kolmogorov-Smirnov test for selecting the extent of permutation according to approximate preservation of the marginal first-difference distribution. Under the periodic idealization, the characteristic function of the random displacement of an individual data value is identical to the transfer function of the ensemble-average operator, thereby providing a destructive interference interpretation of the operator's low-pass behavior. Finally, the individual permuted time series are demonstrated as surrogate inputs to downstream analyses in a proof-of-concept application using empirical mode decomposition.

DOI: 10.1103/g6fz-4fsh

I. INTRODUCTION

The Gaussian moving average (GMA) filter, or Gaussian filter, is one of the most widely used low-pass filters [1–5]. For a time series $x(t)$ of length N with discrete Fourier transform $X(f)$, the filter's response function in the frequency domain is

$$R_G(f) = \exp\left(-2\left(\frac{\pi\sigma}{N}f\right)^2\right), \tag{1}$$

where σ is the smoothing parameter. The characteristic frequency of the filter is defined as $f_c = N/(2\pi\sigma)$. Components with $f \gg f_c$ are exponentially attenuated, while those with $f \ll f_c$ are largely preserved. The degree of smoothing is controlled by f_c. For $x(t) = s(t) + n(t)$, where $s(t)$ is a correlated signal and $n(t)$ is uncorrelated noise, f_c is typically chosen to exceed the characteristic frequencies of $s(t)$ to avoid attenuating the signal.

Smoothing by the Gaussian filter is typically described as a convolution operation. Here, we show that it can alternatively be interpreted as the average of an ensemble of permuted time series, analogous to averaging over repeated experimental measurements. The permuted time series are constructed by swapping randomly selected adjacent data points in the original time series a finite number of times. Such an ensemble is referred to as the synthetic ensemble (SE), with its average referred to as the synthetic ensemble average (SEA).

In the infinite-ensemble limit, the evolution of SEA with the number of swaps can be described exactly by an ensemble-average operator. The resulting evolution follows a discrete diffusion equation with reflecting boundary conditions. Under a periodic idealization, at low frequencies and when the number of swaps is on the order of the data length, its response reduces asymptotically to that of the Gaussian filter. In this sense, SEA provides a more fundamental description of smoothing within the framework of SE, with the Gaussian response emerging as a limiting behavior.

II. PRELIMINARIES

A. Generation of the synthetic ensemble

Given a discrete time series $x(t)$ of length N, permuted time series are produced by randomly swapping adjacent data points. A single swap exchanges the values of $x_i = x(t_i)$ and $x_{i+1} = x(t_{i+1})$ at a randomly chosen index $i \in [1, N-1]$. Repeating this operation ν times produces a permuted time series $x_m(t)$.

Conditional on $x(t)$, $M \geqslant 0.1N$ independent permuted time series are generated to characterize trends in ensemble statistics, forming the SE

$$\{x_{\mathrm{SE}}(t)\} = \{x_m(t)\}_{m=1}^{M}.$$

[*]Contact author: san.chan@oist.jp
[†]Contact author: eliot.fried@oist.jp

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

For any downstream analysis, a larger SE with $M \geqslant 10N$ is used instead. The SEA of $x(t)$ is defined as

$$x_{\mathrm{SEA}}(t) = \langle x_{\mathrm{SE}}(t) \rangle = \frac{1}{M} \sum_{m=1}^{M} x_m(t).$$

A MATLAB implementation of the SE generation is provided in the Supplemental Material [6].

B. Ensemble-average operator

For the analytical treatment of the adjacent-swap process, it is convenient to represent the discrete time series $x(t)$ in vector form as

$$\mathbf{x} = (x_1, \ldots, x_N)^\top.$$

The notations $x(t)$ and $\mathbf{x}$ will be used interchangeably throughout this work, as appropriate. The vector notation allows each adjacent swap to be represented as a linear operator. For each adjacent pair of indices, $(i, i+1)$, with $i \in \{1, \ldots, N-1\}$, define the swap operator

$$S_i = I - (\mathbf{e}_{i+1} - \mathbf{e}_i)(\mathbf{e}_{i+1} - \mathbf{e}_i)^\top, \tag{2}$$

where $\{\mathbf{e}_i; i = 1, \ldots, N\}$ is the natural orthonormal basis. One SE realization of ν swaps can be expressed as

$$\mathbf{x}_{\mathrm{SE}} = S_{I_\nu} \cdots S_{I_1} \mathbf{x},$$

where $I_1, \ldots, I_\nu$ are independent random variables, each uniformly distributed on $\{1, \ldots, N-1\}$. In the infinite-ensemble limit $M \to \infty$, the SEA is

$$\mathbf{x}^{\infty}_{\mathrm{SEA}} = \mathbb{E}[\mathbf{x}_{\mathrm{SE}}] = A_N^\nu \mathbf{x}, \tag{3}$$

where

$$A_N = \frac{1}{N-1} \sum_{i=1}^{N-1} S_i \tag{4}$$

is the one-step ensemble-average operator.

Defining the first-difference operator D such that $(Dx)_i = x_{i+1} - x_i$, we obtain the graph Laplacian $G_N = D^\top D$. Summing Eq. (2) over the allowed adjacent pairs gives

$$\begin{aligned} A_N &= I - \frac{1}{N-1} \sum_{i=1}^{N-1} (\mathbf{e}_{i+1} - \mathbf{e}_i)(\mathbf{e}_{i+1} - \mathbf{e}_i)^\top \\ &= I - \frac{1}{N-1} D^\top D \\ &= I - \frac{1}{N-1} G_N. \end{aligned}$$

Consequently, Eq. (3) can be rewritten as

$$\mathbf{x}^{\infty}_{\mathrm{SEA}} = \left(I - \frac{1}{N-1} G_N \right)^\nu \mathbf{x}, \tag{5}$$

in view of which $\mathbf{x}^{\infty}_{\mathrm{SEA}}$ satisfies the following initial-value problem for the exact finite-N discrete diffusion equation

$$\mathbf{x}^{\infty,(k+1)}_{\mathrm{SEA}} - \mathbf{x}^{\infty,(k)}_{\mathrm{SEA}} = -\frac{1}{N-1} G_N \mathbf{x}^{\infty,(k)}_{\mathrm{SEA}}, \quad \mathbf{x}^{\infty,(0)}_{\mathrm{SEA}} = \mathbf{x}.$$

For the nonperiodic adjacent-swap rule used in the SE construction, G_N is the path-graph Laplacian, defined such that

$$(G_N \mathbf{x})_i = \begin{cases} x_1 - x_2, & i = 1, \\ 2x_i - x_{i-1} - x_{i+1}, & 2 \leqslant i \leqslant N-1, \\ x_N - x_{N-1}, & i = N. \end{cases} \tag{6}$$

In the stochastic description, the end points are reflecting: The first data value has no exterior neighbor on the left, and the last data value has no exterior neighbor on the right. The corresponding eigenvectors are the standing-wave cosine modes

$$\begin{aligned} \phi_0(i) &= \frac{1}{\sqrt{N}}, \\ \phi_f(i) &= \sqrt{\frac{2}{N}} \cos\left(\frac{f\pi}{N} \left(i - \frac{1}{2} \right) \right), \quad f = 1, \ldots, N-1, \end{aligned} \tag{7}$$

with eigenvalues

$$\mu_f = 4 \sin^2\left(\frac{\pi f}{2N} \right), \quad f = 0, \ldots, N-1. \tag{8}$$

These eigenvalues, together with the ensemble-average operator A_N, are central to most analytical derivations of SEA in this work.

Incidentally, Eq. (5) also provides a direct means of computing the SEA in the infinite-ensemble limit without explicitly generating an ensemble of permuted time series. Since A_N is tridiagonal, each application of the operator requires $O(N)$ operations. The computational complexity of Eq. (5) is therefore $O(\nu N)$. By comparison, generating M finite-SE realizations requires $O(M\nu)$ operations, while calculating their ensemble average additionally requires $O(MN)$ operations. The computational complexity of calculating a finite-M SEA is therefore $O(M(\nu + N))$. Evaluation of the SEA through the ensemble-average operator A_N can hence avoid the computational costs associated with generating and averaging a finite ensemble once the number of swaps ν has been specified. Nonetheless, it is important to note that A_N itself does not provide a criterion for identifying a suitable value of ν, which would require examining how the statistical properties of a finite SE change as each of its members is increasingly permuted.

C. Finite-ensemble effect

Most of the analytical formulas in this work refer to the infinite-ensemble limit $M \to \infty$, in which

$$\mathbf{x}^{\infty}_{\mathrm{SEA}} = \mathbb{E}[\mathbf{x}_{\mathrm{SE}}] = \mathbb{E}[\mathbf{x}_{\mathrm{SEA}}]. \tag{9}$$

Then,

$$\begin{aligned} \mathbb{E}[\|\mathbf{x}^{\infty}_{\mathrm{SEA}} - \mathbf{x}_{\mathrm{SEA}}\|^2] &= \frac{1}{M} \mathrm{Tr}\, \mathrm{Cov}(\mathbf{x}_{\mathrm{SE}}) \\ &= \frac{1}{M} \mathbb{E}[\|\mathbf{x}^{\infty}_{\mathrm{SEA}} - \mathbf{x}_{\mathrm{SE}}\|^2]. \end{aligned}$$

Each SE permuted time series is a permutation of $\mathbf{x}$, so $\|\mathbf{x}_{\mathrm{SE}}\| = \|\mathbf{x}\|$. Hence,

$$\mathbb{E}\left[\frac{1}{N} \|\mathbf{x}^{\infty}_{\mathrm{SEA}} - \mathbf{x}_{\mathrm{SEA}}\|^2 \right] \leqslant \frac{\mathrm{rms}(\mathbf{x})^2}{M}. \tag{10}$$

Finite-ensemble deviations therefore scale as $M^{-1/2}$ in root-mean-square (rms) norm.

D. Cross-correlation change induced by a single swap

Consider a discrete time series $\mathbf{x}$ and its SE $\{\mathbf{x}_{\mathrm{SE}}\}$. For the mth permuted time series, a single swap exchanges the values of x_i and x_{i+1} at a randomly chosen index $i \in [1, N-1]$:

$$x_{m,j} = \begin{cases} x_{i+1}, & j = i, \\ x_i, & j = i+1, \\ x_j, & \text{otherwise.} \end{cases}$$

Define the lag-k autocorrelation of $x(t)$ by

$$R_0(k) = \frac{1}{N-k} \sum_{j=1}^{N-k} x_j x_{j+k},$$

where k denotes the discrete lag. Define the cross-correlation between $x(t)$ and $x_m(t)$ by

$$R_m(k) = \frac{1}{N-k} \sum_{j=1}^{N-k} x_j x_{m,j+k}.$$

Suppose that a single swap exchanges the adjacent values at positions $(i, i+1)$ in x_m. For $1 \leqslant k \leqslant N-1$,

$$R_0(k) - R_m(k) = \frac{1}{N-k} \begin{cases} 0, & i < k, \\ x_1 \Delta x_k, & i = k, \\ \Delta x_{i-k} \Delta x_i, & i \geqslant k+1. \end{cases}$$

Averaging uniformly over the swap index gives

$$\mathbb{E}_i[R_0(k) - R_m(k)] = \frac{x_1 \Delta x_k + \sum_{i=k+1}^{N-1} \Delta x_{i-k} \Delta x_i}{(N-1)(N-k)}.$$

Hence, apart from the end-point contribution $x_1 \Delta x_k$, the expected change in the lag-k cross-correlation produced by a single swap is determined by the corresponding sum of lagged products of the first-difference sequence. Consequently, when the end-point contribution and the sum are negligible for the lags of interest, the expected changes in the corresponding cross-correlations are also negligible.

E. Two-sample Kolmogorov-Smirnov test

The single-swap result above identifies the first differences as relevant local quantities for characterizing changes induced by adjacent swapping. We therefore use the two-sample Kolmogorov-Smirnov (KS) statistic [7–10] as a diagnostic for changes in their marginal empirical distribution. The KS statistic does not measure independence or serial correlation.

The test compares the empirical cumulative distribution functions (CDFs) $\hat{F}_1(u)$ and $\hat{F}_2(u)$ of two data vectors, here Δx and $\{\Delta x_{\mathrm{SE}}\}$. The two-sided test statistic is

$$D_{\mathrm{KS}} = \max_u |\hat{F}_1(u) - \hat{F}_2(u)|.$$

For two independent samples, the null hypothesis $H_0 : F_1(u) = F_2(u)$ yields the standard KS reference distribution and its associated p value. In the present setting, however, adjacent first differences in a time series are generally dependent, different SE permuted time series are constructed from the same underlying data, and the pooled sample $\{\Delta x_{\mathrm{SE}}\}$ is therefore not independent in the textbook sense. Accordingly, the reported $p_{\mathrm{KS}} = p_{\mathrm{KS}}(\nu)$ is interpreted as an operational diagnostic score rather than as a strictly calibrated probability.

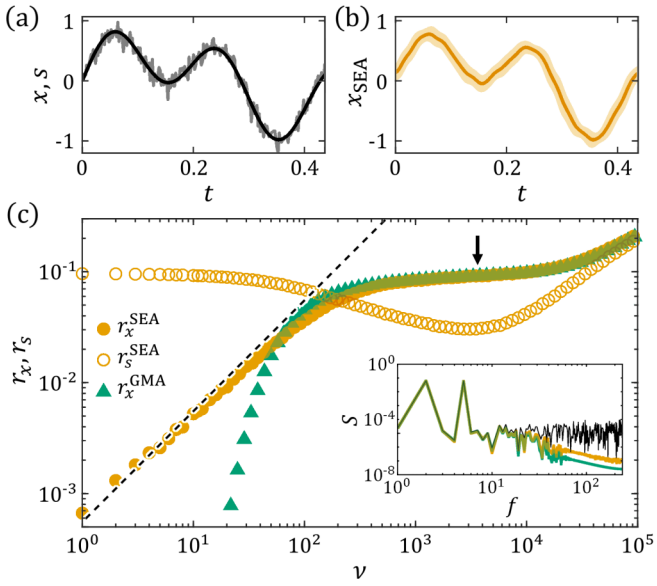


FIG. 1. Low-pass filtering by SEA. (a) Example time series $x(t) = s(t) + n(t)$ of length $N = 500$ (gray line), with ground truth $s(t) = 0.5\sin(4\pi t) + 0.5\sin(10\pi t)$ (black line) and added noise $n(t) = 0.1 n_G(t)$, where $n_G(t) \sim \mathcal{N}(0, 1)$. $x(t)$ is defined for $0 \leqslant t \leqslant 1$; only a portion of it is shown for clarity. (b) SEA result with ensemble mean $x_{\mathrm{SEA}}(t)$ (orange line) bounded by one standard deviation (shaded orange band). The averaging is performed with 25 000 SE permuted time series, each with a number of swaps $\nu = 3500$. (c) Residues $r_x = \mathrm{rms}(y - x)$ and $r_s = \mathrm{rms}(y - s)$ plotted against ν, for $y(t)$ obtained using SEA (orange circles) and the GMA filter (green triangles). For GMA, the smoothing width is defined as $\sigma^2 = 2\nu/N$. The dashed line shows the scaling relation (19). The inset compares the power spectra of the original time series $x(t)$ (black), SEA-filtered signal (orange), and GMA-filtered signal (green), for $\nu = 3500$ (arrow).

For a prescribed threshold p_c, we define the critical number of swaps, ν_c, as the largest tested value of ν satisfying $p_{\mathrm{KS}}(\nu) > p_c$. Unless stated otherwise, p_{KS} is evaluated at 80 logarithmically spaced values of ν over $10^{-2}N \leqslant \nu \leqslant 10^2 N$, rounded to the nearest integer. If the threshold is crossed multiple times, the largest tested value satisfying the criterion is selected. In the examples considered in this work, multiple crossings are observed primarily for intermediate values of p_{KS} and become less frequent as the ensemble size M increases. If no tested value satisfies the criterion within the search range, ν_c is undefined.

III. RESULTS AND DISCUSSION

A. SEA-GMA correspondence

The SEA exhibits low-pass behavior similar to a Gaussian filter. Figure 1(a) shows a time series $x(t) = s(t) + n(t)$, where the ground-truth signal $s(t) = 0.5\sin(4\pi t) + 0.5\sin(10\pi t)$ is contaminated by noise $n(t) = 0.1 n_G(t)$, $n_G(t) \sim \mathcal{N}(0, 1)$. Unless stated otherwise, all time series considered in this study have length $N = 500$ and are defined on $0 \leqslant t \leqslant 1$.

Figure 1(b) displays the SEA result $x_{\mathrm{SEA}}(t)$ (orange line), obtained with 25 000 SE permuted time series, each with a number of swaps $\nu = 3500$. The filtered signal successfully recovers the global trend of $s(t)$ while suppressing the small-scale fluctuations introduced by $n(t)$.

To evaluate smoothing performance, we define the root-mean-square residues between the filtered signal $y(t)$ and either the ground truth $s(t)$ or the original time series $x(t)$ as

$$r_s = \mathrm{rms}(y - s), \quad r_x = \mathrm{rms}(y - x). \tag{11}$$

Figure 1(c) shows that the residue relative to the ground truth, r_s^{SEA}, plateaus for small ν and reaches a minimum near $\nu \approx 3000$. In practice, however, the ground truth $s(t)$ is unknown and only $x(t)$ is observable. The residue r_x^{SEA} increases approximately linearly with ν for small ν and levels off around $\nu \approx 3000$.

For comparison, the residue obtained using the Gaussian filter, r_x^{GMA}, is also shown. When the smoothing parameter σ is set as

$$\sigma = \sqrt{\frac{2\nu}{N}}, \tag{12}$$

the residues r_x^{SEA} and r_x^{GMA} approach one another as ν increases and approximately coincide for $\nu > N = 500$. The inset of Fig. 1 compares the power spectra of the original, SEA-filtered, and GMA-filtered time series. Both SEA and the Gaussian filter exhibit low-pass behavior and their spectra share the same overall shape. High-frequency components are attenuated, and low-frequency components are preserved.

The numerical observations above suggest a correspondence between SEA and the Gaussian filter, which can be analytically derived. Recall that the ensemble-average operator A_N [cf. Eq. (4)] governs how SEA evolves with the number of swaps. To compare with the Fourier response of the Gaussian filter, we now assume periodic boundary conditions, such that an additional swap between x_1 and x_N is enabled and there are N allowed adjacent pairs instead of $N - 1$. Under periodic boundary conditions, G_N has discrete Fourier eigenvectors with wave numbers $2\pi f/N$ and eigenvalues

$$\mu_f = 4\sin^2\left(\frac{\pi f}{N}\right), \quad f = 0, 1, \ldots, N-1.$$

The corresponding ensemble-average operator is

$$A_N^p = I - \frac{1}{N}G_N.$$

Hence, the exact transfer function is

$$H_f^p(\nu) = \left(1 - \frac{4}{N}\sin^2\left(\frac{\pi f}{N}\right)\right)^{\nu}. \tag{13}$$

For $f \ll N$ and $\nu = O(N)$, the transfer function has the asymptotic form

$$H_f^p(\nu) \approx \exp\left(-4\pi^2 \frac{\nu}{N}\frac{f^2}{N^2}\right).$$

Using Eq. (12), this becomes

$$H_f^p(\nu) \approx \exp\left(-2\left(\frac{\pi\sigma f}{N}\right)^2\right),$$

which is the Gaussian response function R_G [cf. Eq. (1)]. Its characteristic frequency $f_c = N/(2\pi\sigma)$ can be expressed as

$$\frac{2\pi f_c}{N} = \left(\frac{2\nu}{N}\right)^{-1/2}.$$

This relation clarifies the physical meaning of the number of swaps ν. Increasing it decreases f_c; the range of frequencies attenuated by SEA extends toward lower frequencies.

B. Sparse-swap asymptotic of SEA

The linear behavior of r_x^{SEA} for $\nu/N \ll 1$ observed in Fig. 1(c) can also be derived from the ensemble-average operator A_N. From Eqs. (5) and (11),

$$(r_x^{\mathrm{SEA}})^2 = \frac{1}{N}\left\|\left(\left(I - \frac{1}{N-1}G_N\right)^{\nu} - I\right)\mathbf{x}\right\|^2. \tag{14}$$

Similarly, the mean-square first difference of $\mathbf{x}$ is

$$\mathrm{rms}(\Delta x)^2 = \frac{1}{N-1}\|D\mathbf{x}\|^2. \tag{15}$$

Define $c_N = 1/(N-1)$. Since the eigenvalues of G_N given by Eq. (8) satisfy $0 \leqslant \mu_f \leqslant 4$,

$$(1 - c_N\mu_f)^{\nu} = 1 - \nu c_N\mu_f + O\big((\nu c_N)^2\mu_f^2\big)$$

uniformly in f whenever $\nu/N \ll 1$. Therefore, by the spectral theorem,

$$(I - c_N G_N)^{\nu} - I = -\nu c_N G_N + O\big((\nu c_N)^2 G_N^2\big). \tag{16}$$

Substituting Eq. (16) into Eq. (14) and dividing by Eq. (15) yields

$$\frac{r_x^{\mathrm{SEA}}}{\mathrm{rms}(\Delta x)} = K_{\Delta,N}(\mathbf{x})\frac{\nu}{N} + O\left(\left(\frac{\nu}{N}\right)^2\frac{\|G_N^2\mathbf{x}\|}{\|D\mathbf{x}\|}\right), \tag{17}$$

where, for nonconstant $\mathbf{x}$,

$$K_{\Delta,N}(\mathbf{x}) = \sqrt{\frac{N}{N-1}}\frac{\|G_N\mathbf{x}\|}{\|D\mathbf{x}\|}. \tag{18}$$

For constant $\mathbf{x}$, both r_x^{SEA} and $\mathrm{rms}(\Delta x)$ vanish, leaving the normalized residue undefined.

The dependence of $K_{\Delta,N}$ on first-difference correlations can be made explicit. From the definition of G_N in Eq. (6),

$$\|G_N\mathbf{x}\|^2 = 2\sum_{i=1}^{N-1}(\Delta x_i)^2 - 2\sum_{i=2}^{N-1}\Delta x_{i-1}\Delta x_i.$$

Therefore,

$$K_{\Delta,N}^2(\mathbf{x}) = \frac{2N}{N-1}\left(1 - \frac{\sum_{i=2}^{N-1}\Delta x_{i-1}\Delta x_i}{\sum_{i=1}^{N-1}(\Delta x_i)^2}\right).$$

If the first-difference sequence is stationary and ergodic with finite second moment, then, in the large-N limit,

$$\frac{1}{N-1}\sum_{i=1}^{N-1}(\Delta x_i)^2 \to \gamma_0 = \mathbb{E}[(\Delta x_i)^2]$$

and

$$\frac{1}{N-1}\sum_{i=2}^{N-1}\Delta x_{i-1}\Delta x_i \to \gamma_1 = \mathbb{E}[\Delta x_i \Delta x_{i-1}].$$

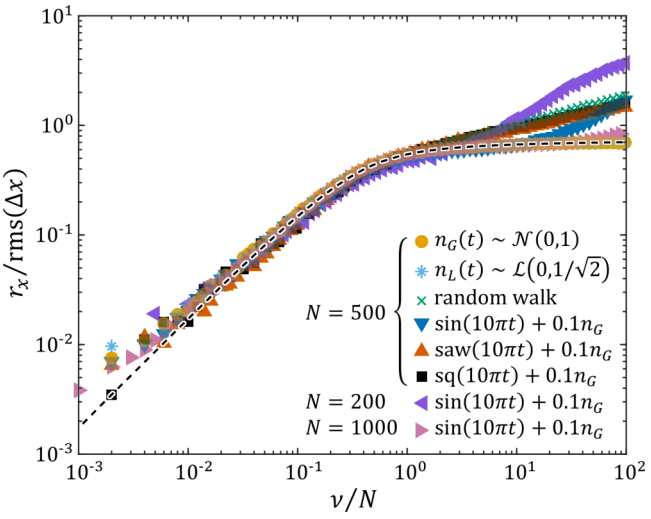


FIG. 2. SEA residue across signal and noise types. Normalized residue $r_x^{\mathrm{SEA}}/\mathrm{rms}(\Delta x)$ plotted against the normalized number of swaps ν/N for diverse synthetic time series $x(t)$, including Gaussian and Laplacian noise, random walk, and noisy sine, sawtooth, and square waves. All time series have length $N = 500$, except for the noisy sine wave, where cases with $N = 200$ and 1000 are also shown. The dashed line indicates the Gaussian similarity function (27).

Hence,

$$K_{\Delta,N}(\mathbf{x}) \to \sqrt{2\left(1 - \frac{\gamma_1}{\gamma_0}\right)}.$$

For independent and identically distributed first differences with mean m and variance $v > 0$,

$$\gamma_0 = v + m^2, \qquad \gamma_1 = m^2,$$

so that

$$K_{\Delta,N}(\mathbf{x}) \to \sqrt{\frac{2v}{v + m^2}}.$$

In particular, independent zero-mean first differences yield $K_{\Delta,N} \to \sqrt{2}$, whereas independent white-noise values yield $K_{\Delta,N} \to \sqrt{3}$, since their first differences satisfy $\gamma_1/\gamma_0 = -1/2$. An alternating high-frequency sequence gives $K_{\Delta,N} \approx 2$.

More generally, since $G_N = D^\top D$ and $\|D\| \leqslant 2$,

$$\|G_N\mathbf{x}\| \leqslant 2\|D\mathbf{x}\|,$$

and therefore

$$0 \leqslant K_{\Delta,N}(\mathbf{x}) \leqslant 2\sqrt{\frac{N}{N-1}} \approx 2.$$

With Eq. (17), this gives the maximal asymptotic scaling

$$\frac{r_x^{\mathrm{SEA}}}{\mathrm{rms}(\Delta x)} \approx 2\frac{\nu}{N}. \tag{19}$$

C. Dimensionless similarity of SEA

The sparse-swap asymptotic predicts a similarity relation between the normalized SEA residue $r_x^{\mathrm{SEA}}/\mathrm{rms}(\Delta x)$ and the normalized number of swaps ν/N for $\nu/N \ll 1$. Figure 2 shows the residues for a range of synthetic time series, including Gaussian noise, Laplacian noise, random walk, and noisy sine, sawtooth, and square waves. After normalization, all data appear to collapse onto the Gaussian-noise case for $\nu/N = O(1)$, pointing to a similarity relation extending beyond the sparse-swap regime:

$$\frac{r_x^{\mathrm{SEA}}}{\mathrm{rms}(\Delta x)} = \Pi\left(\frac{\nu}{N}\right). \tag{20}$$

This numerical observation motivates an analytical treatment of the Gaussian-noise case. Expanding the time series $\mathbf{x}$ in the eigenbasis $\{\phi_f\}$ given by Eq. (7),

$$\mathbf{x} = \sum_{f=0}^{N-1} \hat{x}_f \phi_f,$$

Eq. (14) can be expressed as

$$\left(r_x^{\mathrm{SEA}}\right)^2 = \frac{1}{N}\sum_{f=1}^{N-1}\left[1 - \left(1 - \frac{\mu_f}{N-1}\right)^\nu\right]^2 |\hat{x}_f|^2. \tag{21}$$

Similarly, Eq. (15) can be expressed as

$$\mathrm{rms}(\Delta x)^2 = \frac{1}{N-1}\sum_{f=1}^{N-1} \mu_f |\hat{x}_f|^2. \tag{22}$$

The constant mode does not contribute to either the first difference or the SEA residue, and is therefore omitted from the sums in Eqs. (21) and (22). Combining Eqs. (21) and (22) and approximating $N - 1$ as N for large N yields the normalized residue relation

$$\left(\frac{r_x^{\mathrm{SEA}}}{\mathrm{rms}(\Delta x)}\right)^2 \approx \frac{\sum_{f=1}^{N-1}\left[1 - \left(1 - \frac{\mu_f}{N}\right)^\nu\right]^2 |\hat{x}_f|^2}{\sum_{f=1}^{N-1} \mu_f |\hat{x}_f|^2}.$$

For $\nu = O(N)$,

$$\left(\frac{r_x^{\mathrm{SEA}}}{\mathrm{rms}(\Delta x)}\right)^2 \approx \frac{\sum_{f=1}^{N-1}\left[1 - \exp\left(-\frac{\mu_f \nu}{N}\right)\right]^2 |\hat{x}_f|^2}{\sum_{f=1}^{N-1} \mu_f |\hat{x}_f|^2}. \tag{23}$$

The normalized SEA residue therefore depends not only on ν/N, but also on the spectral energy distribution $|\hat{x}_f|^2$ of the time series.

For Gaussian white noise of zero mean and variance C, the coefficients $\hat{x}_f$ in the orthonormal eigenbasis satisfy $\mathbb{E}[|\hat{x}_f|^2] = C$ for all nonconstant modes. Taking expectations of the squared residue and squared first-difference magnitude gives, for large N and $\nu = O(N)$,

$$\left(\frac{r_x^{\mathrm{SEA}}}{\mathrm{rms}(\Delta x)}\right)^2 \approx \frac{\sum_{f=1}^{N-1}\left[1 - \exp\left(-\frac{\mu_f \nu}{N}\right)\right]^2}{\sum_{f=1}^{N-1} \mu_f} = \frac{Q}{P}, \tag{24}$$

which depends solely on ν/N. Let

$$z = \frac{\nu}{N}, \quad \theta = \frac{\pi f}{N}.$$

Since

$$\mu_f = 4\sin^2\left(\frac{\pi f}{2N}\right) = 2 - 2\cos\theta,$$

the denominator P in Eq. (24) can be approximated as a Riemann sum for large N:

$$P \approx \frac{N}{\pi}\int_0^\pi (2-2\cos\theta)\,d\theta = 2N. \tag{25}$$

Likewise, the numerator Q can be approximated as

$$\begin{aligned} Q &\approx \frac{N}{\pi}\int_0^\pi [1-\exp(-z(2-2\cos\theta))]^2 d\theta \\ &= N[1-2e^{-2z}I_0(2z)+e^{-4z}I_0(4z)], \end{aligned} \tag{26}$$

where $I_0(z)$ is the modified Bessel function of the first kind [11]. Combining Eqs. (24)–(26) then yields the Gaussian similarity function

$$\Pi_G(z) = \left[\frac{1-2e^{-2z}I_0(2z)+e^{-4z}I_0(4z)}{2}\right]^{1/2}. \tag{27}$$

The sparse-swap asymptotic is recovered for $z \ll 1$. Since

$$\begin{aligned} \mathrm{e}^{-2z}I_0(2z) &= 1-2z+3z^2+O(z^3), \\ \mathrm{e}^{-4z}I_0(4z) &= 1-4z+12z^2+O(z^3), \end{aligned}$$

it follows that

$$1-2\mathrm{e}^{-2z}I_0(2z)+\mathrm{e}^{-4z}I_0(4z) = 6z^2+O(z^3),$$

and, thus,

$$\Pi_G(z) = \sqrt{3}\quad z+O(z^2). \tag{28}$$

The coefficient $\sqrt{3}$ is the sparse-swap coefficient for independent Gaussian values. In the examples in Fig. 2, the normalized residues appear to lie close to Π_G for $\nu/N = O(1)$.

D. Departure from the Gaussian similarity function

To understand deviations from the Gaussian similarity function Π_G observed in Fig. 2, consider a noisy time series whose spectral coefficients in the cosine basis can be written as

$$\hat{x}_f = \hat{s}_f + \hat{c}_f,$$

where $\hat{s}_f$ is the structured contribution and $\hat{c}_f$ is a zero-mean Gaussian white noise with variance C. Defining

$$S_f = |\hat{s}_f|^2, \quad W_f(z) = 1-\exp(-\mu_f z),$$

and using the same expectation-level approximation as in the derivation of Π_G, Eq. (23) can be expressed as

$$\left(\frac{r_x^{\mathrm{SEA}}}{\mathrm{rms}(\Delta x)}\right)^2 \approx \frac{C\sum_f W_f^2(z)+\sum_f W_f^2(z)S_f}{C\sum_f \mu_f+\sum_f \mu_f S_f}. \tag{29}$$

Since

$$\Pi_G^2(z) \approx \frac{\sum_f W_f^2(z)}{\sum_f \mu_f},$$

the expression may be rewritten as

$$\left(\frac{r_x^{\mathrm{SEA}}}{\mathrm{rms}(\Delta x)}\right)^2 \approx \Pi_G^2(z)\frac{1+A(z)}{1+B},$$

where

$$A(z) = \frac{\sum_f W_f^2(z)S_f}{C\sum_f W_f^2(z)}, \quad B = \frac{\sum_f \mu_f S_f}{C\sum_f \mu_f}.$$

If $|A(z)| \ll 1$ and $|B| \ll 1$, linearization yields

$$\left(\frac{r_x^{\mathrm{SEA}}}{\mathrm{rms}(\Delta x)}\right)^2 \approx \Pi_G^2(z) + \frac{\sum_f \left(W_f^2(z)-\Pi_G^2(z)\mu_f\right)S_f}{C\sum_f \mu_f},$$

up to corrections of order $A^2+B^2+|AB|$. For the reflecting spectrum, $\sum_f \mu_f \sim 2N$, so the denominator in the linear correction is $O(NC)$. However, this scaling alone does not guarantee a small correction, since $\sum_f S_f$ may also scale with N. The second term remains small when

$$\left|\sum_f \left(W_f^2(z)-\Pi_G^2(z)\mu_f\right)S_f\right| \ll C\sum_f \mu_f. \tag{30}$$

The linearized expression nevertheless clarifies the origin of the collapse observed in Fig. 2. For $z = O(1)$, a fixed low-frequency mode has $\mu_f = O(N^{-2})$, and therefore

$$W_f(z) = 1-\mathrm{e}^{-z\mu_f} = O(N^{-2}), \quad W_f^2(z) = O(N^{-4}).$$

Thus, the residue weighting strongly suppresses smooth low-frequency structured components for moderate z, causing the normalized residue to be dominated by the approximately Gaussian high-frequency contribution. As z increases, or when the structured energy occupies higher frequencies, the weighting becomes increasingly sensitive to the structured spectral contribution S_f. Condition (30) may then no longer hold, and deviations from Π_G become appreciable. In practice, the similarity relation (27) can be expected when several conditions are satisfied: the ensemble size M and data length N are large; the spectral energy of the additive noise can be modeled as uniformly distributed; and the weighted structured spectral contribution remains negligible relative to the noise background. Since the collapse observed in Fig. 2 is visual, here we examine the sensitivity of the similarity relation to those conditions via several quantitative stress tests presented in Fig. 3.

To quantify the agreement between the normalized residue and the Gaussian similarity function, we define the relative deviation

$$\delta_G = \frac{1}{\Pi_G}\left(\frac{r_x^{\mathrm{SEA}}}{\mathrm{rms}(\Delta x)}-\Pi_G\right). \tag{31}$$

Here, $\delta_G = 0$ corresponds to exact agreement with Π_G, while $|\delta_G| < 0.05$ corresponds to a relative deviation of less than 5%. Figure 3 shows δ_G for noisy sine waves of the form $x(t) = 0.5\sin(10\pi t)+\alpha n_G(t)$, $n_G(t) \sim \mathcal{N}(0,1)$. Unless stated otherwise, data length $N = 500$, noise amplitude $\alpha = 0.1$, and ensemble size $M = 100$. Each curve shows the mean over ten independent realizations, with both the additive noise and the SE regenerated independently for each realization, while the shaded region indicates the corresponding minimum and maximum values.

Figure 3(a) shows the effect of the ensemble size M. As M increases, the normalized residue approaches the infinite-ensemble result. For sufficiently large M, agreement with Π_G

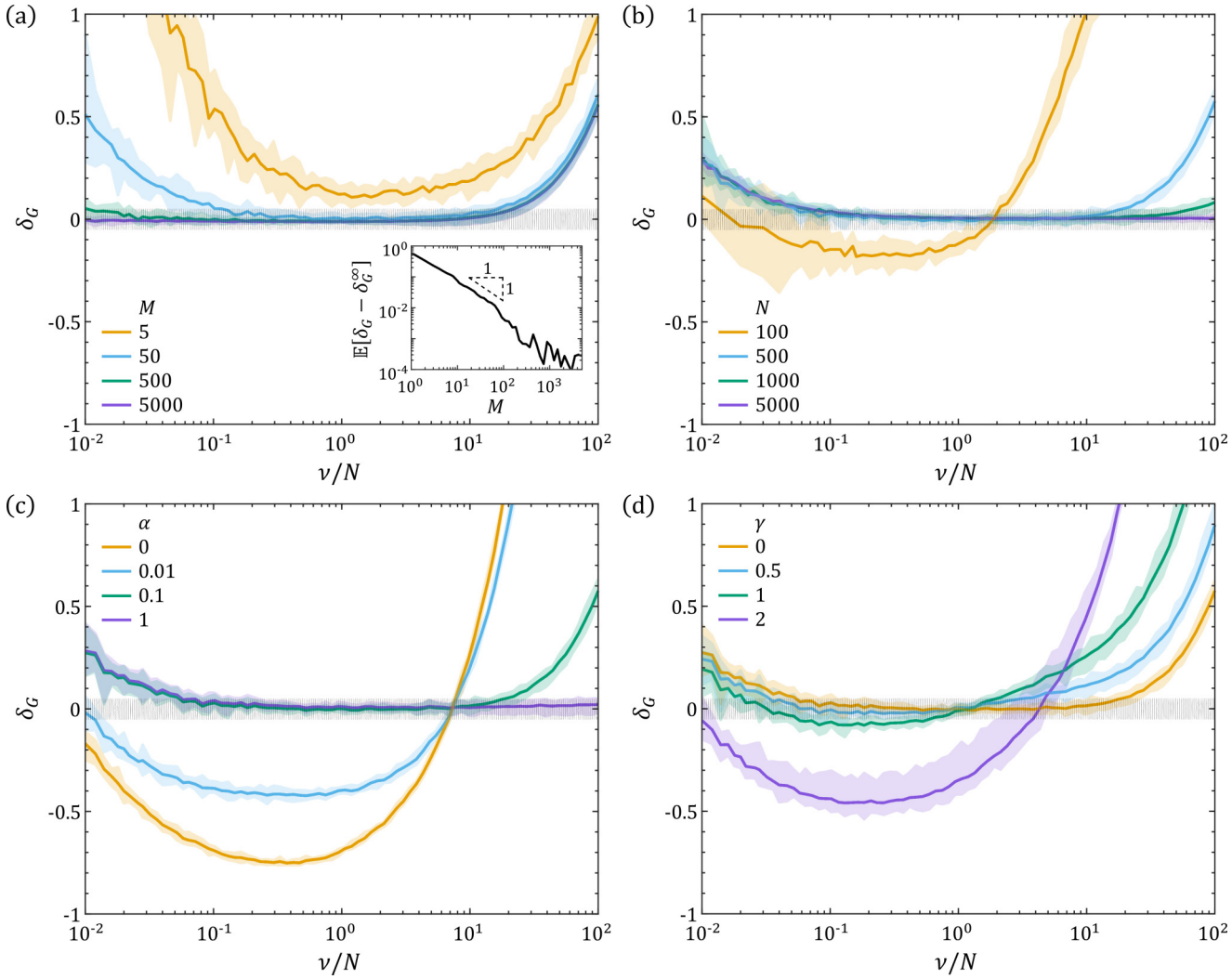


FIG. 3. Departure from the Gaussian similarity function. The relative deviation $\delta_G = \Pi_G^{-1}[r_x^{\mathrm{SEA}}/\,\mathrm{rms}(\Delta x) - \Pi_G]$ as a function of the normalized number of swaps ν/N for a noisy sine wave $x(t) = 0.5\sin(10\pi t) + \alpha n_G(t)$. Unless stated otherwise, $n_G(t) \sim \mathcal{N}(0, 1)$, data length $N = 500$, noise amplitude $\alpha = 0.1$, and ensemble size $M = 100$. Each curve shows the mean over ten independent realizations, and the shaded region indicates the corresponding minimum and maximum values. (a) Dependence on M. The inset shows $|\mathbb{E}[\delta_G - \delta_G^\infty]|$ as a function of M, evaluated at $\nu/N = 1$; δ_G^∞ is the relative deviation calculated assuming $M \to \infty$. (b) Dependence on N. (c) Dependence on α. (d) Dependence on the spectral exponent $\gamma = 0, 0.5, 1, 2$ for an additive $1/f^\gamma$ noise. The gray band indicates $|\delta_G| < 0.05$.

can be observed for $\nu/N \lesssim 10$, whereas the deviation from Π_G persists for large $\nu/N \gtrsim 10$.

The dependence on M can be deduced as follows. Adding and subtracting $\mathbf{x}_{\mathrm{SEA}}^\infty$ in the SEA residue gives

$$\|\mathbf{x}_{\mathrm{SEA}} - \mathbf{x}\| = \|(\mathbf{x}_{\mathrm{SEA}}^\infty - \mathbf{x}) + (\mathbf{x}_{\mathrm{SEA}} - \mathbf{x}_{\mathrm{SEA}}^\infty)\|.$$

Assuming that $\|\mathbf{x}_{\mathrm{SEA}} - \mathbf{x}_{\mathrm{SEA}}^\infty\| \ll \|\mathbf{x}_{\mathrm{SEA}}^\infty - \mathbf{x}\|$ and expanding about $\mathbf{x}_{\mathrm{SEA}}^\infty - \mathbf{x}$, by Eq. (9), the term linear in $\mathbf{x}_{\mathrm{SEA}} - \mathbf{x}_{\mathrm{SEA}}^\infty$ vanishes upon taking the expected value. The leading correction term to the mean residue is therefore quadratic in $\mathbf{x}_{\mathrm{SEA}} - \mathbf{x}_{\mathrm{SEA}}^\infty$. Equation (10) then gives $\mathbb{E}[\delta_G - \delta_G^\infty] = O(M^{-1})$, where δ_G^∞ is obtained from $\mathbf{x}_{\mathrm{SEA}}^\infty$. The inset of Fig. 3(a) shows $|\mathbb{E}[\delta_G - \delta_G^\infty]|$ at $\nu/N = 1$ as a function of M, showing an apparent M^{-1} scaling. Therefore, the disagreement with Π_G at small ν/N observed in Fig. 3(a) can be attributed to the finite-ensemble size.

Figure 3(b) shows the effect of the data length N, whose increase mainly affects the normalized residue at large ν/N. For sufficiently large N, the deviation previously observed for $\nu/N \gtrsim 10$ in Fig. 3(a) disappears over the range of ν/N considered. For noisy sine waves, as N increases, the structured spectral contribution is shifted toward smaller normalized frequencies f/N; its effect on the normalized residue is hence suppressed. Therefore, the disagreement with Π_G at large ν/N observed in Fig. 3(a) can be attributed to the structured spectral contribution.

Figure 3(c) examines the effect of additive Gaussian white noise. In the absence of noise ($\alpha = 0$), a substantial deviation from Π_G is observed. The deviation decreases systematically with increasing α. For the largest value $\alpha = 1$, agreement with Π_G can be observed for $\nu/N \gtrsim 10$, similar to Fig. 3(b), although here the behavior results from the increased noise background relative to the unchanged structured spectral

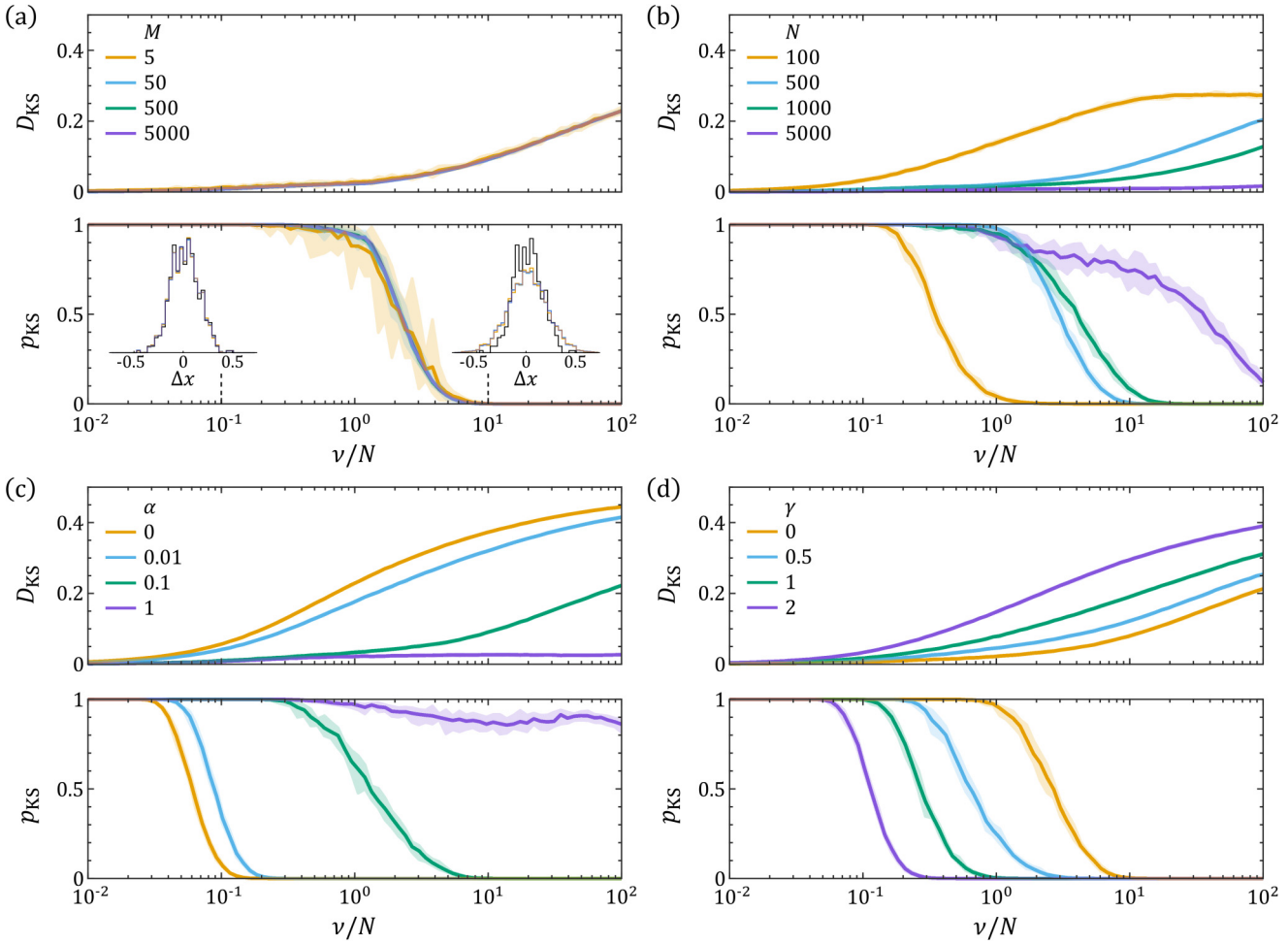


FIG. 4. Sensitivity of the KS operational diagnostic. The KS distance D_{KS} and KS p value p_{KS} as functions of the normalized number of swaps ν/N for a noisy sine wave $x(t) = 0.5\sin(10\pi t) + \alpha n_G(t)$. Unless stated otherwise, $n_G(t) \sim \mathcal{N}(0, 1)$, data length $N = 500$, noise amplitude $\alpha = 0.1$, and ensemble size $M = 100$. For each case, the observed time series is held fixed and the SE is independently regenerated ten times. Each curve shows the mean over the ten ensemble realizations, and the shaded region indicates the corresponding minimum and maximum values. (a) Dependence on M. The insets show the first-difference distribution Δx (black) of the original time series and the pooled distribution $\{\Delta x_{SE}\}$ for the four cases of M at $\nu/N = 0.1$ and 10, with one SE realization shown for each case. (b) Dependence on N. (c) Dependence on α. (d) Dependence on the spectral exponent $\gamma = 0,\ 0.5,\ 1,\ 2$ for an additive $1/f^{\gamma}$ noise.

contribution. Meanwhile, a deviation from Π_G persists for $\nu/N \lesssim 10$, consistent with Fig. 3(a), due to finite-ensemble effects.

Finally, Fig. 3(d) shows the effect of the additive noise's spectra using $1/f^{\gamma}$ noises as examples. For white noise ($\gamma = 0$), agreement with Π_G can be observed over a broad range of ν/N. The finite-ensemble effect at small ν/N and the finite-frequency effect at large ν/N observed in Figs. 3(a) and 3(b), respectively, remain visible. The deviation increases systematically with increasing γ, as the spectral energy of the additive noise becomes increasingly nonuniform. This behavior is consistent with Eqs. (23) and (30), which show that the normalized residue depends on the spectral energy distribution of the time series.

E. Statistical preservation of the first difference

The preceding section establishes that the effect of SE permutation can be characterized by the normalized residue $r_x^{SEA}/\mathrm{rms}(\Delta x)$, whose collapse onto Π_G is observed for $\nu/N = O(1)$, granted that the underlying assumptions are satisfied. However, departures from Π_G can arise from the finite ensemble size M and therefore do not necessarily indicate deterioration due to overswapping. A complementary diagnostic that is less sensitive to M is therefore desirable.

The appearance of $\mathrm{rms}(\Delta x)$ as the natural scale for nondimensionalizing r_x^{SEA} suggests the first difference Δx as the relevant quantity. This is also supported by Sec. II D, where the expected single-swap change in the lag-k cross-correlation is shown to depend on the corresponding sum of lagged products of Δx, modulo an end-point contribution. We therefore employ the KS diagnostic introduced in Sec. II E. Its sensitivity is examined in Fig. 4 using the same noisy sine waves and parameters considered in Fig. 3. For each parameter setting, the observed time series is held fixed while ten SEs are generated independently.

Figure 4(a) shows the effect of the ensemble size M. In contrast to the strong M dependence of the departure from Π_G

in Fig. 3(a), the mean D_{KS} and p_{KS} curves remain close over the tested range of M, while their variability among independently generated SEs increases as M decreases. At $\nu/N = 0.1$, the first-difference distributions $\{\Delta x_{\mathrm{SE}}\}$ remain close to that of Δx for all four values of M, whereas at $\nu/N = 10$ they depart appreciably from the original distribution. Thus, the departure from Π_G at small ν/N is not accompanied by an appreciable change in the marginal first-difference distribution, while the departure at large ν/N is.

To assess the variability of the selected ν_c/N, we determine it separately for each of the ten SEs. For $p_c = 0.05$, the resulting mean and standard deviation are 4.96 ± 0.45, 4.61 ± 0.27, 4.67 ± 0.26, and 4.56 ± 0.28 for $M/N = 0.01$, 0.1, 1, and 10, respectively. For $p_c = 0.95$, the corresponding values are 0.97 ± 0.23, 0.97 ± 0.16, 0.90 ± 0.09, and 0.87 ± 0.05. The standard deviation generally decreases with increasing M, indicating reduced sensitivity of ν_c/N to the particular SE realization. For small relative variations in ν_c/N, the relative variation in the corresponding smoothing parameter σ is approximately half as large, according to Eq. (12).

Figures 4(b)–4(d) show the effects of the data length N, Gaussian-noise amplitude α, and noise spectral exponent γ, respectively. Increasing either N or α shifts the increase in D_{KS} to larger ν/N, whereas increasing γ shifts it toward smaller ν/N; the corresponding transitions in p_{KS} follow the same trends. Increasing N reduces the sample-to-sample contribution of the low-frequency sine wave to the first difference, while increasing α increases the Gaussian-noise contribution, making the first-difference distribution increasingly noise dominated. Conversely, increasing γ makes the additive noise increasingly correlated, resulting in an earlier change in the first-difference distribution under SE permutation. These trends are consistent with the corresponding departures from Π_G in Fig. 3.

The KS diagnostic therefore complements Π_G by indicating whether its departure is accompanied by an appreciable change in the marginal first-difference distribution. It does not, however, explain the mechanism underlying the low-pass behavior of SEA. We next examine this mechanism using the displacement coherence function.

F. Displacement coherence function

Under the periodic idealization, the displacement of a tagged data value during the SE permutation process can be modeled as a random walk. For each swap, the value moves one site to the left with probability $1/N$, one site to the right with probability $1/N$, or remains at the same site with probability $1 - 2/N$. Let ξ_k denote the displacement during the kth swap:

$$\mathbb{P}(\xi_k = \pm 1) = \frac{1}{N}, \quad \mathbb{P}(\xi_k = 0) = 1 - \frac{2}{N}.$$

The displacement after ν swaps is therefore

$$d_\nu = \sum_{k=1}^{\nu} \xi_k.$$

By symmetry, the random walk has mean zero, and its variance is

$$\mathrm{Var}(d_\nu) = \nu \,\mathrm{Var}(\xi_1) = \frac{2\nu}{N}.$$

For a Fourier mode with angular frequency

$$\theta_f = \frac{2\pi f}{N},$$

the displacement d_ν corresponds to a phase shift $\theta_f d_\nu$. We therefore define the displacement coherence function as the characteristic function

$$C_{\mathrm{disp}}(f) = \mathbb{E}[e^{j\theta_f d_\nu}],$$

where $j^2 = -1$ is the imaginary unit (not to be confused with a summation index). Since the increments ξ_k are independent,

$$\begin{aligned} C_{\mathrm{disp}}(f) &= (\mathbb{E}[e^{j\theta_f \xi_1}])^\nu \\ &= \left(1 - \frac{2}{N} + \frac{1}{N}e^{j\theta_f} + \frac{1}{N}e^{-j\theta_f}\right)^\nu \\ &= \left(1 - \frac{2}{N}\{1 - \cos\theta_f\}\right)^\nu \\ &= \left(1 - \frac{4}{N}\sin^2\left(\frac{\pi f}{N}\right)\right)^\nu. \end{aligned} \tag{32}$$

Thus, $C_{\mathrm{disp}}(f)$ is identical to the periodic transfer function H_f^p in Eq. (13). For $f \ll N$ and $\nu = O(N)$,

$$C_{\mathrm{disp}}(f) \approx \exp\left(-4\pi^2 \frac{\nu}{N}\frac{f^2}{N^2}\right).$$

Using Eq. (12), this becomes

$$C_{\mathrm{disp}}(f) \approx \exp\left(-2\left(\frac{\pi\sigma f}{N}\right)^2\right),$$

recovering the Gaussian response (1).

G. SEA and destructive interference

The equivalence between the exact periodic transfer function (13) and the displacement coherence function (32) provides a mechanistic explanation for SEA as a low-pass filter under the periodic idealization. Adjacent swapping randomly displaces the data values, and averaging over the resulting phase shifts leads to destructive interference described by H_f^p. For the SE construction with reflecting boundaries, however, the ensemble-average operator is not diagonal in the Fourier basis, and exact agreement with H_f^p is therefore not expected.

To examine the extent to which the periodic result describes the SEA response with reflecting boundaries, here we compare the numerical SEA response with the corresponding periodic transfer functions. To eliminate finite-ensemble effects in the evaluation of the SEA at each selected ν_c, the SEA is evaluated in the infinite-ensemble limit. The critical number of swaps ν_c is determined for each threshold p_c using the KS diagnostic with an ensemble of size $M = 250$. For this comparison, p_{KS} is evaluated at 160 logarithmically spaced values of ν over the range $[10^{-2}N, 10^2 N]$.

Figure 5 compares the SEA response $|X^{\infty}_{\mathrm{SEA}}|/|X|$ with the exact periodic transfer function H_f^p (13). The discrete Fourier transform is evaluated over the one-sided frequencies $f = 0, \ldots, \lfloor N/2 \rfloor$, and frequencies satisfying $|X(f)| \leqslant 10^{-12} \max_f |X(f)|$ are excluded from the ratio. Natural logarithms are used throughout the current work. Since the

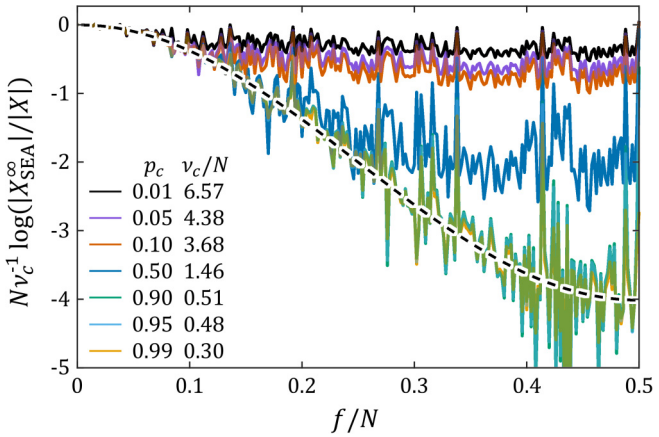


FIG. 5. SEA frequency response. The normalized logarithmic SEA response $Nv_c^{-1}\log(|X^{\infty}_{\mathrm{SEA}}(f)|/|X(f)|)$ as a function of the normalized frequency f/N, obtained for a representative time series $x(t)$ (cf. Fig. 1). Natural logarithms are used throughout the current work. Each curve corresponds to a different threshold p_c and the resulting critical number of swaps v_c. The broken line represents the analytical periodic transfer function (13).

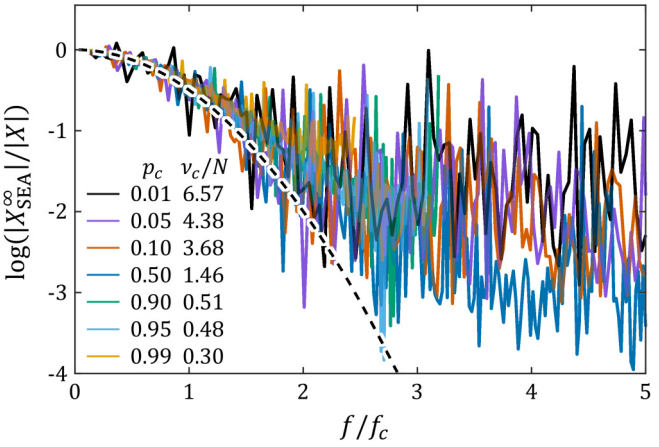


FIG. 6. SEA frequency response replotted. The same SEA response data shown in Fig. 5, replotted as $\log(|X^{\infty}_{\mathrm{SEA}}(f)|/|X(f)|)$ against the normalized frequency f/f_c, where $f_c = N/(2\pi\sqrt{2v_c/N})$. Each curve corresponds to a different threshold p_c and the resulting critical number of swaps v_c. The broken line represents the Gaussian response (1).

logarithm of H_f^p is proportional to ν, normalization by v_c/N causes the analytical transfer functions obtained from different thresholds p_c to collapse onto a single master curve. For lower values of p_c, deviations from the master curve become more apparent at higher frequencies, and the range of adherence becomes restricted to lower frequencies.

These deviations arise because the ensemble-average operator under reflecting boundary conditions, unlike its periodic counterpart, is not diagonal in the Fourier basis. The coupling between Fourier components therefore causes the response at each frequency to depend on other Fourier components of the time series. Consequently, the ratio $X^{\infty}_{\mathrm{SEA}}(f)/X(f)$ under reflecting boundary conditions depends on the time series and does not constitute a universal transfer function. As a side note, the exact periodic transfer function H_f^p may itself be used as a stand-alone low-pass filter if directly applied to $X(f)$. Exploration of this possibility and its implication, however, is outside the scope of the current study.

The same SEA response data are replotted in Fig. 6. $\log(|X^{\infty}_{\mathrm{SEA}}|/|X|)$ is shown as a function of the normalized frequency f/f_c, where $f_c = N/(2\pi\sqrt{2v_c/N})$ is the characteristic frequency. In this representation, the logarithmic SEA response follows the logarithm of the Gaussian response R_G (1) at low f/f_c, consistent with the understanding that R_G is the low-frequency asymptotic of the periodic transfer function H_f^p for $\nu = O(N)$ (cf. Sec. III A). The destructive-interference interpretation of H_f^p therefore also applies to its Gaussian low-frequency limit. At higher frequencies, however, R_G with the same v_c would produce stronger attenuation than H_f^p. Consequently, structured signal components lying outside the low-frequency regime would be attenuated more strongly by the corresponding Gaussian filter than by the periodic SE transfer. In practice, using $\sigma = \sqrt{2v_c/N}$ as the smoothing parameter of a stand-alone Gaussian filter is hence not recommended when preservation of structured signal components outside the low-frequency regime is desired.

H. Comparison to conventional smoothing approaches

Given a noisy time series $x(t) = s(t) + n(t)$, conventional smoothing methods generally seek a representation $y(t)$ that approximates the unknown signal $s(t)$. Methods such as kernel regression, Savitzky-Golay filtering, and smoothing splines achieve this through a smoothing parameter that controls the trade-off between retaining variations in $x(t)$ and suppressing those attributed to noise [12–18]. Ideally, the smoothing parameter would minimize the residue $r_s = \mathrm{rms}(y - s)$ between the smoothed signal and the ground truth. Since $s(t)$ is empirically unknown, however, the smoothing parameter is commonly selected indirectly using criteria such as cross-validation or regularization [19–26]. The meaning of an "optimal" smoothing parameter consequently depends on the criterion used to define optimality.

The SE construction approaches the problem from a different perspective. The starting point is not a parametrized filter applied directly to $x(t)$, but a statistical resampling procedure in which surrogates are generated by permutation of $x(t)$. The number of swaps ν specifies the extent of the permutation, while the corresponding smoothing operation arises only through ensemble averaging. As established in Sec. III A, this averaging process is governed by a discrete diffusion equation in the infinite-ensemble limit, and its periodic transfer function approaches the Gaussian response at low frequencies for $\nu = O(N)$. Gaussian-like smoothing therefore emerges from the statistics of the permuted time series instead of being prescribed as a kernel to be applied on $x(t)$.

This construction also permits the extent of smoothing to be related to the preservation of a relevant statistic. In the present implementation, increasing ν progressively alters the marginal first-difference distribution of the permuted time series, and the KS diagnostic is used to identify a value of ν associated with a prescribed degree of preservation. This cri-

terion is conceptually different from minimizing an estimated reconstruction error or balancing terms in a cost function. It does not attempt to infer the unknown ground truth $s(t)$ or define an optimal approximation to it. Instead, it specifies how extensively the observed data can be permuted while approximately retaining a selected statistical property.

For historical context, several established ideas are relevant to SE. Discrete diffusion smoothing and discrete scale-space theory have long been studied. For instance, signal smoothing has been formulated through a discrete diffusion equation and associated discrete Gaussian kernels [27]. Random adjacent transpositions are likewise a standard shuffle process whose mixing properties have been investigated in probability theory [28]. The same is true in statistical physics; Kawasaki dynamics, for example, evolves spin configurations of the Ising model through exchanges of neighboring spins while conserving the total magnetization [29,30]. There also exist approaches that characterize temporal structure through local ordering patterns. In particular, permutation entropy measures the complexity of time series by quantifying the distribution of ordinal patterns among data values at neighboring time points [31]. The KS diagnostic in the present implementation of SE is similarly concerned with preservation of local temporal structure, but operates on the marginal first-difference distribution rather than ordinal patterns. Permutation-based surrogate methods also exist. In particular, constrained randomization has been introduced in which pair exchanges are accepted or rejected through simulated annealing so as to minimize a prescribed cost function and thereby preserve selected properties of the original time series [32,33]. Such surrogate time series can then be used, for example, in hypothesis testing.

In contrast, the present SE construction accepts uniformly selected adjacent swaps without any externally imposed cost function. The extent of permutation is instead selected afterward using a statistical diagnostic. The contribution of SE therefore lies in combining this local permutation process with a preservation rule based on a relevant time-series statistic. In other words, SE is a statistical resampling scheme for time series; its smoothing behavior arises only as a by-product of ensemble averaging. The resulting ensemble can be used to examine conditional variability or numerical stability of quantities derived from the original time series.

I. Empirical mode decomposition with SE

The preceding sections concern the SEA and its interpretation as a low-pass filter. The SE, however, can also be used without ensemble averaging as a collection of surrogate time series generated from a single observation. This possibility arose from a simple question: Why not generate nearby time series by repeatedly swapping adjacent data points? We were unable to identify an established statistical resampling scheme based on this construction. The SE therefore provides a simple, model-free way of generating surrogate time series without model fitting or training, allowing the sensitivity of a downstream analysis to perturbations of the observed data to be examined.

As a proof of concept, we apply this idea to EMD. EMD is a data-driven method that decomposes a time series into intrinsic mode functions (IMFs) representing oscillatory components at progressively lower characteristic frequencies, together with a residual component [34,35]. Each IMF is extracted through an iterative sifting procedure based on local extrema and interpolated envelopes. All EMD calculations in this work are performed using the `emd` function in MATLAB with spline interpolation and otherwise default parameters [36]. Because the decomposition depends sensitively on the extrema of the input time series, EMD can exhibit mode mixing, in which oscillations of different timescales occur within the same IMF, as well as mode splitting, in which oscillatory content is split among additional IMFs.

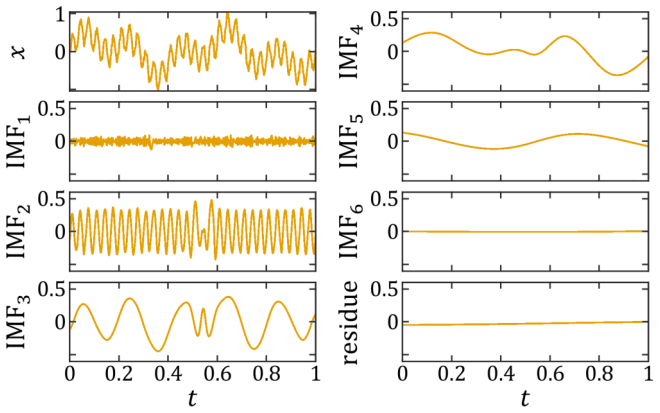


FIG. 7. Empirical mode decomposition. The intrinsic mode functions (IMFs) and residue of a noisy time series $x(t) = s(t) + n(t)$ of length $N = 500$, where $s(t) = \frac{1}{3}[\sin(4\pi t) + \sin(10\pi t) + \sin(60\pi t)]$ and $n(t) = 0.05 n_G(t)$, with $n_G(t) \sim \mathcal{N}(0, 1)$.

Figure 7 illustrates these problems for a time series $x(t) = s(t) + n(t)$ of length $N = 500$, where

$$s(t) = \tfrac{1}{3}[\sin(4\pi t) + \sin(10\pi t) + \sin(60\pi t)]$$

and $n(t) = 0.05 n_G(t)$. The prescribed signal contains three sinusoidal components in addition to the additive noise. The EMD produces six IMFs, with some oscillatory content distributed across multiple modes, thereby signifying mode mixing and mode splitting.

The SE can be used to bias such a decomposition by retaining ensemble members according to the number of IMFs obtained. An example of this procedure, referred to as synthetic ensemble EMD (SE-EMD), is shown in Fig. 8. The critical number of swaps ν_c is first selected using $p_c = 0.99$ with an ensemble of size $M = 100$. A fresh SE is then generated with $M = 10\,000$, and EMD is applied to each permuted time series in the ensemble. Let J_m denote the number of IMFs obtained from the mth member. We define

$$J_{\min} = \min_m J_m$$

and retain the subset

$$\mathcal{M}_* = \{m : J_m = J_{\min}\}.$$

For the present example, $J_{\min} = 4$ and $|\mathcal{M}_*| = 1087$. For IMF index j, the conditional ensemble mean is

$$\overline{\mathrm{IMF}}_j(t) = \frac{1}{|\mathcal{M}_*|} \sum_{m \in \mathcal{M}_*} \mathrm{IMF}_{m,j}(t),$$

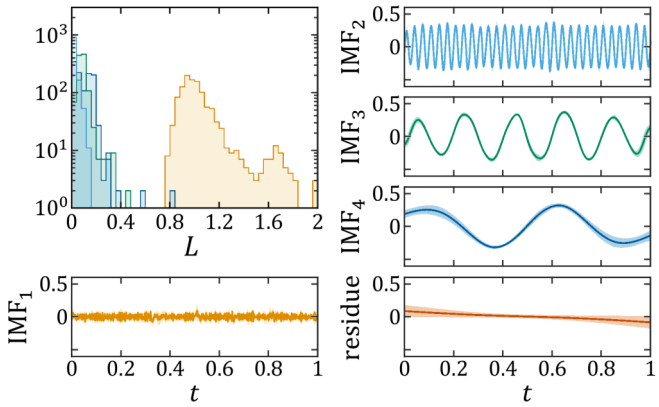


FIG. 8. Synthetic ensemble EMD. An SE containing 10 000 permuted time series of $x(t)$ (cf. Fig. 7) is generated using $p_c = 0.99$. The retained subset consists of the 1087 SE members that yield the minimum number of IMFs, $J_{\min} = 4$. The curves show the subset-averaged IMFs and residue, while the shaded bands indicate one standard deviation within the retained subset and therefore represent conditional SE-EMD variability. The histograms show the distributions of $L_{1,n}$ for the noise-dominated mode and $L_{j,f}$ for the three IMFs corresponding to the prescribed sinusoidal components of frequencies $f = 30$, 5, and 2.

with variance

$$s_j(t)^2 = \frac{1}{|\mathcal{M}_*| - 1} \sum_{m \in \mathcal{M}_*} [\mathrm{IMF}_{m,j}(t) - \overline{\mathrm{IMF}}_j(t)]^2.$$

The shaded bands in Fig. 8 therefore quantify the conditional variability within the retained subset of IMFs.

To assess whether SE-EMD produces cleaner separation of the prescribed oscillatory components, we quantify the fraction of each extracted IMF lying outside the corresponding target-frequency subspace. For a prescribed frequency f, let Z_f denote the N-by-2 matrix whose two columns are the vectors $\sin(2\pi ft)$ and $\cos(2\pi ft)$ evaluated at the N time points. Any sinusoid of frequency f can be expressed as a linear combination of these two vectors. We therefore project IMF_j onto the space spanned by them using the standard least-squares projection operator [37],

$$P_f = Z_f \left(Z_f^{\mathsf{T}} Z_f\right)^{-1} Z_f^{\mathsf{T}}.$$

The product $P_f \times \mathrm{IMF}_j$ is the portion of IMF_j that can be represented by a sinusoid of frequency f, whereas $(I - P_f) \times \mathrm{IMF}_j$ is the remaining portion. This affords a definition of the nontarget fraction

$$L_{j,f} = \frac{\|(I - P_f) \times \mathrm{IMF}_j\|^2}{\|\mathrm{IMF}_j\|^2}.$$

Thus, $L_{j,f} = 0$ corresponds to an IMF containing only the target frequency, whereas larger values indicate increasing nontarget content.

Since the additive noise $n(t)$ has no single target frequency, the nontarget fraction is not defined for the noise-dominated mode. However, since $n(t)$ is known in this numerical example, we define the reconstruction error

$$L_{1,n} = \frac{\|\mathrm{IMF}_1 - n\|}{\|n\|}.$$

TABLE I. Reconstruction error and nontarget fraction for SE-EMD. The retained-subset values are reported as mean ± standard deviation over the retained decompositions, whereas the subset-averaged values are calculated from the subset-averaged IMFs. The corresponding values obtained from standard EMD are shown for comparison.

	Retained subset	Subset averaged	Standard EMD
$L_{1,n}$	1.040 ± 0.195	0.608	0.672
$L_{2,30}$	0.0374 ± 0.0212	0.0192	0.0617
$L_{3,5}$	0.0914 ± 0.0543	0.0686	0.205
$L_{4,2}$	0.102 ± 0.0888	0.0567	0.339

Here, $L_{1,n} = 0$ corresponds to exact recovery of the prescribed noise realization, whereas larger values indicate increasing deviation from it.

For each prescribed frequency f, the corresponding IMF is identified as the one that minimizes $L_{j,f}$. For SE-EMD, IMF_1, IMF_2, IMF_3, and IMF_4 correspond to the noise-dominated mode and the prescribed sinusoidal components with $f = 30$, 5, and 2, respectively. The resulting values of L are summarized in Table I, while the corresponding distributions within the retained subset are shown as histograms in Fig. 8. For comparison, the table also lists the corresponding values for the standard EMD. In the current example, the subset-averaged SE-EMD values are smaller than the corresponding standard EMD values for all four modes. For the three selected sinusoidal modes, this corresponds to smaller nontarget fractions, while for the noise-dominated mode it corresponds to a smaller relative error with respect to the prescribed noise realization. The subset-averaged values are also smaller than the corresponding means within the retained subset for all four modes. These reductions in L, however, are specific to the current example and hence are not expected in general.

While SE offers EMD a mechanism to output fewer IMFs, we stress that this is not intrinsically preferable. Instead, such a decomposition scheme is only one way to apply SE. The present comparison also does not establish superiority of SE-EMD over existing EMD variants such as ensemble EMD (EEMD) [38] or complete EEMD with adaptive noise (CEEMDAN) [39].

IV. CONCLUSIONS AND OUTLOOK

The SE framework, a statistical resampling scheme for time series based on repeatedly swapping randomly selected pairs of adjacent data points, is introduced. The individual SE members are permuted surrogates of the original time series $x(t)$, and their ensemble average defines the SEA. In the infinite-ensemble limit, the SEA is governed exactly by a discrete diffusion equation with reflecting boundary conditions. Under a periodic idealization of the boundaries, an exact transfer function can be obtained. For frequencies $f \ll N$ and a number of swaps $\nu = O(N)$, where N is the data length, this transfer function reduces asymptotically to the Gaussian response.

The SEA residue, $r_x^{\mathrm{SEA}} = \mathrm{rms}(x_{\mathrm{SEA}} - x)$, when normalized by the magnitude of the first difference $\mathrm{rms}(\Delta x)$, can be characterized in terms of the normalized number of swaps ν/N.

When the time series is dominated by white noise, the normalized residue approaches the Gaussian similarity function Π_G, which depends solely on ν/N. Finite-ensemble size, data length, noise amplitude, and noise spectrum affect departures from this relation. The two-sample KS test provides a complementary operational diagnostic by monitoring changes in the marginal first-difference distribution. A critical number of swaps ν_c can then be defined using a prescribed KS p-value threshold.

The low-pass behavior of SEA can be understood through the random displacement of individual data values. Under the periodic idealization, the characteristic function associated with this displacement is identical to the exact SEA transfer function. The resulting attenuation can therefore be interpreted as destructive interference arising from the phase shifts induced by the random motion of data values. For reflecting boundaries, coupling between Fourier components causes the numerical SEA response to deviate from the periodic transfer function as the extent of permutation increases.

Individual SE members can also be used as surrogate inputs to downstream analyses. As a proof of concept, we have applied EMD to each permuted time series and averaged the resulting IMFs over a retained subset whose members have the minimum number of IMFs. In the example considered, the IMFs show improved overlap with the prescribed components of the input time series compared with standard EMD, while variability across the retained subset provides conditional variability bands for the IMFs.

The present SE uses uniformly selected adjacent swaps, with the extent of permutation selected according to preservation of the marginal first-difference distribution. Future work may consider other statistics for this selection, nonuniform swap probabilities informed by local properties of the time series, and extensions to nonstationary, sparse, event-driven, or multidimensional data [40–46].

ACKNOWLEDGMENT

We gratefully acknowledge the support of Okinawa Institute of Science and Technology (OIST) Graduate University with subsidy funding from the Cabinet Office, Government of Japan.

DATA AVAILABILITY

All study data are included in the article and/or Supplemental Material [6].

[1] J. Klapper and C. Harris, On the response and approximation of Gaussian filters, IRE Trans. Audio **AU-7**, 80 (1959).
[2] S. Ohkawa, M. Yoshizawa, and K. Husimi, Direct synthesis of the Gaussian filter for nuclear pulse amplifiers, Nucl. Instrum. Methods **138**, 85 (1976).
[3] M. Basu, Gaussian-based edge-detection methods—A survey, IEEE Trans. Syst. Man Cybern. C **32**, 252 (2002).
[4] F. Dablander and T. M. Bury, Deep learning for tipping points: Preprocessing matters, Proc. Natl. Acad. Sci. USA **119**, e2207720119 (2022).
[5] J. Nako, G. Tsirimokou, C. Psychalinos, A. S. Elwakil, B. J. Maundy, and P. Bertsias, Reduced complexity approximation and design of Gaussian impulse response filters and wavelets, IEEE Access **12**, 160064 (2024).
[6] See Supplemental Material at http://link.aps.org/supplemental/10.1103/g6fz-4fsh for MATLAB code implementing the synthetic ensemble generation workflow.
[7] A. Kolmogorov, Sulla determinazione empirica di una legge di distribuzione, G. Ist. Ital. Attuari **4**, 83 (1933).
[8] N. Smirnov, Table for estimating the goodness of fit of empirical distributions, Ann. Math. Stat. **19**, 279 (1948).
[9] F. J. Massey Jr., The Kolmogorov-Smirnov test for goodness of fit, J. Am. Stat. Assoc. **46**, 68 (1951).
[10] G. Marsaglia, W. W. Tsang, and J. Wang, Evaluating Kolmogorov's distribution, J. Stat. Software **8**, 1 (2003).
[11] G. B. Arfken, H. J. Weber, and F. E. Harris, *Mathematical Methods for Physicists: A Comprehensive Guide* (Academic Press, Massachusetts, USA, 2011).
[12] A. Savitzky and M. J. E. Golay, Smoothing and differentiation of data by simplified least squares procedures, Anal. Chem. **36**, 1627 (1964).
[13] R. W. Schafer, What is a Savitzky–Golay filter? IEEE Signal Process. Mag. **28**, 111 (2011).
[14] J. D. Hart, Kernel regression estimation with time series errors, J. R. Stat. Soc. B: Stat. Methodol. **53**, 173 (1991).
[15] W. Härdle and P. Vieu, Kernel regression smoothing of time series, J. Time Ser. Anal. **13**, 209 (1992).
[16] I. J. Schoenberg, Contributions to the problem of approximation of equidistant data by analytic functions. Part A. On the problem of smoothing or graduation. A first class of analytic approximation formulae, Q. Appl. Math. **4**, 45 (1946).
[17] C. H. Reinsch, Smoothing by spline functions, Numer. Math. **10**, 177 (1967).
[18] P. Craven and G. Wahba, Smoothing noisy data with spline functions, Numer. Math. **31**, 377 (1978).
[19] R. R. Picard and R. D. Cook, Cross-validation of regression models, J. Am. Stat. Assoc. **79**, 575 (1984).
[20] J. Shao, Linear model selection by cross-validation, J. Am. Stat. Assoc. **88**, 486 (1993).
[21] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, Cross validation for model selection: A review with examples from ecology, Ecol. Monogr. **93**, e1557 (2023).
[22] S. Bates, T. Hastie, and R. Tibshirani, Cross-validation: What does it estimate and how well does it do it? J. Am. Stat. Assoc. **119**, 1434 (2024).
[23] A. E. Hoerl and R. W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems, Technometrics **12**, 55 (1970).
[24] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. B: Stat. Methodol. **58**, 267 (1996).
[25] P. H. C. Eilers and B. D. Marx, Flexible smoothing with B-splines and penalties, Stat. Sci. **11**, 89 (1996).
[26] P. H. C. Eilers and B. D. Marx, Splines, knots, and penalties, WIREs Interdiscip. Rev. Comput. Stat. **2**, 637 (2010).
[27] T. Lindeberg, Scale-space for discrete signals, IEEE Trans. Pattern Anal. Mach. Intell. **12**, 234 (1990).

[28] H. Lacoin, Mixing time and cutoff for the adjacent transposition shuffle and the simple exclusion, Ann. Probab. **44**, 1426 (2016).

[29] K. Kawasaki, Diffusion constants near the critical point for time-dependent Ising models. I, Phys. Rev. **145**, 224 (1966).

[30] K. Kawasaki, Diffusion constants near the critical point for time-dependent Ising models. II, Phys. Rev. **148**, 375 (1966).

[31] C. Bandt and B. Pompe, Permutation entropy: A natural complexity measure for time series, Phys. Rev. Lett. **88**, 174102 (2002).

[32] T. Schreiber, Constrained randomization of time series data, Phys. Rev. Lett. **80**, 2105 (1998).

[33] T. Schreiber and A. Schmitz, Surrogate time series, Physica D **142**, 346 (2000).

[34] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N. C. Yen, C. C. Tung, and H. H. Liu, The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis, Proc. R. Soc. A **454**, 903 (1998).

[35] A. Zeiler, R. Faltermeier, I. R. Keck, A. M. Tomé, C. G. Puntonet, and E. W. Lang, Empirical mode decomposition—An introduction, in *The 2010 International Joint Conference on Neural Networks (IJCNN)* (IEEE, Barcelona, 2010), pp. 1–8.

[36] MathWorks, EMD—Empirical mode decomposition, MATLAB Signal Processing Toolbox, https://www.mathworks.com/help/signal/ref/emd.html.

[37] L. N. Trefethen and D. Bau, III, *Numerical Linear Algebra* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1997).

[38] Z. Wu and N. E. Huang, Ensemble empirical mode decomposition: A noise-assisted data analysis method, Adv. Adapt. Data Anal. **01**, 1 (2009).

[39] M. E. Torres, M. A. Colominas, G. Schlotthauer, and P. Flandrin, A complete ensemble empirical mode decomposition with adaptive noise, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Prague, 2011), pp. 4144–4147.

[40] E. J. Kostelich and T. Schreiber, Noise reduction in chaotic time-series data: A survey of common methods, Phys. Rev. E **48**, 1752 (1993).

[41] V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley, Econophysics: Financial time series from a statistical physics point of view, Physica A **279**, 443 (2000).

[42] J. Aljadeff, B. J. Lansdell, A. L. Fairhall, and D. Kleinfeld, Analysis of neuronal spike trains, deconstructed, Neuron **91**, 221 (2016).

[43] H. Y. D. Sigaki, M. Perc, and H. V. Ribeiro, History of art paintings through the lens of entropy and complexity, Proc. Natl. Acad. Sci. USA **115**, E8585 (2018).

[44] B. Lee, M. K. Seo, D. Kim, I. Shin, M. Schich, H. Jeong, and S. K. Han, Dissecting landscape art history with information theory, Proc. Natl. Acad. Sci. USA **117**, 26580 (2020).

[45] A. A. Bagrov, I. A. Iakovlev, A. A. Iliasov, M. I. Katsnelson, and V. V. Mazurenko, Multiscale structural complexity of natural patterns, Proc. Natl. Acad. Sci. USA **117**, 30241 (2020).

[46] S. T. Chan and E. Fried, Structural stability and thermodynamics of artistic composition, Proc. Natl. Acad. Sci. USA **121**, e2406735121 (2024).