Signals as departures from random walks

Glenn Ierley [*]
Department of Mathematical Sciences, Michigan Technological University, 1400 Townsend Drive, Houghton, Michigan 49931, USA and Scripps Institution of Oceanography, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093-0225, USA

Alex Kostinski [†]
Department of Physics, Michigan Technological University, 1400 Townsend Drive, Houghton, Michigan 49931, USA

(Received 21 January 2022; accepted 18 May 2022; published 14 June 2022)

We study statistics of data ranking, focusing on the recently discovered distribution-invariant discrete eigenvalue spectrum for an independent and identically distributed (IID) process. We employ a variant of a cumulative distribution function in rank and time that maps the sampling variability for an IID process onto a set of random walks. This mapping admits confidence bounds on whether the residual (data with signal removed) arises solely from IID sampling variability. Any deviations judged significant are regarded as signals, whether deterministic, chaotic, or random. Unlike our recent work on extracting unknown signals in arbitrary noise, here we focus on aspects that are easily combined with any other methods of signal extraction. The ubiquitous case of a single trace receives particular attention. The approach is illustrated on dark current and gamma-ray arrival datasets where we examine the residual for consistency with the expected sampling variability of IID noise.

DOI: 10.1103/PhysRevE.105.064114

I. INTRODUCTION

Perhaps no task is more ubiquitous in physical science and data analysis than detection, separation, and extraction of signals from noise, e.g., the detection of a gravitational wave (weak) signal, arrival of a high energy cosmic ray, or laser-induced fluorescence. Consequently, the literature on the subject is vast, and spans many disciplines such as an interdisciplinary field of statistical signal processing and nonparametric statistics [1]. Throughout this literature, the noise is typically assumed to be additive, Gaussian, and white. In contrast, here we focus on developing a method capable of handling arbitrary and unknown noise, i.e., of an unknown arbitrary distribution, including heavy-tailed and infinite variance, undefined mean such as Cauchy (Lorentzian), etc.

The problem is significant because the outliers ("rare events") unduly influence the conventional methods such as least squares, but the corresponding literature is considerably narrower [2]. Unlike our earlier work [3–5], throughout this paper, we stress the modularity of the approach. The reader may use any method to extract a signal from data. We then furnish simple means based on a universal spectrum of eigenvalues for independent and identically distributed (IID) noise for answering whether the residual data (after signal removal) has arisen solely from the ever-present sample-to-sample fluctuations (sampling variability).

Our route to distribution-invariant results is to rank the data. Here ranking by magnitude is meant; i.e., a sequence [0.94, 1.87, 0.60] converts to 2,3,1 where the lowest value maps to rank 1. Although much effort has been devoted to rank-based decision tests in the nonparametric statistics literature [6], there is a dearth of rank-based approaches to signal retrieval. For example, the three-volume set [1] does not include any ordinal methods.

To that end, we developed such a method in a sequence of recent papers [3–5]. This article aims to expand on the part of our method that is readily combined with any other method of signal extraction and/or fitting. First, we elucidate the relationship between ranking and conditions imposed on white noise and map sampling variability of an IID random process to realizations of a pinned (returning or clamped, all used interchangeably) random walk. We then focus on the ubiquitous and practically important case of a single time series ($n = 1$) and illustrate the new method on dark current noise data from a camera and from an astrophysical time series.

II. RANKING AND WHITE NOISE

Because of the scarcity of rank-based approaches to signal analysis and to illustrate counterintuitive features of rank, before embarking on our main thrust of evaluating quality of signal extraction, we begin with a brief excursion on ranking of white noise. Typically, in the physics literature, $\delta-$correlation in time defines whiteness as the spectrum flatness then follows from the Wiener-Khinchin theorem (e.g., p. 237 of [7]). However, we begin with the stricter case of perfect white noise, that is, a random process of identically distributed, meaning the same probability density function (pdf), and independently drawn trials, the IID process. Let n denote the number of series and each series be of length

[*]Emeritus, Scripps Institution of Oceanography; grierley@ucsd.edu
[†]kostinsk@mtu.edu

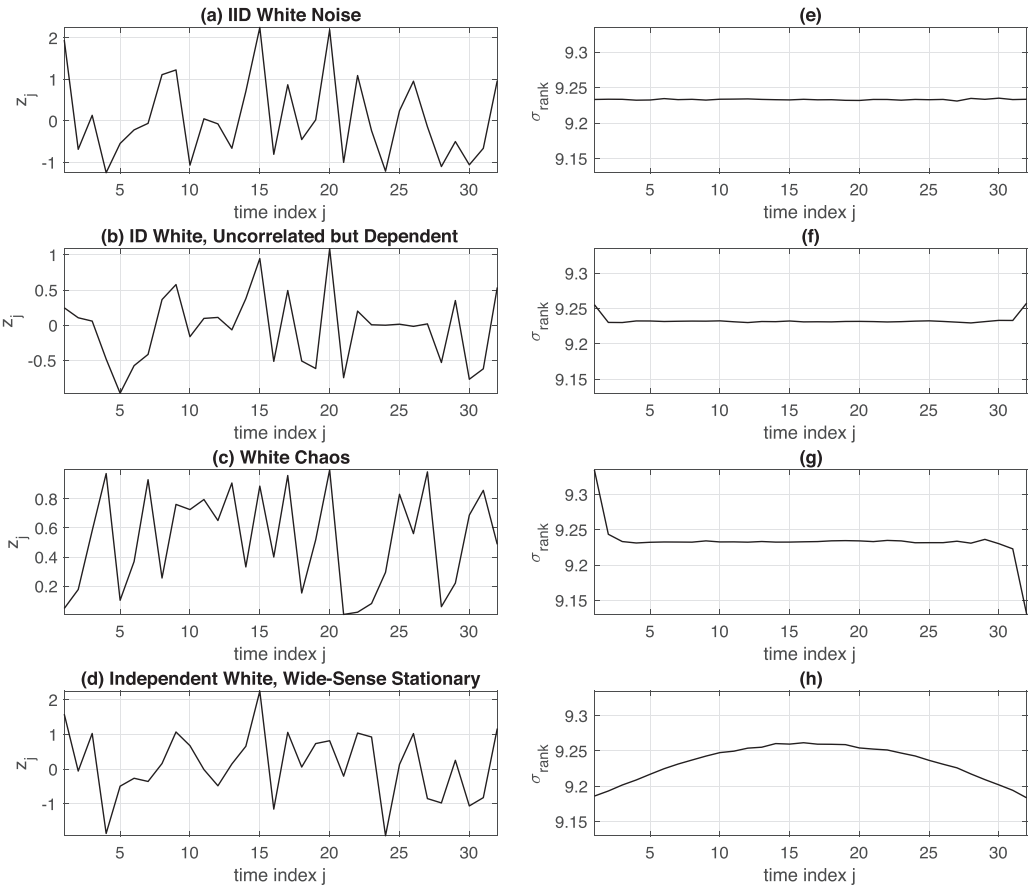

FIG. 1. Ranking patterns distinguish variants of white noise. [(a)–(d)] Four traces of length $N = 32$. Panel (a) is an output of a strict (IID) Gaussian white noise whereas (b)–(d) are white (δ correlated) but not IID. Panel (b) ("strong but not strict white noise") is an uncorrelated but dependent random variable $z_j \equiv x_j\, x_{j-1}$, with IID x drawn from $\mathcal{N}(0, 1)$. Panel (c) ("white chaos") is the output of logistic map $x_{j+1} = a x_j\,(1 - x_j)$, uncorrelated at the chosen $a = 4.0$. Panel (d) ("wide-sense stationary") is from a Pearson pdf with a quadratically varying kurtosis. The four data traces look similar but panels (e)–(h), paired with (a)–(d) and based on $n = 10^7$ realizations each, differ in standard deviation of rank patterns. (e) The mean value of σ_{rank} for IID noise for equally likely integers $1, 2, \ldots, 32$ is given by $\sqrt{(N-1)(N+1)/12} = \sqrt{341}/2 \approx 9.23$. However, despite the whiteness, this statistically uniform σ_{rank} profile for IID noise is disturbed by the spikes in (f) and (g) at $j = 1, 32$ (see Appendix A for details). (h) σ_{rank} follows the subtle kurtosis pattern, despite the constant variance of the data itself.

N. When averaged over a large number of series, the sample mean of the data, by the law of large numbers, will approach the true mean (a constant) and so will the mean rank (and all higher moments) as time slots in a series are indistinguishable in the IID case and can be reshuffled.

Note, however, that not all white noise is created equal, e.g., δ correlation can hold but statistical independence be lacking (e.g., see Fig. 1). Indeed, the notion of "white noise" is not treated uniformly throughout the literature (e.g., pp. 114–115 of [8]) but we reserve the phrase "purely random sequence" or a "strictly white noise" for the IID process as in, e.g., pp. 254–256 of [9]. This raises an interesting question: Is the ranking operation sensitive only to whiteness or to other attributes, such as the lack of statistical independence? The purpose of Fig. 1 is to demonstrate that the answer is the latter and that departure from uniformity of rank variance transcends the scope of correlation. Although there are white noise tests in the literature, e.g., the Bartlett test [8] and rank correlation tests such as the Kendall τ test [10], we have not seen this question raised about patterns of rank. In Fig. 1 we highlight the differences in the response of the ranking operation to variants of white noise and chaos.

Specifically, Fig. 1 shows that departing from the IID conditions, e.g., introducing uncorrelated but dependent noise or white chaos, creates patterns in rank statistics [Figs. 1(f)–1(h)]. These patterns constitute signals that the conventional correlation analysis would miss. For example, Fig. 1(b) is a series of trials from $z_k \equiv x_k\, x_{k-1}$ with the x being IID and $\mathcal{N}(0, 1)$ [11]. Hence, z is white insofar as entries are not correlated, zero mean, and identically distributed. However, statistical independence is gone so this is a strong but not strict white noise [9]. For this stationary, uncorrelated but dependent process, Fig. 1(f) shows enhanced fluctuations of rank at each end, presumably because each has only one neighbor. Based on this hunch, one would expect an increase in end effects with decreasing data segment length N and this is, indeed, the case. Once independence is lost, asking whether the entry in, say, the first time slot is the smallest becomes a question of conditional probability. This is discussed explicitly for the example of white chaos in Appendix A.

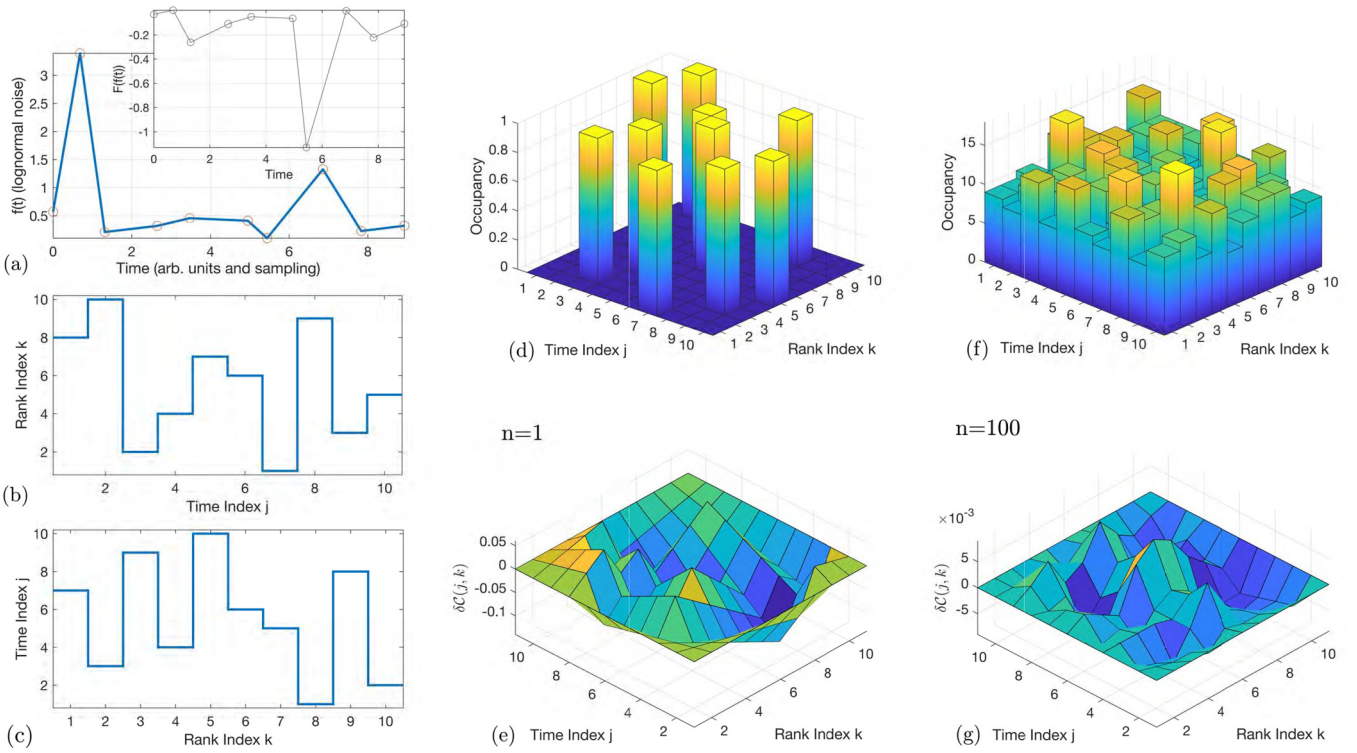

FIG. 2. A new distribution-invariant description of an IID discrete random process: joint rank-time empirical cumulative distribution. (a) Magnitude $f(t)$ of a log-normally fluctuating quantity vs time t. To emphasize the general setting, $f(t)$ is measured in arbitrary units and sampled unevenly. (b) Rank r vs time t for the data of panel (a). Monotonic transformation of data, inset $F(f(t))$, yields the same r vs t as rank depends solely on serial order. Nor does the nonuniform sampling in time affect the abscissa as the time index is also integer valued from 1 to 10. (c) Time vs rank for the data of panel (a): t and r are *statistically* indistinguishable for a discrete IID process. (d) Joint rank-time histogram, whose normalized version is a probability mass function (pmf). (e) Joint deviation cumulative distribution function (cdf), $\delta\mathcal{C}$, defined in Eq. (1), i.e., the difference between the actual and theoretical IID cdf, for data of panel (d), designed to contrast the actual sampling variability with that of an IID process. (f) Same as panel (d) but extending the dataset in (a) to $n = 100$, yielding a histogram closer to the ensemble limit pmf $U(1, N) \times U(1, N)$. (g) *Relative* fluctuations of $\delta\mathcal{C}$, similar to those in (e) but with the decrease in amplitude by a factor of $\sqrt{n} = 10$. For any n, each row and column slice of the augmented $\delta\mathcal{C}$ is a pinned random walk (see text).

In Figs. 1(d) and 1(h) we highlight detection of a subtle nonstationarity by ranking. Here the data are wide-sense stationary, drawn from a Pearson distribution whose four moments are independently adjusted parameters. In this case, the first three moments—mean, variance, and skewness—are all constant with time but the fourth (kurtosis) has a parabolic profile in time. Figure 1(h) shows the standard deviation of rank, which detects the parabolic profile of the raw data kurtosis. The global nature of the ranking operation causes this spillover of high-order moment information into lower ones and suggests *simple, quick, and robust (distribution-invariant) means of detecting high-order nonstationarity in raw data*.

The above comments and Fig. 1 are meant to interest the reader in the unusual properties of rank, but the approach described here is of conceptual rather than practical interest as the required number of series n is large [12]. We note that there is an extensive literature on practically useful detection of statistical dependence (e.g., [13,14]) extending beyond the single time series case discussed here, probing for dependence between distinct random variables (RVs) X and Y.

Next, we proceed with the expansion of the ranking into the rank-time plane, as shown in Fig. 2. Our goal here is to describe a fundamental modal expansion of the IID process, a complete characterization that can be used with any method of signal extraction to answer the question: Is the residual (data with signal removed) IID noise? To that end, Fig. 2 supplies a detailed description of our procedure, initially proposed in [5]. In typical applications the required n is often quite modest including the common case of $n = 1$, emphasized below. The guiding principle is simple: invariance with respect to data *reshuffling* in the IID case. In other words, all permutations of N ranks among the N indexed "time" slots are equally likely, occurring with probability $1/N!$. This is the context for Figs. 2(a)–2(c). Figure 2(b) shows the result of ranking by magnitude [15]. The setting is rather general insofar as the data can be sampled nonuniformly and the units are arbitrary because ranking is not affected. Nor are the results affected by any monotone transformation of the data as illustrated by the inset of Fig. 2(a) where the data are barely recognizable yet the same rank is obtained. This kind of robustness is lacking in conventional methods such as least squares.

The purpose of flipping the axes in Figs. 2(b) and 2(c) is to emphasize that rank and time are statistically indistinguishable for IID processes: Figs. 2(b) and 2(c) occur with the same frequency. This rank-time parity does not hold for all variants of stationary white noise as Figs. 1(b) and 1(f) show, an insight we missed in earlier work [5].

III. EMPIRICAL RANK-TIME CUMULATIVE DISTRIBUTION FUNCTION

Guided by this indistinguishability and invariance under exchange of rank (r) and time indices (t) for the IID process, we introduce a mapping to the rank-time (r, t) plane, as shown in Figs. 2(d) and 2(f), for $n = 1$ and $n = 100$, respectively. This rank-time distribution function must satisfy additional constraints that stem from the global nature of rank. For $n = 1$ in Fig. 2(d), there is one unique rank k for each time interval j, that is, one unit entry per each row and column. Thus, the row and column sums are all unity. For the sum of n permutation matrices, each of the rows and columns add up to n.

Initially, we searched for a representation of an IID random process that takes maximal advantage of the available symmetries such as the interchange of time and rank, their reshuffling, etc. This is the reason for the move from the customary pdf to a (deviation) cdf [Figs. 2(e) and 2(g)] and from one to two dimensions: rank and time, to examine joint distributions. [16]. In the large number of time series limit, $n \to \infty$, the rank-time joint pdf is jointly uniform, i.e., a perfect horizontal surface at the value of $1/N^2$ in the two-dimensional (2D) rank-time space, $U(1, N) \times U(1, N)$, for N points in a data sequence where U denotes the uniform distribution. However, for a finite n, sampling variability is present and signal-induced departure from uniformity must be distinguished from this sampling variability. Thus signal detection becomes a "judgment call" that must be supplemented or quantified with some level of confidence.

For example, Fig. 2(d) is the 2D histogram (pmf) in the rank-time space for $n = 1$ where the bar heights are either 0 or 1, with a single unit entry in each row and column. The relative height spread in the analogous Fig. 2(f) is $\sqrt{100} = 10$ smaller, as the sampling variability is $\mathcal{O}(\sqrt{n})$. For the single series case ($n = 1$), the 2D rank-time discrete pdf is anything but a uniform horizontal planar surface, and sampling variability is extreme. However, as shown below, the sampling variability of the smoother cdf is amenable to an analytic treatment and follows a simple convergent eigenfunction expansion for all n, even $n = 1$. In fact, sampling variability of the deviation cdf, to be defined shortly, maps onto pinned (or clamped, the two terms used interchangeably) random walks—hence the title of this paper. Our next move to Figs. 2(e) and 2(g) is therefore to construct the 2D rank-time cumulative *deviation* empirical distribution, denoted as $\delta\mathcal{C}$.

To motivate the construction, recall that a uniform pdf $p(x, y) = 1$ over a square is the continuous analog to $U(1, N) \times U(1, N)$, for N data points and the associated cdf is simply $\mathcal{C}(x, y) = xy$, paralleling the one-dimensional (1D) case, $\mathcal{C}(x) = x$ for the uniform pdf in one dimension. Therefore, to focus on the contrast between the actual and IID sampling variability, given the cdf $\mathcal{C}$ for a sample time series, the deviation cdf is defined on the discrete lattice as

$$\delta\mathcal{C}_{k,l} = \sum_{i=1}^{k} \sum_{j=1}^{l} \left(p_{i,j} - \frac{1}{N^2} \right), \quad \{k, l\} = 1, 2, \ldots, (N-1), \tag{1}$$

where $p_{i,j}$ is the probability mass function (pmf) and the symbol δ indicates deviation [17]. Again, the main motivation is to zoom in on the deviation from the IID limit. In order to clarify the link to random walks (IID increments implied throughout when using this term), and to minimize a vexing artifact of discretization, in all figures, we augment $\delta\mathcal{C}$ with a zero boundary so that there are $N + 1$ (rather than $N - 1$) points on the rank and time axes of all $\delta\mathcal{C}$ surface plots. However, all calculations, including the eigenfunction expansion, rely on Eq. (1).

Signals, trends, correlations, etc., occur in time, thereby lacking the rank-time symmetries of pure randomness. We exploit this contrast to extract signals from data via departures from various symmetries such as rank-time parity. The probabilities of such rank-time symmetries define the IID process uniquely and departures from this baseline can then be used to address the perpetual difficulty of separating true signals from sample-to-sample fluctuations (sampling variability). In this regard the $\delta\mathcal{C}$ construct proves crucial as it furnishes a universal analytic and convergent eigenfunction expansion for the IID case [5], thus allowing efficient signal extraction, even in the case of a single time series, $n = 1$.

IV. IID PROCESS IN RANK-TIME AND PINNED RANDOM WALKS

At any n, rows and columns of an IID sample $\delta\mathcal{C}$s such as Figs. 2(e) and 2(g) or Figs. 3(b) and 3(e) are pinned at the ends, whether IID or not. This pinning stems from the rank-time pdf constraint that rows and columns of the normalized probability mass function add up to $1/N$ and then subtraction in the definition (1) of $\delta\mathcal{C}$ annuls it. This is illustrated in Fig. 3 where Figs. 3(c) and 3(f) project out individual sheets of the associated $\delta\mathcal{C}$ surfaces in Figs. 3(b) and 3(e). These IID-based $\delta\mathcal{C}$ slices (rows and columns) are samples of a pinned random walk.

The data in Fig. 3 are for dark current, collected with the lens blocked and due to the thermally generated electrons flowing in the absence of incident photons. Figure 3(a) shows a 128 pixel sample of this shot noise (from spontaneously generated electrons within the 5464×8192 CMOS chip). These are the raw data: output from a 14 bit A-D converter on a Canon R5 camera. We ask: Is there an instrumental bias or inhomogeneity on this spatial scale or are the data from an IID process? As it turns out (see below), the answer is the latter but, given the sampling variability, the answer must be statistical, i.e., a judgment call, supplemented by some specified level of confidence. How should such a judgment be made?

One simple albeit crude approach is to pick a scalar metric, e.g., some integral characteristic of $\delta\mathcal{C}$, for an IID process and compare it to the data at hand. For example, a simple average over the $(N-1) \times (N-1)$ rank-time grid cells, $\overline{\delta\mathcal{C}}$, defined by

$$\overline{\delta\mathcal{C}} \equiv \frac{1}{(N-1)^2} \sum_{j=1}^{N-1} \sum_{k=1}^{N-1} \delta\mathcal{C}_{j,k}, \tag{2}$$

is zero for an IID process in the $n \to \infty$ limit. Given a dataset consisting of a finite number of time series n (each of length N) the question then is whether the value of $\overline{\delta\mathcal{C}}$ differs from zero significantly beyond the expected sampling variability, $\mathcal{O}(1/\sqrt{n})$. In [4] we showed that signals can be

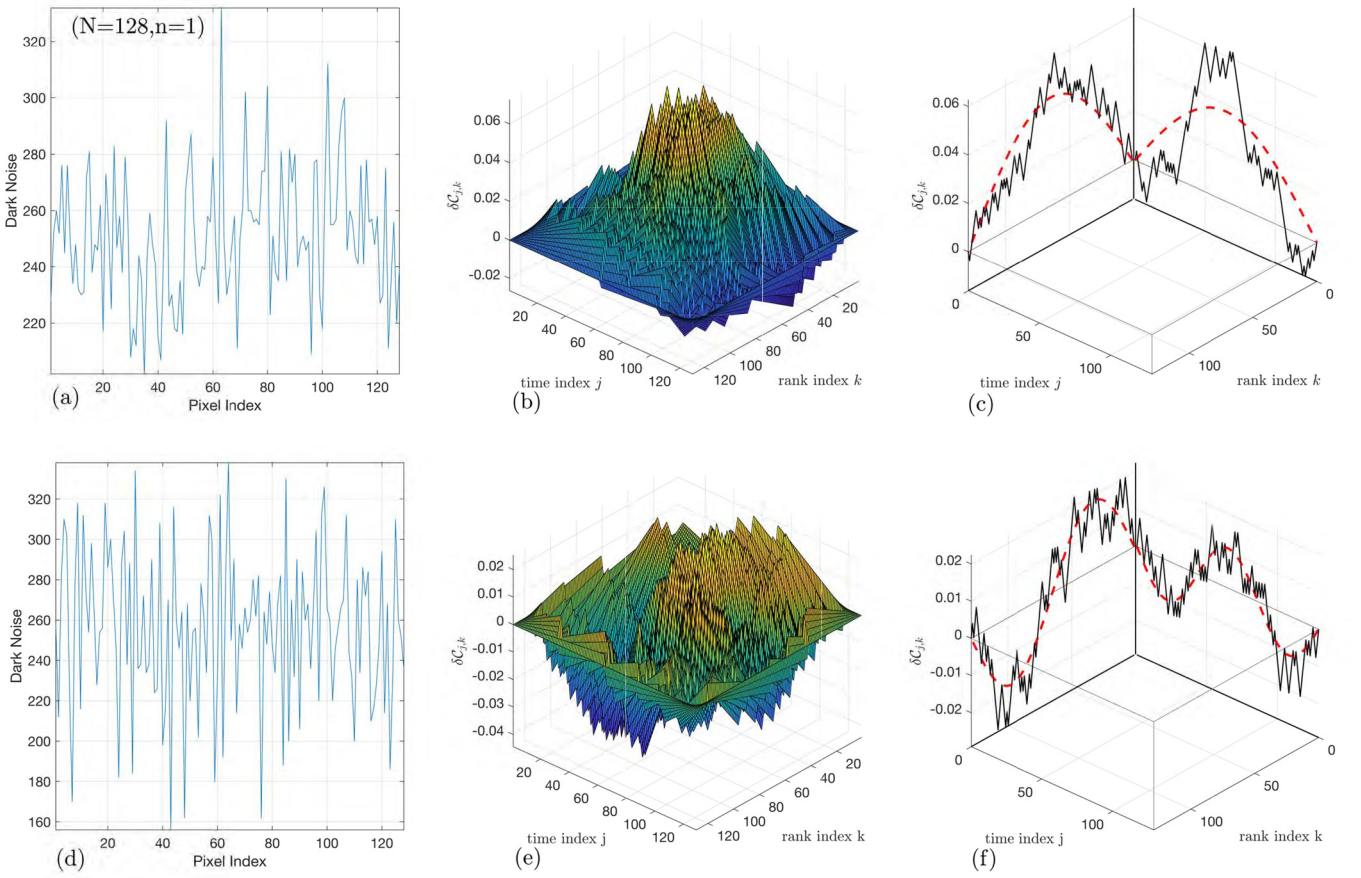

FIG. 3. Rank and time slices of IID-based sample $\delta\mathcal{C}$ [empirical deviation cdf; see Eq. (1)] are pinned random walks. (a) A 128 pixel sample of dark current (IID proxy), recorded as the raw output of a 14 bit A-D converter from 1 s exposure on a Canon R5 camera, in the absence of incident photons (lid closed). (b) $\delta\mathcal{C}$ calculated from panel (a) data. (c) The middle row and column slices of panel (b) $\delta\mathcal{C}$: rank and time are statistically equivalent for IID processes. Panels (d)–(f) parallel the top three panels but for a 128 pixel segment taken from a different row of the image. As illustrated in panels (c) and (f), all IID-based $\delta\mathcal{C}$ slices (rows and columns) execute pinned random walks (see text). The universal eigenfunction expansion (see goodness of fit section) shows that lowest frequency mode ["half wave," red dashed curves in panel (c)] is prominent, accounting for about 61% and 37% of the variance in 1D and 2D IID variability, respectively. In panel (f), the $\delta\mathcal{C}$ slices for panel (d) project most on the second and third mode (on average 15.2% and 6.8% of 1D variance, respectively). For illustrative purposes, traces in panels (c) and (f) were picked among several hundreds for the closest resemblance to pure modes.

detected efficiently on this basis, particularly when employing $\delta\mathcal{C}$-based metrics, supplemented by symmetry considerations, e.g., separating odd and even parity components. In the case of dark current, or residual examination generally, where one is searching for subtle departures from an IID process, variance metrics are more suitable.

In the ensemble limit $n \to \infty$, $\delta\mathcal{C}$ is a perfectly smooth horizontal surface at zero. But Fig. 3 presents single data sequences ($n = 1$). Are Figs. 3(b) and 3(e) rugged $\delta\mathcal{C}$ surfaces within the typical range of IID variability? Analogous traditional metrics suggest $\delta\mathcal{C}_{\rm rms}$ as a measure of such roughness. We define it as

$$\delta\mathcal{C}_{\rm rms} \equiv \sqrt{\overline{\delta\mathcal{C}^2}} \equiv \frac{1}{(N-1)}\left[\sum_{j=1}^{N-1}\sum_{k=1}^{N-1}(\delta\mathcal{C}_{j,k})^2\right]^{1/2}. \quad (3)$$

For the dark current data in Fig. 3(a), Eq. (3) yields a value of $\delta\mathcal{C}_{\rm rms}$ that ≈6% of IID-based samples exceed, but for data in Fig. 3(d), ≈76% of IID-based samples exceed this rms variability. Insofar as both values are within the 2σ (95%) rule, one might conclude on this limited basis that the dark current noise variability is within a typical range for an IID process. As it turns out, a more comprehensive test of a central patch of 128 × 128 pixels confirms this conclusion overwhelmingly. We note in passing that this method reveals significant instrumental bias such as spatial inhomogeneity in the mean count, variance, etc., but only near the sensor edges in a border about 20 pixels wide. For clarity, we do not digress here but press on towards the fundamental characterization of the IID process and its link to random walks.

The slices in Figs. 3(c) and 3(f) differ in the number of zero crossings, prompting one to ask which patterns are more likely to occur for an IID process. Are zero crossings frequent in the IID case? Another natural question is: What kind of time series yields extremes of $\delta\mathcal{C}_{\rm rms}$? A monotonic rank sequence, e.g., $1, 2, \ldots, N$, maximizes it while a rapidly oscillating time series yields minimal $\delta\mathcal{C}_{\rm rms}$ values. The $\delta\mathcal{C}$ shape for the monotonic data sequence is a single hill centered in the rank-time square (zero crossings only at the boundary). This can be seen at the level of the Fig. 2 histograms. While the trace

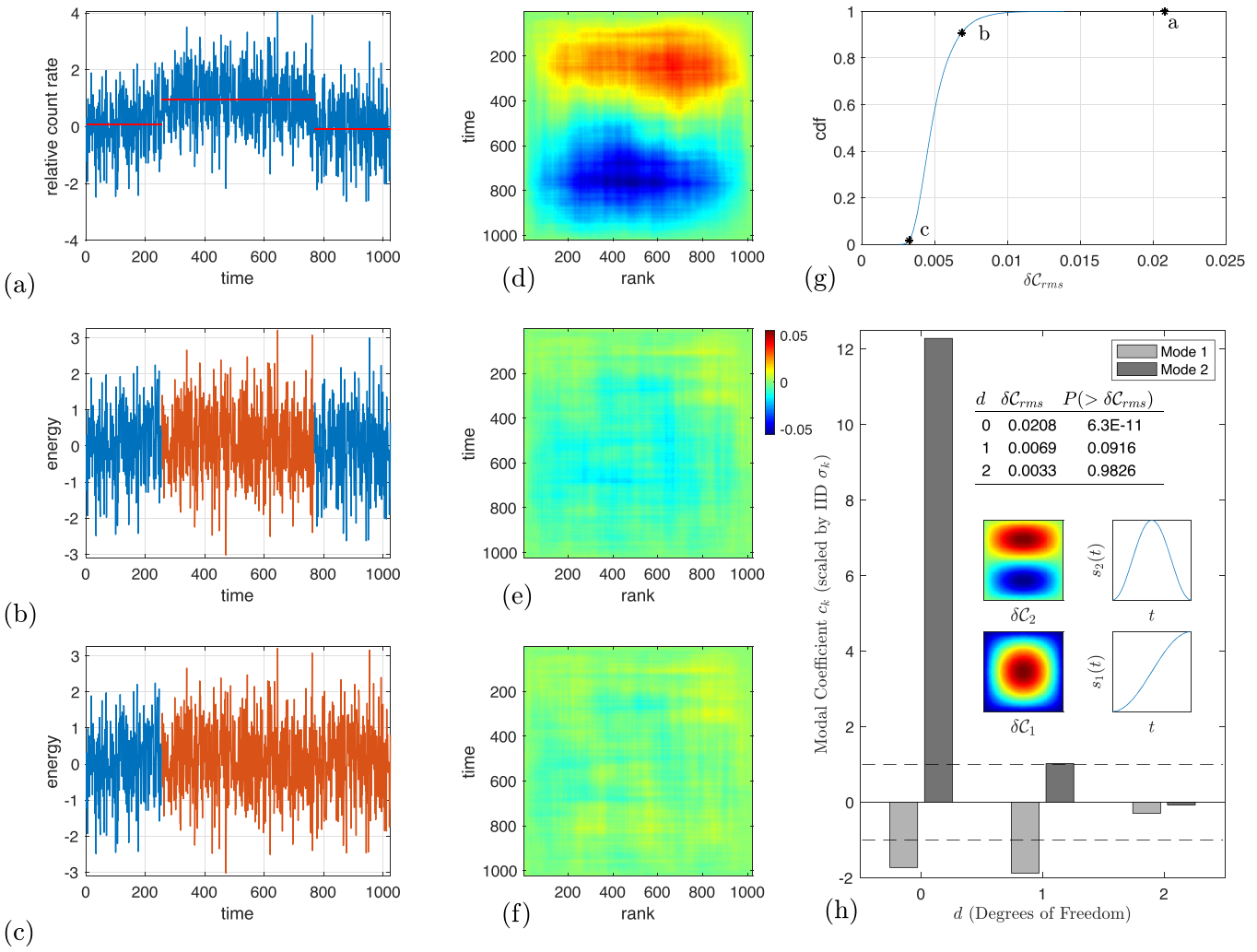

FIG. 4. An agnostic criterion for judging fits to data: To what confidence level is the residual IID? (a) Simulated counts of γ-ray photons, received at the BATSE instrument [21], (apj457785), with three constant rate intervals marked by red and durations determined in [21] via Bayesian change point analysis. As ranking is unaffected by a constant offset, either one or two rate constants are to be fitted. (b) A single parameter to fit ($d = 1$) the middle excess rate relative to the first segment in (a) is subtracted. The residual is shown in red. (c) Data in (a) after removing two excess rates ($d = 2$) as in [21], now second and third segments shown in red. (d) Top view (color map) of the deviation cdf $\delta\mathcal{C}$ for the raw data in (a), with a pronounced "hill-valley" pattern, signaling a strong deviation from IID. (e) $\delta\mathcal{C}$ for the data in (b) (same color range) and no clear pattern. (f) Same as (e) but for data in (c) and no pattern evident. (g) Values of $\delta\mathcal{C}_{\rm rms}$ are overlaid on the IID distribution to answer the question: How likely is a given trace to have arisen from pure noise? Raw data $\delta\mathcal{C}_{\rm rms} = 0.0208$ are exceedingly unlikely, but which is "better": the $d = 1$ or $d = 2$ fit? (In both cases, the residual data are white as judged by δ correlation and a flat FFT.) (h) The $d = 2$ residual results in a rather small $\delta\mathcal{C}_{\rm rms}$, exceeded by pure noise 98% of the time whereas the $d = 1$ fit $\delta\mathcal{C}_{\rm rms}$ residual variance is exceeded by 9.1% of IID realizations (see inset table). The insets are rank-time color maps (top view) of the outstanding modes of $\delta\mathcal{C}$ and the associated signal patterns in time. These results suggest that two parameters are not needed. The new expansion (see text) pinpoints oversuppression of the second mode (darker gray bar) as the cause.

$1, 2, \ldots, N$ maximizes $\delta C_{\rm rms}$, data series that minimize $\delta C_{\rm rms}$ do so by inducing frequent zero crossings in both directions, thus bounding the amplitude of the height swings.

Although the data in Fig. 3(a) are not in a monotonic pattern, even a tiny spurious trend causes the hill in the middle of the associated $\delta\mathcal{C}$ surface in Fig. 3(b), albeit with some asperities. This hill (lowest mode as discussed below) contributes on average 37% to the variance of $\delta\mathcal{C}$. From the random walk perspective, such lowest mode prominence is analogous to the gambler's ruin and arcsin laws, which render frequent zero crossings unlikely in any single realization.

In earlier work on signal extraction [3] we took for granted proper signal extraction minimizing $\delta\mathcal{C}_{\rm rms}$ of the residual. However, we have realized that overfitting is also possible insofar as the residual (data with signal removed) may result in too little variability, i.e., with $\delta\mathcal{C}_{\rm rms}$ that is almost always exceeded by IID noise, as will be illustrated shortly in Fig. 4(g). Furthermore, the $\delta\mathcal{C}_{\rm rms}$ is a restricted metric because its value is disproportionately affected by the "low frequency" components such as the slices of Fig. 3(c) [18]. Both of these concerns are resolved by the universal eigenfunction characterization of the rank-time IID process to which we now turn.

A formal proof of the link between fluctuations of $\delta\mathcal{C}$ and random walks in two dimensions was given in [5] but here we focus primarily on the $n = 1$ case and strive for a more conceptual and intuitive understanding via the notion of $\delta\mathcal{C}$ slices such as those in Figs. 3(c) and 3(f). Consider δC vs time, for instance. Why the clamped ends? All cumulative distributions begin at zero by definition, 1D slices as well as 2D cdf edges. But generally, slices of 2D cdfs are not cdfs themselves, nor are those of the more restricted class of rank-time cdfs. However, as we prove next, even in the case of $n = 1$ (extreme sampling variability), the *deviation* slices and deviation *rank-time* cdfs end at zero as well. This is so because the pure noise cdf "*xy*" baseline is known and can be subtracted off. In addition, the constant row and column sum constraints ensure that the subtraction annuls the ends for each $\delta\mathcal{C}$ realization. Every column sum of $\delta\mathcal{C}$ vanishes because

$$\sum_{i=1}^{N}\left(p_{i,j}-\frac{1}{N^2}\right)=0, \tag{4}$$

as row and column sums of the rank-time matrix are all equal to $1/N$ (for $n = 1$ only one rank can occupy any given time slot). The raw occupancy (permutation) matrix has a single "1" in every row and column. Neither superposition of rank-time (permutation) matrices ($n > 1$) nor normalization (pmf) alters that row and column sum constancy. The present sum of $1/N$ is then canceled by the N-fold sum of the negative constant term in Eq. (4).

Returning to the slices in Figs. 3(c) and 3(f), for the middle rank slices, $\delta\mathcal{C}_{64,k}$, one can separate their calculation into a first "vertical" sum, described by the auxiliary variable

$$\tau_j=\sum_{i=1}^{64}\left(p_{i,j}-\frac{1}{N^2}\right),\quad j=1,\ldots,128, \tag{5}$$

where τ is meant to suggest integration in time. Now the slices plotted in Figs. 3(c) and 3(f) are given by

$$\delta\mathcal{C}_{64,k}=\sum_{j=1}^{k}\tau_j \tag{6}$$

so that the walk is a sum over the RV τ_j. Why is it pinned at the end, i.e., $\delta\mathcal{C}_{64,128} = 0$? Because upon substituting Eq. (5) into Eq. (6), setting $k = 128$, and interchanging the order of the two sums, the inner sum $\sum_{j=1}^{128}$ becomes a complete row sum in the form (4) which, as noted above, vanishes. The remaining sum $\sum_{i=1}^{64}$ is hence a sum of 64 zeros and thus the walk—each rank slice—is pinned. One simply reverses the argument for the time slices [19]. In summary, the additional (permutation matrix) constraints ensure that pinning holds for each realization.

We now turn to the "random" element in the random walk (understood throughout the paper as a walk with IID increments) as in the slices in Figs. 3(c) and 3(f). Whether the steps τ_j come from a true IID process, a correlated one, or one with trends, the individual slices may look similar at first glance but the process information is stored in the texture; e.g., positively correlated RVs might produce a smoother trajectory than, say, $\delta\mathcal{C}_{64,k}$ in Fig. 3(c). It is only for an IID process that the increments τ_j are independent and each $\delta\mathcal{C}$ slice such as in Eq. (6) accumulates τ_js as a pinned walk with independent increments.

Rank-time exchange symmetry and independence for IID processes permit factorization of the 2D underlying covariance matrices as a Kronecker product of their 1D counterparts [5]. However, our extension to the rank-time plane is beyond a simple outer product as all the $2(N-1)$ slices contribute to the detection of departures from IID noise. For all slices, sample-to-sample fluctuations of IID-based walks obey the laws of Brownian bridges [5]. To see this link for the discrete rank-time case, define

$$\tilde{X}_k=\sum_{j=1}^{k}r_j,\quad k=1,\ldots,N, \tag{7}$$

where r_j is the jth rank, chosen by sampling without replacement from the integers 1 to N. The ensemble consists of the $N!$ permutations of rank. For IID noise, all such permutations are equally likely. We can render Eq. (7) a random walk beginning at the origin by defining $\tilde{X}_0 = 0$. But the terminus is always $\tilde{X}_N = N(N+1)/2$ as the ordering of the fixed complement of steps varies.

A pinned (returning to the origin) random walk results if we subtract the mean step size, $(N+1)/2$, from each term. Dropping the tilde, we have

$$X_k=\sum_{j=1}^{k}(r_j-(N+1)/2),\quad k=1,\ldots,N,\quad X_0=0. \tag{8}$$

The covariance matrix for X_k is then

$$\mathcal{K}_{X_j,X_k}=\mathbf{E}[(X_j-\mathbf{E}[X_j])\,(X_k-\mathbf{E}[X_k])],$$

where $\mathbf{E}$ is the expectation value over an ensemble. With Eq. (8) we have that $\mathbf{E}(X_i) = 0$ and hence

$$\mathcal{K}_{X_j,X_k}=\mathbf{E}[X_jX_k].$$

After some algebra, the following results:

$$\mathcal{K}_{X_j,X_k}\equiv\mathcal{K}_{j,k}:=\frac{N+1}{12}\,(N\,\min(j,k)-j\,k), \tag{9}$$

and the associated eigenvalue problem

$$\mathcal{K}\,\psi_m=\mu_m\psi_m, \tag{10}$$

for which

$$\psi_m(x_k)=\sqrt{\frac{2}{N}}\,\sin(m\,x_k), \tag{11}$$

$$\mu_m=\frac{N\,(N+1)}{24\,(1-\cos(m\pi/N))},\qquad (k,m)=1,\ldots,N-1, \tag{12}$$

where $x_k = k\pi/N$.

In common application of the discrete Karhunen-Loève transform (or principal component analysis), the covariance matrix is based on observation. In that context, the eigenvalues μ_m are the main concern as they indicate the amount of signal variance captured by a proposed low-order approximation. Higher modes are discarded, reflecting either noise or unresolved structure. Our approach here is distinct, a hybrid. The covariance matrix is exact, deriving from the $N!$ permutations

of Eq. (8). But, unlike the continuous eigenvalue problem for the Brownian bridge, the eigenvalue problem (10) is discrete.

Moreover, the approach to signal extraction is not tendered as a problem in variance capture because, unlike the empirical setting, here one knows that the modal expansion coefficients in the case of IID noise are independent RVs. When suitably normalized they follow an $\mathcal{N}(0,\sigma)$ distribution, where σ is the standard deviation of the noise. The relevant standard for a signal is hence the determination of which normalized components exceed a stated confidence level. That normalization is not that given by μ_m. Rather, we appeal to the singular value decomposition, for which the columns of U are the eigenvectors (modes) ψ_m, and then the appropriate weight is $\sigma_m/\sqrt{N!} = \sqrt{\mu_m}$ where σ_m is the mth singular value.

The low frequency modes, $m = 1, 2$, are shown in dashed red lines in Figs. 3(c) and 3(f), respectively. It can be seen from these panels that the projections on the low frequency modes are larger than on the high frequency ones. In fact, the discrete IID process is uniquely characterized by the decay of such projections; that is, the eigenvalues in Eq. (12) initially decay as m^{-1}. With increasing m, the eigenvalues exceed the pure algebraic rate of m^{-1} by a factor which reaches $\pi/2$ for $m = (N-1)$ (for Brownian bridges the decay is pure m^{-1} [20]).

Due to the IID rank-time equivalence and the associated Kronecker factorization of the underlying covariance matrices, the eigenvalues for fluctuations of the IID $\delta\mathcal{C}$ in two dimensions are given by the product of the 1D solutions:

$$\lambda_{n,m} = \frac{1}{2N\sqrt{N-1}} \sqrt{\frac{1}{(1-\cos(m\pi/N))\,(1-\cos(n\pi/N))}}, \tag{13}$$

where $\{m, n\} = 1, \ldots, N-1$ and the eigenfunction associated with $\lambda_{n,m}$ is

$$\psi_{n,m} = \psi_n(t_j)\,\psi_m(r_k), \quad \{j,k\} = 1, 2, \ldots, N-1, \tag{14}$$

with the component eigenfunctions on the right, given by Eq. (11). Because of the rank-time symmetry of IID processes, there is the $\lambda_{n,m} = \lambda_{m,n}$ degeneracy.

We can now understand the emergence of a discernible hill in both Figs. 3(b) and 3(e). The average variance captured by the hill, that is, the lowest 2D mode $\psi(1,1)$, is given by

$$\frac{\lambda_{1,1}^2}{\sum_{m=1}^{N-1}\sum_{n=1}^{N-1}\lambda_{n,m}^2}.$$

This ratio can be approximated by

$$\left[\sum_{m=1}^{\infty}\sum_{n=1}^{\infty}\frac{1}{m^2\,n^2}\right]^{-1} = \frac{36}{\pi^4} \approx 0.37\ldots.$$

Loosely speaking, so significant a fraction in one mode means that quite often, as in the dark current noise surface plot of Fig. 3(b), one sees a large mound in the center. Correlations (positive or negative), trends, and all manner of signals will alter the decay of the modal coefficients. Thus, *the modal spectrum of Eq. (13) constitutes a signature of the IID process as represented by the fluctuating $\delta\mathcal{C}$ in rank time.* There are $2(N-1)$ slices, e.g., $2\times(127)$ in Figs. 3(b) and 3(e) and $(127)^2$ modes and sample eigenvalues. All departures of the data modal coefficients from the IID eigenvalues carry information about possible departures from an IID process.

TABLE I. The lowest mode, $\psi_{1,1}$ (hill) captures 37% of the IID variance, on average. The $\delta\mathcal{C}$ of Fig. 3(b) has nearly twice the expected amount and mode $\psi_{3,1}$ at 14% triples the expected value. With these two accounting for 84% of the total variance, there is a deficit of other modes, e.g., the mode $\psi_{1,2}$ accounting only for 0.5% instead of the expected 9%. In contrast, Fig. 3(e) has both modes $\psi_{1,2}$ and $\psi_{3,1}$ in abundance.

	$\psi_{1,1}$ (% var)	$\psi_{1,2}$ (% var)	$\psi_{3,1}$ (% var)
Fig. 3(b)	2.233 (0.700)	−0.187 (0.005)	0.995 (0.139)
Fig. 3(e)	−0.675 (0.197)	−0.745 (0.239)	0.622 (0.166)
$\sigma = \lambda_{j,k}$	1.151 (0.370)	0.576 (0.092)	0.384 (0.041)

Goodness of fit

The IID spectrum can serve as a "goodness of fit" matrix, once a signal has been removed. For example, the expected 37% of $\psi_{1,1}$ for the IID process differs from the actual projection of the data in Fig. 3(b) onto the lowest mode (the hill) which accounts for 70% of the total variance. In contrast, as Table I details, Fig. 3(b) has a striking deficit of $\psi_{1,2}$ whereas Fig. 3(e) has a considerable deficit of the hill mode, $\psi_{1,1}$, but an abundance of $\psi_{1,2}$ and $\psi_{3,1}$. Is this a typical sampling variability of a pure IID process? Not quite so because the authors went through several hundred samples in search of the visually striking slices of Figs. 3(c) and 3(f). In the absence of any *a priori* information, when faced with such data, one would be compelled to test the data further for a presence of a signal. We advocate an exploratory and heuristic application of the modal spectrum in such circumstances, particularly because the results are distribution invariant and the modes are associated with signal forms.

As an example of such association, given the improbably large hill in Fig. 3(b), is there a hint of a monotone (e.g., linear) trend associated with the time series of rank in Fig. 3(a) but not at all in Fig. 3(d)? Indeed, least squares fits to the two series yield a positive slope for Fig. 3(a) and a negative one for Fig. 3(d), but with huge 95% confidence intervals for both, consistent with the sampling variability.

A more thorough examination of dark current noise was carried out with data taken from a 128×128 patch in the center of the CMOS sensor, thus $(N = 128, n = 128)$. The full modal expansion for this $\delta\mathcal{C}$ has then $(127)^2$ components and we verified that the distribution of expansion coefficients was consistent with IID noise. Specifically, as $N\to\infty$, the distribution of each of the modal coefficients, scaled by their respective eigenvalues, tends to $\mathcal{N}(0,1)$: a zero mean, unit variance, normal pdf. One would miss this, were one to scale the modal coefficients by the m^{-1} eigenvalues of the continuous Wiener process, e.g., as in [20]. Having the spectrum (13) opens up a variety of tests for departure from pure IID noise including, e.g., the Kolmogorov-Smirnov and Anderson-Darling tests [22]. The spectrum may also serve the needs of instrument calibration and monitoring [23,24] such as, say, occasionally pointing a spaceborne radiation detector at a local source of noise and testing for an IID response.

Note that a simple average, $\overline{\delta\mathcal{C}}$, of Eq. (2) can also be used as a quick additional test on a plausible presence of the signal. How close is it to zero (the ensemble average for the IID process)? For Figs. 3(b) and 3(e), the calculation yields, in units of standard deviation, -0.69 and 1.63, respectively, neither large enough to suspect a signal-based trend.

We next consider an application where some *a priori* information about a signal is available and signal extraction is parametric. The goodness of the signal extraction is then judged by asking: Is the residual (data with signal removed) compatible with IID noise, within the bounds of sampling variability? The data come from an astrophysical context: modeling arrivals of gamma-ray photons at the BATSE instrument [21], as shown in Fig. 4(a). The *a priori* information here is that the signal is piecewise constant and the breakpoints are known and confirmed in [21] by Bayesian statistical analysis. In other words, the extent of the constant intervals is given, and the question is: How many rates are needed? Recall that ranking is indifferent to an overall constant offset and, therefore, our question boils down to: *One or two rate constants (parameters) to fit*?

To that end, consider the traces shown in Figs. 4(b) and 4(c). In Fig. 4(b), a single parameter ($d = 1$) measuring the excess rate of the middle segment compared to the first segment in Fig. 4(a) is subtracted, with the middle residual shown in red. In Fig. 4(c), parameters for two excess rates ($d = 2$) are determined, as in [21] with both residual segments now shown in red. Figure 4(d) presents the deviation cumulative distribution $\delta\mathcal{C}$ for the raw data in Fig. 4(a) as a color map, that is, a top view on the surface plot, with the z-axis values color coded as prescribed by the color bar in Fig. 4(e) and the same for all three middle panels. The dipolar pattern in Fig. 4(d) is pronounced, signifying a large departure from IID noise and associated with the "down-up-down" pattern in raw data [or, more formally, the plot of s_2 in the inset of Fig. 4(h)]. The eigenfunction, closest to the $\delta\mathcal{C}(r, t)$ pattern in Fig. 4(d) is $\psi_{2,1}$, with the peak and valley in rank slices. In contrast, $\delta\mathcal{C}$ in Fig. 4(e), corresponding to the Fig. 4(b) trace, shows no such pattern and neither does Fig. 4(f), for the data in Fig. 4(c). So which one to prefer: the two-parameter fit ($d = 2$) or the one-parameter fit ($d = 1$)? Is it the $\delta\mathcal{C}(r, t)$ in Fig. 4(e) or Fig. 4(f)?

To that end, consider the rms values for the three $\delta\mathcal{C}$s, placed on the IID distribution shown in Fig. 4(g). This curve quantifies the likelihood that a given trace arose from an IID process. Although $\delta\mathcal{C}_{\rm rms} = 0.0208$ of the raw data (point a on the curve) is exceedingly unlikely as documented in inset table of Fig. 4(h), points b and c are less obvious. Which is "better": $d = 1$ (point b) or $d = 2$ (point c) fit? In both cases, the residual data are white, confirmed by δ correlation in time and flat (FFT) spectrum. The $d = 2$ residual results in a rather small $\delta\mathcal{C}_{\rm rms}$, exceeded by 98% of the IID noise cases, whereas the residual variance of $d = 1$ fit $\delta\mathcal{C}_{\rm rms}$ residual variance is exceeded by only 9.1% of IID realizations as documented in the inset table of Fig. 4(h). These results hint that the fit of Fig. 4(c) might not be necessary because the fit of Fig. 4(b) is already in the "vicinity" of IID noise, i.e., within normal sampling variability. By Occam's razor, one favors the simpler option.

To explore further, we compare the modal spectrum of the data to that of Eq. (13) for IID noise and the results are shown in gray bars of Fig. 4(h) for all three cases. Note that the ordinate gives modal amplitudes, normalized by their respective standard deviations. As noted above, for IID noise, these scaled amplitudes are normally distributed $\mathcal{N}(0, 1)$ RVs. It is apparent that the tall bar of the raw data ($d = 0$ on the x axis), associated with the dipolar pattern (second mode), $\psi_{2,1}$ is of overwhelming statistical significance, $>12\sigma$. In contrast, the $\psi_{2,1}$ projection for the one-parameter fit data of Fig. 4(b) is already well within the sampling variability of IID noise, $\approx \sigma$, and it is suppressed in the two-parameter fit by another order of magnitude. Insofar as the $d = 1$ option is already within the range for both modes, $d = 2$ is not required. Note that the sinusoidal signal forms $s_1(t)$ and $s_2(t)$, shown in the insets of Fig. 4(h), in the (practically most important) weak signal limit and for large n, correspond exactly to the time series that excite modes $\psi_{1,1}$ and $\psi_{2,1}$, respectively. Unlike the strong dipolar pattern of $\psi_{2,1}$ in Fig. 4(d), which distorts rank-time symmetry, $\psi_{1,2}$ should not be much affected by the excess rate, but remain within IID bounds and indeed, the $\psi_{1,2}$ projection is 1.95σ for the raw data ($d = 0$) and 0.44σ for ($d = 1$).

For the data in Fig. 4(a) we have limited ourselves to a small number of modes. However, if one seeks to explore the full spectrum of modes, a value $N \sim 10^4$ may become prohibitive and so we developed 1D alternatives, described in Appendix B.

Depending on the context, further testing may be warranted and the method supplies ample means to do so. The metric $\overline{\delta\mathcal{C}}$ exactly cancels the large antisymmetric ("hill-valley") $\psi_{2,1}$ contribution in Fig. 4(d). This suggests a weighted mean version of Eq. (2) as follows:

$$\frac{1}{(N-1)^2}\sum_{j=1}^{N-1}\sum_{k=1}^{N-1}\mathrm{sgn}(j-(N+1)/2)\,\delta\mathcal{C}_{j,k}, \qquad (15)$$

which responds to contributions from $\psi_{2m,1}$ including $\psi_{2,1}$, and annuls those from $\psi_{2m-1,1}$, e.g., $\psi_{1,1}$. When applied to Figs. 4(a)–4(c), this alternate form, which is sensitive to symmetric signals in time, yields results similar to Fig. 4(h), with similar conclusions.

To summarize the evolution of our views on $\delta\mathcal{C}$-based criteria, in [3] we annulled a quantity similar to $\overline{\delta\mathcal{C}}$ as the criterion for signal extraction of a trend, while symmetry-based $\delta\mathcal{C}$ criteria such as odd, even, and other 2D symmetry group-based metrics were used in [4] for detection (as opposed to extraction) of arbitrary signals. Here, on the other hand, our emphasis is to complement and test any method of signal extraction (including our own developed in [5]) by using the universal spectrum of Eq. (13) for testing the residual.

V. CONCLUDING REMARKS

In contrast to our earlier work on detection and extraction of unknown signals in arbitrary noise [3–5], here we examined those properties of data ranking that can be used in parallel with any other method of fitting data, extracting signals, or for instrument calibration purposes. To that end, we characterized an IID process and its sampling variability in the rank-time

space by finding the distribution-invariant spectrum (12) and by employing it to evaluate residuals (data with signal removed). The relevant question is whether further extraction or fitting is warranted. This has been illustrated on dark current and gamma-ray arrival data of Figs. 3 and 4.

A skeptic might argue that rank-based distribution invariance does not add much to existing methods: after all, the very notion of white noise is distribution invariant. But, as we point out in Fig. 1, not all white noise is created equal and our method distinguishes the variants of white noise, thus reaching beyond the traditional spectral analysis. Furthermore, our distribution-invariant eigenvalue spectrum extends well beyond rank, applying to pinned walks with nearly any imaginable steps and derived in two distinct settings: discrete integer steps and order statistics [5]. Also, unlike the white spectrum, ours is convergent. In fact, the 2D spectrum furnishes a tool to address goodness of fit well beyond any simple scalar metrics. It allows for *a priori* information about signals to be incorporated in a heuristic and exploratory manner, e.g., as in the parametric fits of Fig. 4, which drew on only a small sample of that spectrum.

Another practical question that is often raised in applications, e.g., hydrology [25,26], is whether the given data are only wide-sense stationary (WSS) or strictly stationary. As Figs. 1(d) and 1(h) demonstrate, ranking can deliver the answer, detecting time-dependent kurtosis in manifestly wide-sense stationary data (and stationary skewness as well). This type of rank-induced entanglement of moments is due to the global nature of rank and holds promise for many practical applications.

Rank-based signal-noise decomposition is well suited for instrument calibration and monitoring as well as for weak signal extraction. Even when a weak and white chaotic signal such as in Fig. 1(c) is embedded in strong IID noise, rank still senses the subtle white chaos-induced perturbation well beyond the sensitivity of the common delay plot. Of the three departures from an IID process in Fig. 1, surprisingly, it is the "strong white noise" for which the nonindependence is detected with the smallest number of required time series n.

To conclude, we think that linking such disparate fields as ranking in data analysis and random walks in science will prove to be fruitful because both have such a broad range of applications. Who would have guessed that a random walk with independent increments picture (e.g., a freely jointed chain model in polymer physics) would have anything to contribute to characterizing the (purest noise) IID process in rank time as in Figs. 3 and 4? It is hoped that the community of statistical physicists, well versed with random walks, will find this link of interest.

ACKNOWLEDGMENTS

This work was supported by NSF Grant No. AGS-1639868. We are grateful to Professor Yeonwoo Rho for helpful comments and Daniel Savio for a careful reading of the manuscript.

APPENDIX A: RANKING PATTERNS VERSUS DEPARTURES FROM STRICT WHITE NOISE

Towards interpreting Figs. 1(e)–1(h) we ask: What does it take to disturb the rank uniformity of Fig. 1(e)? To that end, consider ranking the shortest possible time series of just two successive continuous RVs, (x, y), for time slots 1 and 2, respectively. In each trial the lesser of the two is assigned rank 1, the higher rank 2. After accumulating trials, we build a 2D histogram whose entries are the probability estimates vs time and rank [27]. As the number of trials $n \to \infty$, the probability for time slot 1 and rank 2 approaches

$$P(X > Y) = \int_{x=-\infty}^{\infty} dx \int_{y=-\infty}^{x} dy\, f_{X,Y}(x, y), \tag{A1}$$

where $f_{X,Y}(x, y)$ is the joint pdf and, without any loss of generality, we take $x, y \in \mathbb{R}$. For the case of IID white noise, we have

$$f_{X,Y}(x, y) = f(x)\, f(y)$$

with no subscript needed on the right as there is but a single 1D pdf. Then the expression reduces to

$$\int_{x=-\infty}^{\infty} dx\, f(x) \int_{y=-\infty}^{x} dy\, f(y) = \int_{x=-\infty}^{\infty} dx\, f(x)\, F(x) = \frac{1}{2} F(x)^2 \Big|_{-\infty}^{\infty} = \frac{1}{2} \tag{A2}$$

as the integrand is a perfect derivative.

We now loosen the IID constraint by keeping the independence (pdf factorization) but dropping the identical distribution requirement, e.g., by introducing subtle nonstationarity in our short time series, as in Figs. 1(d) and 1(h) where the kurtosis parabola is a random signal. In this case

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y)$$

and subscripts are needed on the right-hand side as the pdf evolves from one time slot to the next. Substitution of this as in Eq. (A2) leaves again a 1D integral

$$\int_{x=-\infty}^{\infty} dx\, f_X(x)\, F_Y(x),$$

but there is no longer an exact differential. Our pmf for the nonstationary variables is thus no longer the uniform distribution [with value of $1/4$ for the 2D pdf or value of $1/2$ for the 1D version of Fig. 1(e)].

Strong but not strict (IID) white noise can be strictly stationary where x and y are uncorrelated but not independent as in Figs. 1(b) and 1(f). Here too Eq. (A1) has an irreducible double integral yielding generally nonuniform pmf. One thus concludes that this pmf and, by extension, any $N \times N$ pmf, such as the 32×32 rank-time matrix of Fig. 2, is guaranteed of being uniformly populated only when the RVs are IID.

To address Figs. 1(c) and 1(g), we evaluate the departure of the ranking distribution from uniformity for the logistic map $x_{k+1} = a x_k\,(1 - x_k)$, uncorrelated at the chosen $a = 4$ [28]. The 1D pdf for that logistic case is given by

$$p(x) = \frac{2}{\pi \sqrt{1 - 4\,(x - 1/2)^2}} \tag{A3}$$

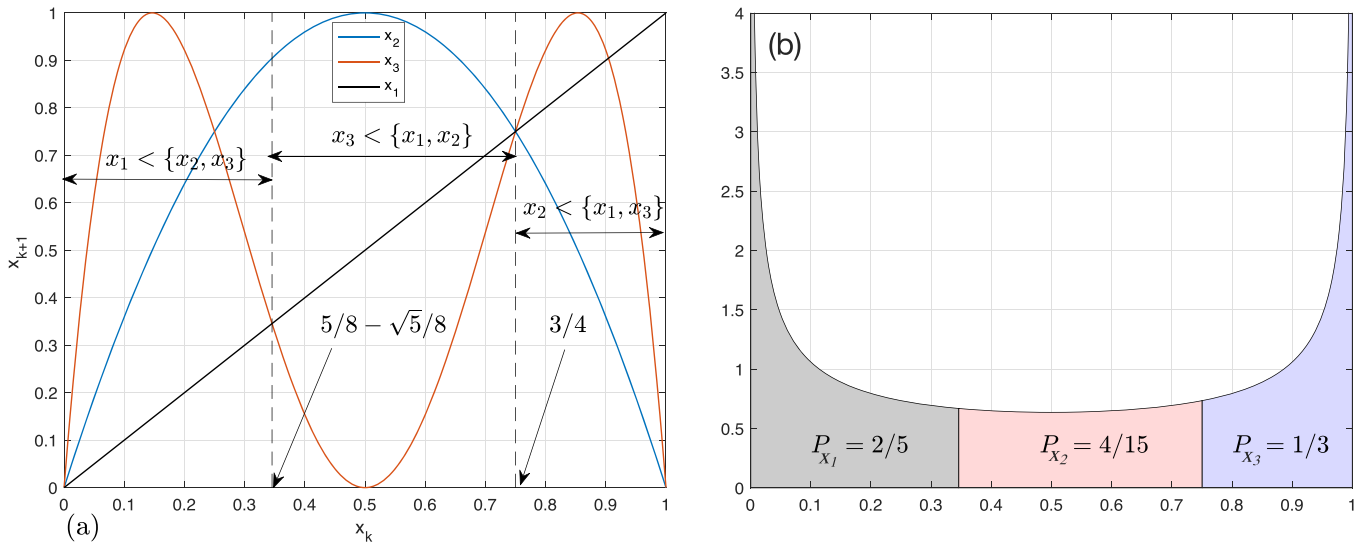

FIG. 5. Ranking of the logistic map $x_{k+1} = 4x_k\,(1 - x_k)$ for $N = 3$ (ranks 1,2,3). (a) Three domains where each of the consecutive variables $\{x_1, x_2, x_3\}$ is the smallest (rank 1). (b) The probability that x_1 is rank 1: $\int_0^{5/8-\sqrt{5}/8} p(x)\,dx = 2/5$. Integrating from 3/4 to 1 gives a probability of 1/3 for x_2 to be rank 1, and the middle third yields 4/15 for the probability that x_3 is the lowest rank. Note that the middle point x_2 attains the probability expected for all three time slots with an IID process. Hence, the departure from uniformity is an "end effect." The three areas in panel (b) show that a symmetrical scheme for rank would obtain were the first border moved back to 1/4, thereby yielding three equal probabilities of 1/3.

and consequently the baseline value of σ_{data} in Fig. 1(g) is

$$\sqrt{\int_{-1/2}^{1/2} \frac{2y^2\,dy}{\pi\sqrt{1-4\,y^2}}} = \frac{1}{2^{3/2}} \approx 0.35.$$

Yet the ranking pattern is not uniform even for this *white* chaos, as we now demonstrate for the case of $N = 3$.

Denote the RVs in time slots 1,2,3 as x_1, x_2, x_3. The argument, presented in Fig. 5, relies on the logistic map and Eq. (A3). In Fig. 5(a) we iterate the map $x_{k+1} = 4\,x_k\,(1 - x_k)$ twice to record the relations among the x_k and one sees that x_1 lies the lowest of the three in the range $[0, 5/8 - \sqrt{5}/8]$, succeeded by a second range in which x_3 is the lowest rank and then to the right, the final interval where x_2 falls the lowest. The pdf (A3) for x_k is plotted in Fig. 5(b), where we then integrate over the three domains to find the probability that each of the x_k is rank 1, resulting in $[2/5, 1/3, 4/15]$ for the probabilities that x_1, x_2, x_3 attain the lowest rank $r = 1$, as recorded in the first column in Table II. The same reasoning is used for the ranks $r = 2$ and $r = 3$, giving the second and third columns of the probability matrix.

TABLE II. Entries in the first three columns constitute the 3×3 pmf $p_{j,k}$, where j is the row index referring to slots labeled as $x_{1,2,3}$, and k the column index referring to rank slots $r_{1,2,3}$. With this matrix one can calculate all rank moments. The first two are the mean rank, in the fourth column, and mean square rank in the fifth.

	$r = 1$	$r = 2$	$r = 3$	$\langle \mathbf{r} \rangle$	$\langle \mathbf{r}^2 \rangle$
x_1	2/5	2/5	1/5	9/5	19/5
x_2	1/3	1/3	1/3	2	14/3
x_3	4/15	4/15	7/15	11/5	83/15

This discrete 2D pmf is all one needs to compute any desired rank moment. For example, the mean rank at time index j is as follows:

$$\langle r_j \rangle = \sum_{k=1}^{N} k p_{j,k}, \tag{A4}$$

where $p_{j,k}$ is the entry from Table II indicating the probability that index x_j achieves rank k [29]. The second moment follows in the same manner, simply replacing k by k^2 in Eq. (A4) with the result as given in Table II. From $\sigma = \sqrt{E[X^2] - (E[X])^2}$, we can then compute, e.g., σ_{rank} for the second slot as $\sqrt{14/3 - (2)^2} = \sqrt{2/3}$, which is the standard deviation of [1,2,3], as similarly for the $N = 32$ result stated in regard to Fig. 1(e).

Whenever the pmf is symmetric with respect to rotation about its vertical midline, mean rank $\langle \mathbf{r} \rangle$ is uniform, namely, $(N + 1)/2$. Because this occurs for two of the three non-IID cases in Fig. 1, rank variability σ_{rank} was chosen instead as it is nonuniform in all three cases.

APPENDIX B: ONE-DIMENSIONAL APPROXIMATIONS

Here, our concern is with the *finite* discrete process of a pinned random walk of N steps, for which the corresponding discrete covariance function assumes the general form (9). It is reminiscent of the result in the theory of *continuous* stochastic processes, the so-called Brownian bridge of a Wiener (Brownian particle) process. The bridge (pinned at both ends) for W (Wiener process RV) is defined as $B_t = W_t - tW_1$. It has a continuous covariance function of the form [20,30]

$$K(t, s) = \min(t, s) - t\,s. \tag{B1}$$

The principal components of an orthogonal function expansion of the Karhunen-Loève type [8,31] are then eigensolution pairs

$$f(x) = \sin(k\pi x), \quad g(y) = \frac{1}{k^2\pi^2}\sin(k\pi y), \quad k = 1, 2, \ldots$$

for the integral equation [20]

$$g(y) = \int_0^1 K(x, y)\, f(x)\, dx.$$

The Gaussian assumption, via an application of the central limit theorem, underpins these results. In contrast, our discrete results (11) and (13) require no such assumption as discussed next [32].

Motivated by the notion of accumulated rank in $\delta\mathcal{C}$, our 1D random walks are defined by the steps chosen from a finite set $\{z_k\}$ for $k = 1, \ldots, N$ by sampling without replacement. The (time) slots are indistinguishable, meaning that at every slot all ranks are equally likely to occur. Thus, all permutations of the integers 1 to N occur with probability of $1/N!$. When this holds, the covariance matrix is of a universal form (9). For a given rank vector $\mathbf{r}$ with elements r_k we hence define a pinned walk as

$$\rho_k = \begin{cases} 0, & k = 0 \\ \sum_{j=1}^{k} (r_j - (N+1)/2), & k = 1, \ldots N, \end{cases} \tag{B2}$$

where the zeroth element is introduced to pin the walk at the origin, $\rho_0 = 0$. All walks contain the same elements: only one rank per time slot, no replacement. Thus, subtracting the mean of the integers 1 to N yields $\rho_N = 0$. For the elementary example used in the Introduction, [0.94, 1.87, 0.60] converts to 2,3,1, $(N+1)/2 = 2$, $\rho_1 = 0$, $\rho_2 = 1$, $\rho_3 = 0$, pinned as expected.

Note that the *ensemble* ($n \to \infty$) average of r_j is $(N+1)/2$—not because of the single trial average invoked above, but because for an IID process, all ranks are equally likely at all sites. Thus, ρ_k approaches zero pointwise, just as $\delta\mathcal{C}$ tends everywhere to zero $\sim n^{-1/2}$.

As the covariance matrix (9) holds for the pinned walk ρ, the vectors in Eq. (11) form the eigenset and one can use the expansion coefficients

$$c_m = \sum_{k=1}^{N-1} \psi_m(x_k)\, \rho_k \tag{B3}$$

to explore departures of a given dataset from an IID process. The normalized set $\{c_m/\lambda_m\}$ is, for IID noise, a set of $\mathcal{N}(0, \sigma)$ RVs. Departures from this normality beyond sampling variability may suggest the presence of a signal. For the example of rank, $\sigma = \sqrt{N(N+1)/12}$. There is an associated Plancherel- or Parseval-type identity of note:

$$\sum_{m=1}^{N-1} \left(\frac{c_m}{\lambda_m}\right)^2 = (N-1)\,\sigma^2, \tag{B4}$$

which places a bound on the maximal range for any c_m. Strictly speaking, Eq. (B4) shows that the distribution of $\{c_m\}$ cannot be $\mathcal{N}(0, \sigma)$ as the latter can have no finite bound. However, that distribution is asymptotic with N and is entirely adequate in practice, even for modest N.

This 1D scheme can be linked to the 2D formulation in terms of $\delta\mathcal{C}$ via

$$\rho_j = -(N-1) \sum_{k=1}^{N-1} \delta\mathcal{C}_{j,k}. \tag{B5}$$

Notation for the continuous version is simpler:

$$\rho(t) = \int \delta\mathcal{C}(t, r)\, dr.$$

In other words, this random walk based on rank is the horizontal (rank) average of $\delta\mathcal{C}$, that is an average over all the pinned walk rank slices in, say, Fig. 3(b), which would include the middle slice in Fig. 3(c). One can anticipate from the surface plot that the resulting rank average will retain a hill of intermediate height.

The random walk may also be based on the plotted entries in Fig. 2(b). Recalling rank-time equivalence of IID noise, Fig. 2(c) then suggests a parallel line of development, namely, a pinned random walk based on time as a function of rank:

$$\tau_k = \sum_{j=1}^{k} (t_j - (N+1)/2) = -(N-1) \sum_{j=1}^{N-1} \delta\mathcal{C}_{j,k}. \tag{B6}$$

The mean of $\delta\mathcal{C}$ is now alternately given by

$$\overline{\delta\mathcal{C}} = -\frac{1}{(N-1)^3} \sum_{k=1}^{N-1} \tau_k = -\frac{1}{(N-1)^3} \sum_{j=1}^{N-1} \rho_j. \tag{B7}$$

So the net area under the curve of the pinned random walk ρ is identical to that for the pinned random walk τ. However, individual area distributions are wholly independent and the two sets of projections may be assessed independently for departures from IID noise.

This formulation addresses the most difficult case for $n = 1$. For $n > 1$, we simply form

$$\rho_k = \frac{1}{n} \sum_{j=1}^{n} \rho_k^{(j)}, \tag{B8}$$

an average over the pinned random walk for each trial, and then

$$\rho_j = -(N-1) \sum_{k=1}^{N-1} \delta\mathcal{C}_{j,k} \tag{B9}$$

and similarly for τ. The identity (B4) no longer holds and the sum fluctuates via

$$\sum_{m=1}^{N-1} \left(\frac{\mathbf{c}_m}{\lambda_m}\right)^2 = \sigma^2 \left[\frac{N-1}{n} \pm \frac{\sqrt{2\,(N-1)}}{n}\right]. \tag{B10}$$

In addition, the $c_m^{(j)}$ from individual trials approach

$$\lim_{n\to\infty} \mathrm{std}\left(\left\{c_m^{(j)}\right\}\right) \to \sigma\lambda_m. \tag{B11}$$

The pair of walks (ρ, τ) does not amount to the information content of $\delta\mathcal{C}$ as they each smear out content in one dimension by averaging [33]. Yet, they each offer an $N-1$ component "fingerprint" with which to assess departure from IID noise. While shy of the exhaustive $(N-1)^2$ matrix fingerprint of $\delta\mathcal{C}$, this condensed version may often suffice. Also, they are

more quickly computed and require negligible storage. Note that either of the last two expressions in Eq. (B7) offers an $\mathcal{O}(N)$ order algorithm to compute the mean value of $\delta\mathcal{C}$. Reliable detection of weak trends in a large data set is thereby facilitated.

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing Volume I: Estimation Theory; Volume II: Detection Theory; Volume III: Practical Algorithm Development* (Prentice-Hall, Englewood Cliffs, NJ, 1998).

[2] S. R. K. Vadali, P. Ray, S. Mula, and P. K. Varshney, IEEE Trans. Commun. **65**, 1061 (2017).

[3] G. Ierley and A. Kostinski, Phys. Rev. X **9**, 031039 (2019).

[4] G. Ierley and A. Kostinski, Phys. Rev. E **102**, 032221 (2020).

[5] G. Ierley and A. Kostinski, Phys. Rev. E **103**, 022130 (2021).

[6] P. J. Bickel and K. A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics* (Simon and Schuster, New York, 1977), pp. 352–353.

[7] N. G. Van Kampen, *Stochastic Processes in Physics and Chemistry*, Vol. 1 (Elsevier, Amsterdam, 1992).

[8] M. B. Priestley, *Spectral Analysis and Time Series* (Academic Press, New York, 1981).

[9] B. Picinbono, *Random Signals and Systems* (Prentice-Hall, Englewood Cliffs, NJ, 1993).

[10] V. Voinov and M. Nikulin, *Unbiased Estimators and Their Applications: Volume 1: Univariate Case*, Vol. 263 (Springer, Berlin, 2012).

[11] This construction is related to the family of GARCH random processes used widely in the statistics of finance (Dr. Yeonwoo Rho, private communication). The pdf for z is given by $p(z) = K_0(|z|)/pi$ where K_0 is the Bessel function and the joint pdf for (z_k, z_{k-j}) is the quadratic product of $p(z)$ for all $j \geqslant 2$ but fails for $j = 1$.

[12] However, we have discovered efficient rank-rank methods where the required n is modest.

[13] Y. A. Reshef, D. N. Reshef, H. K. Finucane, P. C. Sabeti, and M. Mitzenmacher, J. Mach. Learn. Res. **17**, 7406 (2016).

[14] G. J. Székely, M. L. Rizzo, and N. K. Bakirov, Ann. Stat. **35**, 2769 (2007).

[15] Although there are no ties here, when faced with data of limited resolution, throughout this work we break the ties by adding a tiny amount of IID noise.

[16] Unlike a 1D cdf, the 2D counterpart lacks serial ordering, rendering applications, e.g., of the Kolmogorov-Smirnov test in two dimensions [34], difficult. However, ordering ambiguities are absent in rank-time space due to the row-column sum constancy of the probability mass function (pmf).

[17] To avoid cumbersome notation, we use $\mathcal{C}$ to denote either the ensemble cdf or the empirical one, the latter based on finite n including a single realization, depending on the context. Note that the ordinates at the jumps of the empirical cdf are ranks divided by N and, fundamentally, binning is not required.

[18] Improved single number metrics, based on five distinct symmetry classes of $\delta\mathcal{C}_{\rm rms}$, were used with some success for signal detection in [4].

[19] The row and column sum constancy is of broader relevance and turns up any time permutation matrices arise, extending to the "Birkhoff polytope," a set of $N \times N$ matrices formed from convex combinations of permutation matrices, that is, doubly stochastic matrices whose entries are non-negative real numbers, with rows and columns adding up to unity. In 1926 Van der Waerden conjectured that the minimum "matrix permanent" among all doubly stochastic matrices was that with all entries equal: our pmf in the ensemble limit of pure noise! A proof did not appear until 1980.

[20] I. Gikhman, A. Skorokhod, and M. Yadrenko, *Probability Theory and Mathematical Statistics* (Vyshcha Shkola, Kiev, 1988).

[21] J. D. Scargle, J. P. Norris, B. Jackson, and J. Chiang, Astrophys. J. **764**, 167 (2013).

[22] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing* (Cambridge University Press, New York, 2007).

[23] B. Suits, A. Kostinski, and M. Kulkarni, J. Magn. Reson., Ser. A **108**, 230 (1994).

[24] M. Kulkarni and A. Kostinski, IEEE Trans. Geosci. Remote Sens. **33**, 799 (1995).

[25] A. Jameson and A. Kostinski, J. Appl. Meteorol. Climatol. **41**, 83 (2002).

[26] A. Jameson and A. Kostinski, Bull. Am. Meteorol. Soc. **82**, 1169 (2001).

[27] The rows and columns sum to unity because, e.g., rank 1 must be attained by one or the other of the time slots, and the first time slot must be either rank 1 or 2.

[28] Deterministic chaos violates independence and, seemingly, even randomness. Nonetheless, owing to a positive Lyapunov exponent of $\lambda = \ln 2$ for $a = 4$, for j sufficiently large ($j \approx 55$ for double precision arithmetic), the joint pdf for (x_k, x_{k-j}) approaches a product form, as expected for a chaotic process.

[29] The pattern seen in Fig. 1(g) does not emerge until $N \geqslant 7$ but this is minor. Our main point is that the chaotic map provides an elementary demonstration of how the lack of independence causes departures from rank moment uniformity, despite being uncorrelated.

[30] R. Chicheportiche and J.-P. Bouchaud, Phys. Rev. E **86**, 041115 (2012).

[31] A. Papoulis, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, New York, 1984).

[32] Numerical evidence confirms the statistical independence of the modal coefficients, $\langle c_m c_l \rangle = 0$ in our case.

[33] In this connection, it is interesting to recall that the probability of a return for a 2D random walk is factorizable into two corresponding 1D probabilities, but this is not so in three dimensions, e.g., see p. 127 in [35].

[34] J. A. Peacock, Mon. Not. R. Astron. Soc. **202**, 615 (1983).

[35] P. G. Doyle and J. L. Snell, *Random Walks and Electric Networks*, Vol. 22 (American Mathematical Society, Providence, RI, 1984).