Behavior of the scaling correlation functions under severe subsampling

Sabrina Camargo,[1,2,*] Nahuel Zamponi,[3,†] Daniel A. Martin,[1,2] Tatyana Turova,[4]
Tomás S. Grigera,[2,5,6,7] Qian-Yuan Tang,[8] and Dante R. Chialvo[1,2,8]

[1]*Instituto de Ciencias Físicas (ICIFI-CONICET), Center for Complex Systems and Brain Sciences (CEMSC3),*
Escuela de Ciencia y Tecnología, Universidad Nacional de Gral. San Martín, Campus Miguelete, 25 de Mayo y Francia, 1650,
San Martín, Buenos Aires, Argentina

[2]*Consejo Nacional de Investigaciones Científcas y Técnicas (CONICET), Godoy Cruz 2290, 1425, Buenos Aires, Argentina*

[3]*Division of Hematology and Medical Oncology, Department of Medicine, Weill Cornell Medicine,*
1300 York Avenue, New York, New York 10065, USA

[4]*Mathematical Statistics, University of Lund, Box 118, 221 00 Lund, Sweden*

[5]*Instituto de Física de Líquidos y Sistemas Biológicos (IFLySiB), Universidad Nacional de La Plata,*
Calle 59 n 789, 1900 La Plata, Argentina

[6]*Departamento de Física, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, 1900 La Plata, Argentina*

[7]*Istituto dei Sistemi Complessi, Consiglio Nazionale delle Ricerche, via dei Taurini 19, 00185 Rome, Italy*

[8]*Department of Physics, Centre for Nonlinear Studies, Hong Kong Baptist University, Hong Kong SAR, China*

(Received 10 April 2025; accepted 10 June 2025; published 2 July 2025)

Scale invariance is a ubiquitous observation in the dynamics of large distributed complex systems. The computation of its scaling exponents, which provide clues on its origin, is often hampered by the limited available sampling data, making an appropriate mathematical description a challenge. This work investigates the behavior of correlation functions in fractal systems under conditions of severe subsampling. Analytical and numerical results reveal a striking robustness: the correlation functions continue to capture the expected scaling exponents despite substantial data reduction. This behavior is demonstrated numerically for the random 2D Cantor set and the Sierpinski gasket, both consistent with exact analytical predictions. Similar robustness is observed in 1D time series both synthetic and experimental, as well as in high resolution images of a neuronal structure. Overall, these findings are broadly relevant for the structural characterization of biological systems under realistic sampling constraints.

DOI: 10.1103/mdf8-6w38

I. INTRODUCTION

Scale invariance and long-range correlations are regarded as hallmarks of complexity in physical and biological systems [1–3]. In particular, the scaling behavior of correlation functions often signals proximity to a critical point, where large susceptibility and diverging correlation lengths dominate the system's dynamics [1,4]. Consequently, identifying and quantifying such scaling behavior has become a standard approach in the study of collective phenomena ranging from neural activity [5–10] to mitochondrial networks [11,12], flocking [13–15], and gene regulation [16].

While numerical models offer complete control over system size and sampling, they may neglect important real-world constraints and omit biologically relevant variables as modeling simplifications. In contrast, experimental data are inherently noisy, finite, and limited by technical constraints such as resolution, field-of-view, and imaging duration [8]. This leads to incomplete or coarse-grained observations of a system's structure or dynamics, raising the important question: To what extent are inferred scaling properties robust to subsampling [9]?

In real-world applications, observations are often restricted to a small fraction of the system. This spatial subsampling often cannot be compensated by longer measurements. It has been argued that such incomplete sampling may bias the estimation of scaling exponents or even produce apparent signatures of criticality as artifacts of the inference methodology [9,10].

Despite this, relatively few studies have addressed how correlation-based estimators behave under conditions of severe data reduction. In particular, while estimators such as box-counting dimension are regarded as sensitive to finite size and noise [17], correlation-based metrics like the radial distribution function or detrended fluctuation analysis (DFA) [18,19] have not been systematically assessed in this context. Understanding whether these quantities remain stable under subsampling is essential for both theoretical consistency and the practical analysis of experimental systems.

In this work, we investigate the effect of stochastic subsampling on correlation-based scaling measures across a range of synthetic and experimental datasets. We analytically show that estimators such as the radial distribution function are self-averaging, and numerically confirm their robustness under substantial subsampling in fractal sets, correlated time

[*] Contact author: scamargo@unsam.edu.ar
[†] Contact author: zamponi.n@gmail.com

series, and high-resolution neuronal images. In contrast, box-counting methods show higher variability and sensitivity to sampling fraction. These results provide practical guidance for researchers dealing with limited datasets and support the use of correlation-based estimators as robust tools for characterizing scale-invariant systems.

II. FRACTAL DIMENSION

The fractal dimension D_f of a set can be estimated by the box-counting algorithm which analyzes how the number of covering boxes scales with box size [2]. The steps involve covering the set with a grid of boxes of side length ε, and counting the number of boxes $N(\varepsilon)$ that contain part of the set. The process is repeated for a range of decreasing ε values and finally the slope of the linear regression line through the $\ln N(\varepsilon)$ versus $\ln(1/\varepsilon)$ points is computed. The fractal dimension is given by

$$D_f = \lim_{\varepsilon\to 0} \frac{\ln N(\varepsilon)}{\ln(1/\varepsilon)}. \tag{1}$$

In practice, D_f is approximated as the negative slope of the best-fit line in the log-log plot,

$$D_f \approx -\frac{\Delta \ln N(\varepsilon)}{\Delta \ln \varepsilon}, \tag{2}$$

for a reasonably small ε.

III. RADIAL DISTRIBUTION FUNCTION

The radial distribution function $g(r)$ is widely used to quantify spatial correlations in systems ranging from colloids [20] and granular media [21] to neuronal networks [5,22] and spatial tree patterns [23].

For homogeneous systems it can be written as

$$g^{(2)}(\mathbf{r}) = \frac{1}{\rho N} \sum_{i\neq j} \langle \delta[\mathbf{r} - (\mathbf{x}_i - \mathbf{x}_j)] \rangle, \tag{3}$$

where $\rho = N/V$ is the number density and $\mathbf{x_i}$ are the particles' positions. The radial distribution function is essentially the two-point density correlation function, normalized to unity at long distances,

$$g(r) = \frac{1}{\rho^2} \langle \rho(0)\rho(r) \rangle, \quad r > 0, \tag{4}$$

where $\rho(\mathbf{r}) = \sum_i \delta(\mathbf{r} - \mathbf{x}_i)$, and the equality is valid except at $r = 0$ where $g(r = 0) = 0$ and the two-point density correlation is singular.

IV. SUBSAMPLING

For each of the structures analyzed below, quantities are computed for the full sample, and then at several subsamplings defined by the probability s of including a given particle of the original structure in the subsample. The quantities are computed exactly in the same way for the original structure and all subsamplings (when involved, the sample size is clearly recomputed for the subsample, i.e., $N_s \approx sN$).

V. ANALYTICAL CONSIDERATIONS

Given an N-point snapshot of a system's configuration, the radial distribution function is estimated as (we consider periodic boundary conditions for simplicity)

$$\hat{g}(r_k) = \frac{1}{\rho N} \sum_{ij} \frac{\Delta[r_{ij} - (k + 1/2)\delta r]}{V_k}, \tag{5}$$

where δr is the bin width, $r_k = k\delta r$ is the position of the center of the kth bin and V_k its volume, and $\Delta[r]$ is the interval indicator function

$$\Delta[r] = \begin{cases} 1 & \text{if } -\delta r/2 < r \leqslant \delta r/2, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Defining the number of points in the kth bin centered on particle i, $N_i(k) = \sum_j \Delta[r_{ij} - (k + 1/2)\delta r]$, we can write the estimator as

$$\hat{g}(r_k) = \frac{1}{\rho N} \frac{1}{V_k} \sum_i N_i(k). \tag{7}$$

When the snapshot is subsampled, each point is included with probability p (here $p = s$). Let's call $N_i^{(p)}(k)$ the number of points in the kth bin for a subsampled configuration, given that particle i belongs to the subsample. The $N_i^{(p)}(k)$ are correlated random variables, but the marginal probability of a single of these quantities should be binomial if the subsampling is homogeneous and independent (i.e., if the probability of picking a particle is independent of its position and of whether its neighbors have been picked). Then

$$\left\langle N_i^{(p)}(k) \right\rangle_p = pN_i(k), \tag{8}$$

where the average is over all possible subsamplings with probability p. The estimate of the radial distribution function from a single subsample is

$$\hat{g}^{(p)}(r_k) = \frac{1}{V_k} \frac{1}{\rho^{(p)} N^{(p)}} {\sum_i}' N_i^{(p)}(k), \tag{9}$$

where the sum is primed to remind that the number of terms fluctuates with the subsampling.

Averaging over all possible subsamples,

$$\left\langle \hat{g}^{(p)}(r_k) \right\rangle \approx \frac{V}{V_k} \left\langle \frac{N_i^{(p)}(k)}{N^{(p)}} \right\rangle_p = \hat{g}(r_k), \tag{10}$$

where we have assumed homogeneity [as does Eq. (5)], and the $\approx$ sign is because it is not the single snapshot that is homogeneous, but the ensemble where it comes from.

The main point is that the single-subsample estimate Eq. (9) will actually be a good approximation to the full-sample estimate Eq. (5) as long as $N^{(p)}$ is not too small, because due to double homogeneity of the original sample and of the subsampling, the average over the subsamples can be approximately realized by a space average (average over focal particles), i.e., $\hat{g}(r_k)$ can be regarded as self-averaging over the ensemble of subsamples.

We can estimate the extent of the fluctuations of $\hat{g}^{(p)}(r_k)$,

$$\mathrm{Var}\Big[\big(\rho^{(p)}\big)^2\hat{g}^{(p)}(r_k)\Big] = \frac{1}{V^2V_k^2}\Bigg[\sum_i{}' \mathrm{Var}\big(N_i^{(p)}\big) + \sum_{i\neq j}{}' \big\langle\big(N_i^{(p)}-\big\langle N_i^{(p)}\big\rangle\big)\big(N_j^{(p)}-\big\langle N_j^{(p)}\big\rangle\big)\big\rangle\Bigg]. \tag{11}$$

Of the $N(N-1)$ terms contributing to the double sum, only a number of order N will be different from zero since $N_i^{(p)}$ and $N_j^{(p)}$ become decorrelated when the centers are far apart. Since $\mathrm{Var}[N_i^{(p)}(k)] = p(1-p)N_i(k)$, we have

$$\mathrm{Var}\big[\hat{g}^{(p)}(r_k)\big] \approx \frac{1-p}{Np}\frac{\hat{g}(r_k)}{V_k\rho} + O\Big(\frac{1}{N}\Big), \tag{12}$$

which confirms that $\hat{g}(r_k)$ is self-averaging for $V\to\infty$.

Notably, the estimator $\hat{g}(r_k)$ is computed by binning all pairwise distances into intervals of width δr, so it can be formally written as a convolution of the exact radial distribution function $g(r)$ with a rectangular kernel:

$$\hat{g}(r_k) = \int g(r)\,K_{\delta r}(r_k - r)\,dr, \tag{13}$$

where $K_{\delta r}(r) = \frac{1}{\delta r}\,\Delta(r)$ is the normalized box function centered at zero. In Fourier space, this becomes

$$\widehat{\hat{g}}(q) = \widehat{g}(q)\cdot \mathrm{sinc}(q\delta r/2), \tag{14}$$

where q denotes the spatial frequency (wavenumber) dual to r. This expression makes the low-pass filtering effect explicit: the sinc-shaped transfer function attenuates high-frequency components of $g(r)$, effectively imposing a spatial frequency cutoff near $1/\delta r$. As a result, the estimator suppresses small-scale fluctuations while preserving the low-frequency content that governs the scaling behavior.

Now we turn to describe numerical results of the effect of subsampling synthetic fractal structures using two well-known fractal sets where the fractal dimension is known analytically.

VI. THE SIERPINSKI GASKET

The Sierpinski gasket, or Sierpinski triangle, is a self-similar set with fractal dimension $D_f = \log_2 3$. It can be built in several ways; here we have used the so-called chaos game [24,25] where starting from a point that belongs to the set, one obtains another point by moving halfway toward a randomly selected vertex of the equilateral triangle containing the set. Specifically, taking an equilateral triangle defined by the vertices $\mathbf{v}_1 = (0,0)$, $\mathbf{v}_2 = (1,0)$, $\mathbf{v}_3 = (1/2, \sqrt{3}/2)$, and setting $\mathbf{p}_1 = \mathbf{v}_1$, successive points belonging to the fractal are obtained by the random sequence

$$\mathbf{p}_{n+1} = \tfrac{1}{2}(\mathbf{p}_n + \mathbf{v}_r), \tag{15}$$

where r is a random integer between 1 and 3.

The results for the Sierpinski gasket are shown in Fig 1 depicting the $g(r)$ and the initial decay exponent for the full sample of the fractal and for several subsamplings. Notice that the different curves in Fig. 1 are almost indistinguishable from each other. The theoretical value of η is $\eta = D_f - 2 = \log 3/\log 2 - 2 \approx -0.415$, a numerical estimate of D_f which is within 0.5% of the theoretical value, and very stable against (even severe) subsampling. In passing, we note that the correlation estimation based on box counting are less stable than $g(r)$, an observation also made with other sets commented later on.

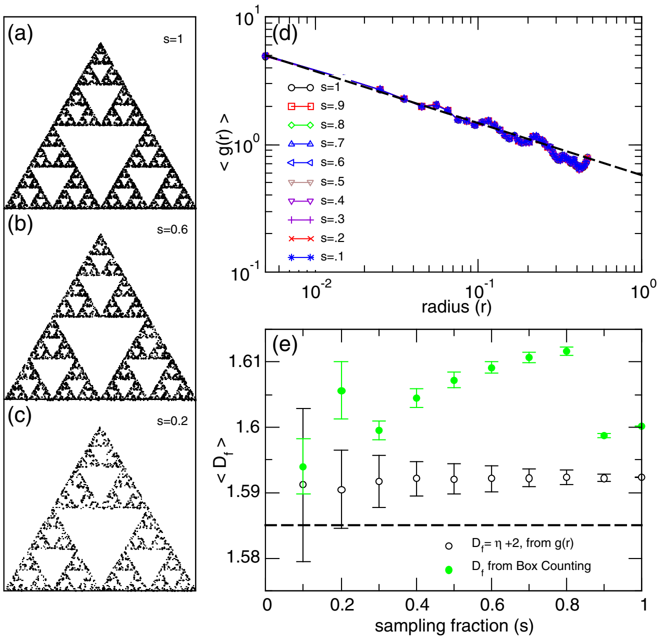

FIG. 1. Subsampling of the Sierpinski gasket does not severely affect the numerical estimation of the scaling exponents. A set of 10 000 points belonging to the Sierpinski triangle was created by iterating the chaos game rule. Typical sets for full, 0.6, and 0.2 sampling fractions are shown in panels (a)–(c), respectively. The results in panel (d) show the radial distribution functions $g(r)$ computed for the sampling fractions depicted in the legend, from which the initial power-law decay of $g(r)\sim r^{\eta}$ was computed by a log-linear fit of the estimated $g(r)$. The dashed line corresponds to a power law with the exact exponent $\eta = D_f - 2 = -0.415$ and $D_f = \log(3)/\log(2)$. Panel (e) illustrates the scaling exponents vs sampling ratio derived both from the $g(r)$ function as well as from the box counting method. Error bars correspond to mean +/− sd values computed from 100 realizations. The dashed line corresponds to the exact $D_f = \eta + 2 = 1.585$ value.

VII. THE 2D RANDOM CANTOR SET

We also consider the random Cantor set [24]. Its stochastic character entails the concept of the almost sure Hausdorff dimension, meaning it holds with probability 1 for a random realization of the set. The random Cantor set in 2D is constructed through an iterative process that generalizes the 1D random Cantor set to two dimensions. Starting with a unit square (or any initial square) in 2D space, the construction process involves the following steps. (i) Division step: Divide the square into smaller subsquares. The number of subsquares depends on the scaling factor r. For example, if $r = \frac{1}{2}$, the square is divided into $2\times 2 = 4$ smaller squares. (ii) Random removal step: Each subsquare is removed with probability p_C or kept with probability $1 - p_C$. In this case,

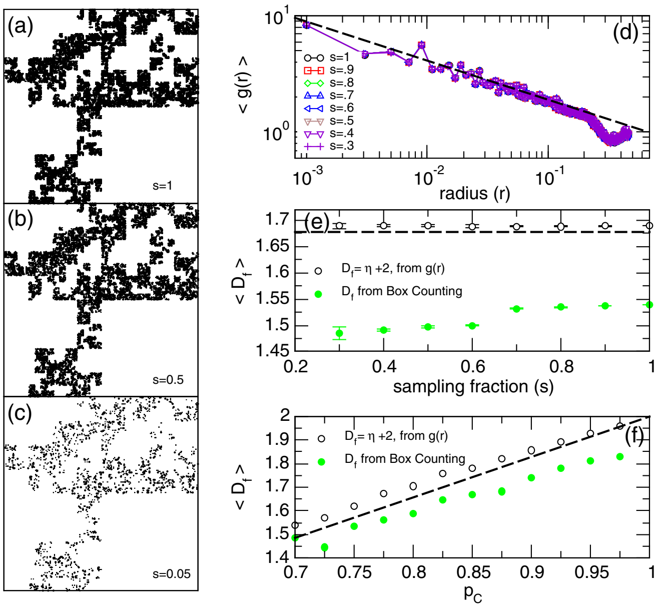

FIG. 2. Numerical estimation of the scaling exponents of the 2D random Cantor set under a wide range of subsampling. Examples of the Cantor sets (for $p_C = 0.8$) are shown in panel (a) (fully sampled) as well as two subsampled versions (s = 0.5 and 0.05) in panels (b) and (c), respectively. Panel (d): Radial distribution functions $g(r)$ for $p_C = 0.8$ at different subsamplings. Panel (e): Fractal dimension estimated from the $g(r)$ and from the box counting computations of the 2D random Cantor set with $p_C = 0.8$ at different sampling fractions. Panel (f): Same estimations as in panel (e) for the fully sampled 2D random Cantor set generated with different values of p_C. Statistics [means +/− standard deviation in panels (d) and (e), means in panel (f)] computed from ten independent realizations; dashed lines in panels (d)–(f) correspond to the analytical scaling values.

$p_C = 0.8$, so each subsquare has an 80% chance of being removed and a 20% chance of being kept. (iii) Iteration: Repeat the process for each remaining subsquare from the previous step. At each iteration, the remaining squares are further divided, and subsquares are randomly removed or kept, continuing until a desired level of detail is reached. The expected fractal dimension of the random Cantor set in 2D is $D_f = 2 + \log(p_C)/\log(2)$.

Figure 2 shows the numerical estimation of the scaling exponents for the 2D random Cantor set ($p_C = 0.8$ and $r = \frac{1}{2}$, where $D_f \approx 1.678$) under a wide range of subsampling. Panels (a)–(c) show the fully sampled Cantor set ($s = 1$) and two subsampled versions ($s = 0.5$ and $s = 0.05$), respectively. The radial distribution functions $g(r)$ in panel (d) and the estimated fractal dimensions in panel (e) (for $p_C = 0.8$) exhibit the same robust behavior observed for the Sierpinski gasket: the curves remain nearly indistinguishable across sampling levels, and the estimates based on $g(r)$ are notably more stable than those from box counting.

VIII. 1D TIME SERIES' CORRELATIONS BEHAVIOR UNDER SUBSAMPLING

The correlations scale invariance discussed in the above sections can be defined as well for the case of 1D time series which is expected to show similar behavior under subsampling. This can be demonstrated by computing the scaling of the fluctuations inside segments of increasing length as implemented by the detrended fluctuation analysis (DFA) [18]. The DFA method is commonly used to determine the statistical self-affinity of a time series, which may exhibit long range correlations. The DFA scaling exponent (commonly denoted as α) equals the Hurst exponent H [26] in the case of stationary processes, but unlike traditional methods, DFA can also assess scaling in nonstationary processes. The relationships among the relevant scaling exponents—the autocorrelation decay exponent γ, the power spectral exponent β, H, and α—can be derived from the Wiener-Khinchin theorem [19,27,28]: $\gamma = 2 - 2\alpha$; $\beta = 2\alpha - 1$; $\gamma = 1 - \beta$; $\alpha = (\beta + 1)/2$; and $H = \alpha$ only for fractional Gaussian noise (i.e., for $-1 \leqslant \beta \leqslant 1$).

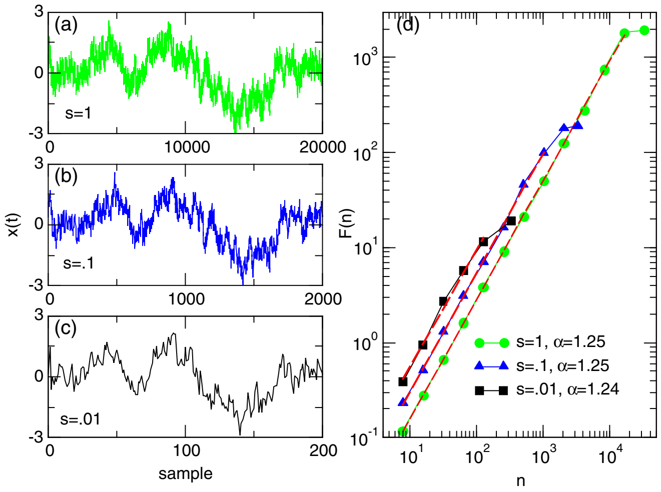

FIG. 3. Time series long range correlations behavior under random subsampling. Panel (a) illustrates a typical example of a Gaussian correlated time series with $1/f^\beta$ spectral decay ($\beta = 1.5$, $\sigma^2 = 1$). Panels (b) and (c) depict the subsampled versions at $s = 0.1$ and s $= 0.01$, respectively. Panel (d) shows the detrended fluctuation analysis of each time series, where $F(n)$ is the total fluctuation computed from segments of length n, and a behavior $F \propto n^\alpha$ is expected. The computed value of the α scaling exponent remains very close to the predicted dependency on the spectral β, i.e., $\alpha = (\beta + 1)/2 = 1.25$, even for the case of extreme subsampling. Note that the subsampling seems to remove high frequency fluctuations, thus limiting the scaling regions to smaller n values [calculation done with a fully sampled ($s = 1$) time series of $N = 2^{15}$].

The relative persistence of the long range correlations under subsampling can be readily demonstrated for 1D time series. This is evident already by simple visual inspection of the time series, as shown in panels (b) and (c) of Fig. 3 where the overall shape of the signal is preserved, even for a sampling rate hundreds of times smaller.

IX. NEURONAL STRUCTURE

The correlation behavior under subsampling is now briefly explored for the scaling of neuronal structures. The correlation analysis of this type of data can be computationally demanding, especially for high resolution images which implies the

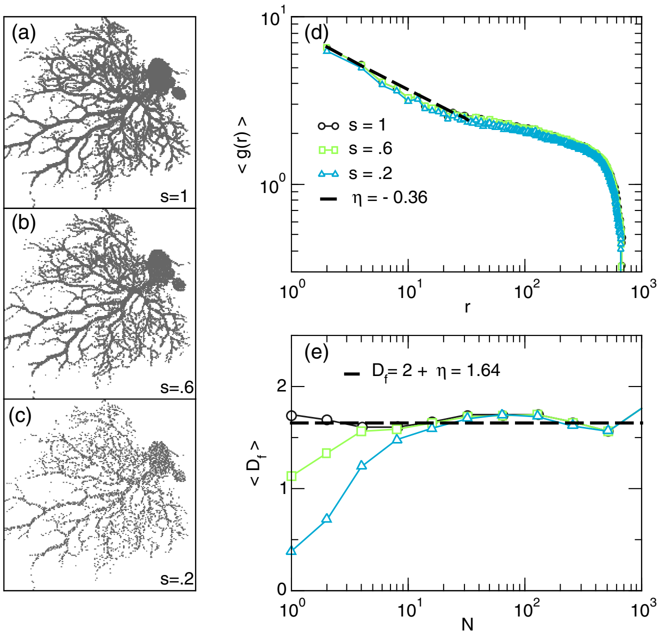

FIG. 4. Maximum intensity projection of a set of optical sections acquired with a multiphoton microscope of a Purkinje neuron from the mouse cerebellar cortex injected with Lucifer yellow fluorescent dye. Panels on the left shown the fully sampled binarized image [labeled (a)], as well as for two different sampling fractions [s = 0.6 in panel (b) and s = 0.2 in panel (c)]. Panel (d) shows the radial distribution function $g(r)$ computed for a range of sampling fractions s, and panel (e) the local fractal dimension for increasing sampling fractions of the raw binarized image data. Data freely available from [29].

calculation of products of several million pixels. Therefore it is relevant to demonstrate that similar correlation results can be obtained at subsampled images. The results of the analysis is presented in Fig. 4. The fluorescence images were binarized following the methods in [11,12] and processed in the same manner as in the synthetic fractal discussed already. It can be seen that η, the exponent estimated from the initial decay of $g(r)$, is not severely affected by the subsampling while agreeing with the expected value for D_f. In passing, note that the obtained values are consistent with earlier estimations of D_f for this type of cerebellum neuron reported in Ref. [31]. Box counting here again shows greater sensitivity to subsampling, especially at small box sizes where data sparsity has a stronger effect.

X. LONG TERM CORRELATIONS IN FLY'S MOTION BEHAVIOR

Recent studies monitored the movements of fruit fly *Drosophila melanogaster* in an effort to understand how universally conserved phenomenon is sleep among the animal kingdom [30]. Using machine learning-based video-tracking technology, they conducted a detailed high-throughput analysis of sleep. To quantify walking dynamics, flies were introduced on a thin glass tube and loaded into an ethoscope, which is a self-contained machine able to record the activity of flies in real time using computerized video tracking with a resolution of 1920 × 1080 pixels, at 30 frames per second (see [32] for additional details).

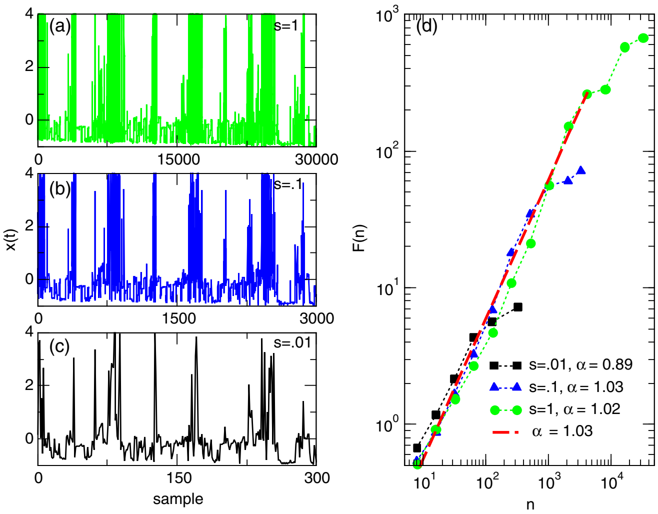

FIG. 5. Analysis of longterm records of flies positional tracings. Timeseries shown refers to the position of a single *Drosophila melanogaster* inside a thin (3 mm) glass tube (70 mm length) registered every 10 seconds for more than 30 days. Panel (a) shows the raw fully sample data. Panels (b) and (c) depict the subsampled versions at $s = 0.1$, and s = 0.01, respectively, and panel (d) shows the detrended fluctuation analysis of each time series. Data from Ref. [30] kindly provided by the authors.

In Fig. 5 we analyze one of the data sets presented in Ref. [30]. DFA results show similar behavior to Fig. 3: the estimation of characteristic exponent α remains robust even under subsampling by a factor of 100.

XI. DISCUSSION

Work closely related to the present study merits discussion. Reissa *et al.* [17] systematically investigated noise-induced biases in the estimation of fractal dimensions from 2D images, reporting a ∼20% inflation in the computed values depending on the algorithm used and the underlying spatial structure. These findings underscore the sensitivity of local geometric estimators to noise and perturbations—an issue less pronounced in correlation-based methods. Another relevant line of inquiry involves the estimation of critical exponents from avalanche statistics. In systems such as the brain, where only a subset of nodes is typically observable, power-law exponents derived from avalanche size distributions are known to be biased by subsampling [6–10]. Building on the work of Kuntz and Sethna [33], who showed that the avalanche exponent γ corresponds to the spectral exponent β in branching processes, Conte and de Candia [34] addressed this bias by demonstrating that both power spectral and DFA-based exponents remain stable under subsampling. Their results are in agreement with ours and reinforce the idea that estimators based on long-range correlations naturally suppress sampling-related noise.

In this study, we systematically examined the impact of uniform stochastic subsampling on the estimation of correlation-based scaling exponents across a diverse set of scale-invariant systems, including 2D spatial patterns, 1D time series (both synthetic and experimental), and biological image data. In all cases, we found that the estimated exponents—whether obtained from the initial decay of the radial distribution function $g(r)$ or via DFA—remained remarkably stable even under substantial data reduction. This robustness arises because both the subsampling procedure and the estimators themselves act to suppress short-range fluctuations, thereby preserving the large-scale structures that underpin scaling behavior.

A central challenge in many biological experiments is the limited sampling imposed by practical, technical, or ethical constraints. High-resolution imaging of cellular structures, large-scale neural recordings, and long-term behavioral monitoring typically capture only a small subset of the full system, either due to limited spatial coverage, finite recording durations, or restricted access to internal variables. These limitations raise important concerns about the reliability of inferred structural and dynamical properties, particularly when attempting to characterize scale-invariant features. Our results show that, under the assumption of statistical homogeneity, correlation-based measures such as $g(r)$ and DFA remain reliable even in the presence of significant data loss. This robustness establishes a principled foundation for analyzing complex biological systems when only partial observations are available.

These findings offer practical guidance for empirical studies in which full data acquisition is infeasible. In domains such as high-resolution microscopy, large-scale neural recordings, and long-term behavioral tracking, subsampling is often unavoidable. Nevertheless, our results demonstrate that robust inference of scale-dependent properties remains attainable in such settings, provided the system exhibits statistical homogeneity and appropriate estimators are used. This opens the door to rigorous data reduction strategies that do not compromise the integrity of the system's essential features.

At a more fundamental level, the observed robustness can be traced to a physical mechanism: both stochastic subsampling and correlation-based estimators effectively act as low-pass filters. By attenuating high-frequency fluctuations, they preserve the long-range structure responsible for scaling behavior. This phenomenon is analogous to the aliasing effect in signal processing [35], where undersampling distorts high-frequency components unless prefiltered. In our case, the intrinsic smoothing properties of correlation estimators mitigate such distortions, enabling reliable recovery of macroscopic features. This observation resonates with the renormalization group theory of critical phenomena, where coarse-graining preserves the key descriptors of the system despite microscopic variability. A related perspective is offered by compressed sensing [36], which shows that global structure can be reconstructed from sparse observations when guided by suitable priors, albeit through different mechanisms.

A key caveat to our findings is the assumption of statistical homogeneity. Systems exhibiting pronounced spatial or temporal inhomogeneities may violate the conditions under which our conclusions hold. While our focus here has been on scale-invariant correlations—due to their broad relevance across physical and biological systems—similar robustness may also apply to systems with short-range correlations, provided the correlation length exceeds the sampling resolution and the system remains homogeneous. More generally, the preservation of macroscopic statistical descriptors under subsampling is expected to depend on the interplay between correlation scale and sampling resolution.

In summary, our results provide a solid framework for structure-preserving subsampling in the analysis of complex systems. They demonstrate that scaling exponents, rather than being fragile or overly sensitive to data loss, can serve as resilient markers of collective organization. As datasets grow in complexity and scale, the ability to perform principled data reduction without compromising scientific insight will become increasingly important.

ACKNOWLEDGMENTS

D.R.C. thanks the support of Hong Kong Baptist University through the Distinguished Professorship of Science awarded during the course of these studies.

DATA AVAILABILITY

The data that support the findings of this article are openly available [37].

[1] H. E. Stanley, Scaling, universality, and renormalization: Three pillars of modern critical phenomena, Rev. Mod. Phys. **71**, S358 (1999).

[2] T. Vicsek, *Fractal Growth Phenomena*, 2nd ed. (World Scientific, Singapore, 1992).

[3] T. Mora and W. Bialek, Are biological systems poised at criticality? J. Stat. Phys. **144**, 268 (2011).

[4] A. Attanasi, A. Cavagna, L. Del Castello, I. Giardina, S. Melillo, L. Parisi, O. Pohl, B. Rossaro, E. Shen, E. Silvestri, and M. Viale, Finite-size scaling as a way to probe near-criticality in natural swarms, Phys. Rev. Lett. **113**, 238102 (2014).

[5] S. Camargo, D. A. Martin, E. J. Aguilar Trejo, A. de Florian, M. A. Nowak, S. A. Cannas, T. S. Grigera and D. R. Chialvo, Scale-free correlations in the dynamics of a small (N∼10000) cortical network, Phys. Rev. E **108**, 034302 (2023).

[6] T. L. Ribeiro, M. Copelli, F. Caixeta, H. Belchior, D. R. Chialvo, M. A. L. Nicolelis, and S. Ribeiro, Spike avalanches exhibit universal dynamics across the sleep-wake cycle, PLoS ONE **5**, e14129 (2010).

[7] T. L. Ribeiro, S. Ribeiro, H. Belchior, F. Caixeta, and M. Copelli, Undersampled critical branching processes on small-world and random networks fail to reproduce the statistics of spike avalanches, PLoS ONE **9**, e94992 (2014).

[8] V. Priesemann, M. H. J. Munk, M. Wibral, Subsampling effects in neuronal avalanche distributions recorded in vivo, BMC Neurosci. **10**, 40 (2009).

[9] A. Levina and V. Priesemann, Subsampling scaling, Nat. Commun. **8**, 15140 (2017).

[10] T. T. A. Carvalho, A. J. Fontenele, M. Girardi-Schappo, T. Feliciano, L. A. A. Aguiar, T. P. L. Silva, N. A. P. de Vasconcelos, P. V. Carelli, M. Copelli, Subsampled directed percolation models explain scaling relations experimentally observed in the brain, Front. Neural Circuits **14**, 576727 (2021).

[11] N. Zamponi, E. Zamponi, S. A. Cannas, O. V. Billoni, P. R. Helguera, D. R. Chialvo, Mitochondrial network complexity emerges from fission/fusion dynamics, Sci. Rep. **8**, 363 (2018).

[12] N. Zamponi, E. Zamponi, and S. A. Cannas, D. R. Chialvo, Universal dynamics of mitochondrial networks: A finite-size scaling analysis, Sci. Rep. **12**, 17074 (2022).

[13] A. Cavagna, A. Cimarelli, I. Giardina, G. Parisi, R. Santagati, F. Stefanini, and M. Viale, Scale-free correlations in starling flocks, Proc. Natl. Acad. Sci. USA **107**, 11865 (2010).

[14] A. Cavagna, D. Conti, C. Creato, L. Del Castello, I. Giardina, T. S. Grigera, S. Melillo, L. Parisi, and M. Viale, Dynamic scaling in natural swarms, Nat. Phys. **13**, 914 (2017).

[15] J. Múgica, J. Torrents, J. Cristín, A. Puy, M. C. Miguel, and R. Pastor-Santorras, Scale-free behavioral cascades and effective leadership in schooling fish, Sci. Rep. **12**, 10783 (2022).

[16] D. B. Brückner, H. Chen, L. Barinov, B. Zoller, and T. Gregor, Stochastic motion and transcriptional dynamics of pairs of distal DNA loci on a compacted chromosome, Science **380**, 1357 (2023).

[17] M. A. Reissa, N. Sabathielb, H. Ahammer, Noise dependency of algorithms for calculating fractal dimensions in digital images, Chaos, Solitons Fractals **78**, 39 (2015).

[18] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Mosaic organization of DNA nucleotides, Phys. Rev. E **49**, 1685 (1994).

[19] C. Heneghan and G. McDarby, Establishing the relation between detrended fluctuation analysis and power spectral density analysis for stochastic processes, Phys. Rev. E **62**, 6103 (2000).

[20] E. R. Weeks, J. C. Crocker, A. C. Levitt, A. Schofield, and D. A. Weitz, Science **287**, 627 (2000).

[21] R. Al-Raoush, Physica A **377**, 545 (2007).

[22] T. S. Grigera, Correlation functions as a tool to study collective behaviour phenomena in biological systems, J. Phys.: Complex. **2**, 045016 (2021).

[23] P. Villegas, A. Cavagna, M. Cencini, H. Fort, and T. S. Grigera, Joint assessment of density correlations and fluctuations for analysing spatial tree patterns, R. Soc. Open Sci. **8** 202200 (2021).

[24] M. Barnsley, *Fractal Everywhere* (Academic Press, San Diego, USA, 1988).

[25] M. F. Barnsley and A. Vince, The chaos game on a general iterated function system, Ergod. Theory Dyn. Syst. **31**, 1073 (2011).

[26] H. E. Hurst, Long term storage capacity of reservoirs, Trans. Am. Soc. Eng. **116**, 770 (1951).

[27] N. Wiener, Generalized harmonic analysis, Acta Math. **55**, 117 (1930).

[28] A. Khintchine, Korrelationstheorie der stationären stochastischen Prozesse, Math. Ann. **109**, 604 (1934).

[29] The Cell Image Library https://www.cellimagelibrary.org/images/CCDB_3687.

[30] Q. Geissmann, E. J. Beckwith, G. F. Gilestro, Most sleep does not serve a vital function: Evidence from Drosophila melanogaster, Sci. Adv. **5**, eaau9253 (2019).

[31] B. R. Krauss, B. J. Serog, D. R. Chialvo, and A. V. Apkarian, Dendritic complexity and the evolution of cerebellar Purkinje cells, Fractals **02**, 95 (1994).

[32] Q. Geissmann, L. Garcia Rodriguez, E. J. Beckwith, A. S. French, A. R. Jamasb, and G. F. Gilestro, Ethoscopes: An open platform for high-throughput ethomics, PLoS Biol. **15**, e2003026 (2017).

[33] M. C. Kuntz and J. P. Sethna, Noise in disordered systems: The power spectrum and dynamic exponents in avalanche models, Phys. Rev. B **62**, 11699 (2000).

[34] D. Conte and A. de Candia, Inferring global exponents in subsampled neural systems, bioRxiv, doi: https://doi.org/10.1101/2024.11.29.626005.

[35] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, *Signals and Systems*, 2nd ed. (Prentice Hall, Upper Saddle River, NJ, 1999).

[36] D. L. Donoho, IEEE Trans. Inf. Theory **52**, 1289 (2006).

[37] https://github.com/DanielAlejandroMartin/Subsampling.