Constrained surrogates for arbitrary families of continuous probability distributions

Jack Murdoch Moore [1,2] and Eduardo G. Altmann [3,*]

[1] *MOE Key Laboratory of Advanced Micro-Structured Materials, and School of Physical Science and Engineering, Tongji University, Shanghai 200092, People's Republic of China*

[2] *National Key Laboratory of Autonomous Intelligent Unmanned Systems, MOE Frontiers Science Center for Intelligent Autonomous Systems, Tongji University, Shanghai 200092, People's Republic of China*

[3] *School of Mathematics and Statistics, University of Sydney, New South Wales 2006, Australia*

(Received 5 February 2026; accepted 29 September 2026; published 28 October 2026)

Empirical regularities described by parametrized families of probability distributions can shape how complex systems are interpreted and explained. A recurring problem is to determine the extent to which these descriptions are supported by observations, and how they can be used to characterize the underlying system. Here we introduce a general framework for constructing constrained surrogates capable of representing arbitrary families of continuous distributions. We demonstrate its utility by providing explicit algorithms for several canonical distributions in statistical physics, including continuous power-law, lognormal, exponential, and truncated Gaussian distributions. These constrained surrogates enable unbiased estimation and accurate tests of model validity based on arbitrary (physically meaningful) sample statistics, without requiring specification or estimation of model parameters. We show how these theoretical advantages translate into more informative and nuanced conclusions in applications ranging from bushfire magnitudes and word usage in texts to brain connectomes and city populations.

DOI: 10.1103/fs71-p5y8

I. INTRODUCTION

Computation enables exploration of complex systems that are inaccessible to purely analytic approaches or direct experimentation. For this reason, it has come to stand alongside theory and experiment as a third pillar of science [1,2]. A computational strategy with close ties to both theory and experiment is the method of surrogates, which generates synthetic datasets, known as surrogates, by imposing a theoretical hypothesis or model on empirical data [3–5]. Surrogates can be used to estimate properties of observed systems [6,7], test model validity [8,9], and compare candidate models [10,11]. As a result, surrogate methods are applied across scientific domains, including neuroscience [12], climate science [13], epidemiology [14], and artificial intelligence [15].

Both the theoretical support and the performance of surrogate methods depend on how they are constructed. Surrogates methods relying on a fully specified generative model risk bias when the assumed model is misspecified [4,5,16]. Constrained surrogates address this limitation by avoiding explicit specification of a single generative process and instead constraining quantities observed in the data whose preservation allows representation of the relevant class of models in generality [17–27]. By reducing sensitivity to model misspecification, constrained surrogates can—especially at low sample length—provide higher accuracy and power in significance tests, avoid biasing estimates of arbitrary quantities, and enable the principled use of arbitrary (test) statistics [4,16].

A central question in statistical physics and complex systems is whether observations are consistent with a given parametric probability distribution $p(x)$, and how such distributions can be used to characterize and model the underlying system [28–33]. In particular, fat-tailed distributions are often taken as evidence for underlying critical behavior [34–40]. While constrained surrogates have been developed for discrete power-law models [16], many other parametric distributions of comparable importance lack constrained surrogates which would allow theoretically supported assessment and application to empirical data. These include exponential [41,42], Gaussian [43,44], and lognormal [45,46] distributions, which arise from distinct generative mechanisms and are widely used to model diverse empirical systems.

In this paper, we introduce an approach for generating constrained surrogates representing arbitrary families of continuous distributions. Applying this approach, we obtain constrained-surrogate algorithms for the most central distributions in complex systems and statistical physics: continuous power-law, lognormal, exponential, and truncated Gaussian distributions. We then show how the theoretical license to employ arbitrary statistics can enhance both hypothesis testing and statistical estimation in practice. We illustrate these advantages across several empirical systems. In particular, for data on the magnitude of Australian

*Contact author: eduardo.altmann@sydney.edu.au

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

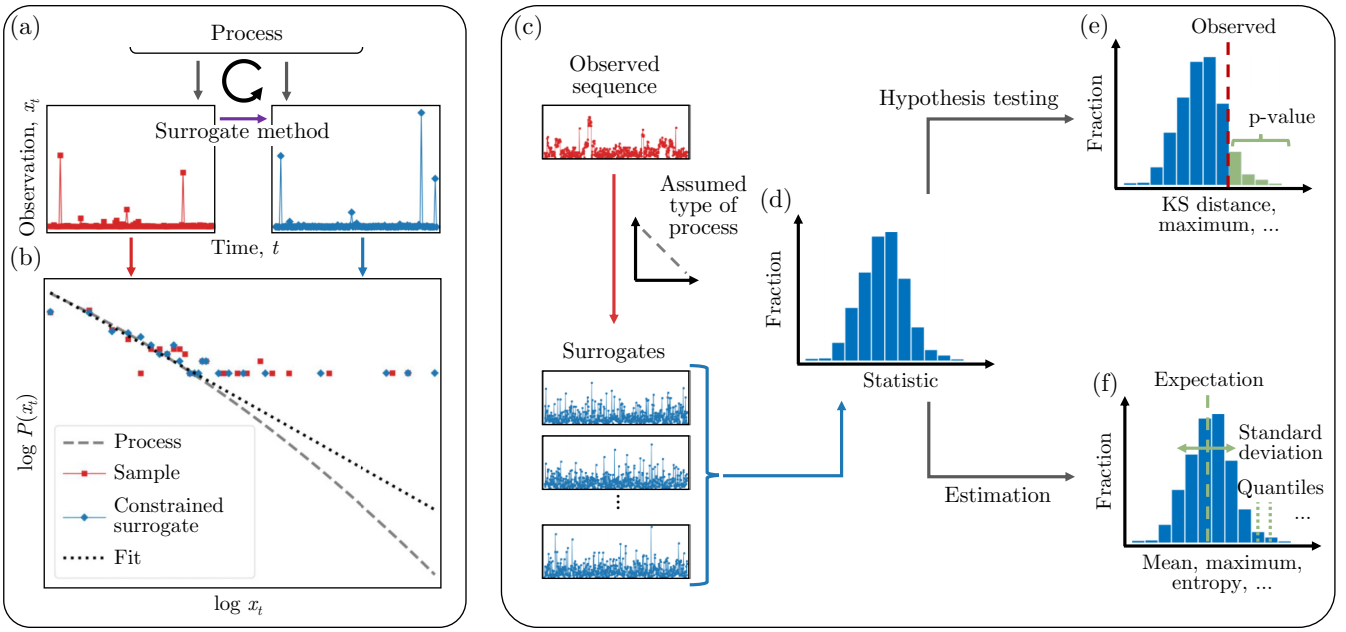


FIG. 1. Surrogates in theory and practice. (a), (b) Theory. (a) Constrained surrogates provide commutativity (circular arrow) in the sense that generating from a probabilistic process and then applying a constrained surrogate algorithm which properly represents that process is equivalent to generating directly from the original process: the two procedures yield the same marginal distribution. (b) A constrained surrogate preserves the likelihood and hence the maximum likelihood fit of an input sequence. Ideally, it is chosen uniformly at random from among all sequences which exhibit this likelihood function (and maximum likelihood fit). (c)–(e) Practice. (c) An assumed type of process (e.g., power-law, lognormal, or exponential) applied to an observed sequence gives rise to an ensemble of surrogates. (d) A statistic of interest is calculated separately from each generated surrogate sequence. (e) Hypothesis testing is performed by comparing the value of the statistic calculated from the original sequence with the distribution estimated from surrogates and can lead to a rejection of the assumed process. (f) The distribution of statistics under surrogates is used to estimate the properties of the observed system, assuming it follows the hypothesized type of process.

bushfires, constrained surrogates allow us to distinguish agreement in the bulk of the distribution from systematic overestimation of the probability of extreme events. Applications to word usage in texts and brain connectomes further illustrate how constrained surrogates enable tests based on physically meaningful quantities that reveal both the presence and the nature of discrepancies invisible to standard goodness-of-fit tests. Across all cases, these improvements are achieved without specifying or restricting model parameters.

II. TYPICAL AND CONSTRAINED SURROGATES

A surrogate is a version of an observed dataset that was artificially generated under a hypothesized model. While the process underlying empirical observations is often *a priori* unknown, properly generated surrogates represent a hypothesis or model unambiguously, and this concreteness is key to their utility. Surrogates inform choice of model because they allow the observed dataset to be compared with datasets known to have arisen under a hypothesis. Once a model has been chosen, surrogates allow estimation of properties of the observed system based on the observed dataset, sometimes without the need to estimate parameters. We illustrate the role of surrogates in scientific research in Fig. 1.

The *typical* approach for obtaining surrogates involves estimating the parameters ξ of the hypothesized model from the observed time series, e.g., through maximum likelihood, and then generating new time series as realizations of the model with fitted parameters $\xi = \hat{\xi}$. A prominent example of this approach is the test of power-law distribution, in which the parameter corresponds to the power-law exponent $\xi = \gamma$ and the model is sampling independent and identically distributed (i.i.d.) from the distribution with the estimated exponent [47]. This method is convenient and intuitive, but has the limitation that it does not represent the hypothesized process in general (e.g., a power-law process with unknown exponent γ) but only the specific process with the parameters of best fit (e.g., power-law process with exponent $\gamma = \hat{\gamma}$). The ubiquity of correlations in real-world complex systems and the difficulty of accurately estimating parameters from limited samples can limit the accuracy and power of hypothesis tests based on typical surrogates and lead to bias in statistical estimates. The popular bootstrapping method (sampling with replacement) can be seen as an example of a typical surrogate method, built on the hypothesis that observations arose i.i.d. under some unknown distribution $p(x)$. Bootstrapping treats the empirical distribution of the observed data, the maximum likelihood estimate of $p(x)$, as the true distribution and generates surrogates by resampling from it.

The *constrained* surrogate approach was originally described as an artificial dataset which exhibits the same parameters of best fit under the considered model as the observed data from which it arose [4]. This definition has been refined to a dataset generated from the hypothesized

process while conditioning on a sufficient statistic, which renders irrelevant the parameters of the underlying process [48]. We will employ a restriction of this definition which in many cases of interest is equivalent. Denoting by $\mathcal{L}_{\mathbf{x}}$ the likelihood of a sequence $\mathbf{x}$, a constrained surrogate for an input sequence $\mathbf{x}^{\text{input}}$ is chosen uniformly at random from a set of sequences $\mathcal{S}(\mathbf{x}^{\text{input}})$ satisfying the following three conditions [16]:

(1) Each sequence in the set has the same likelihood under the hypothesized process, i.e., $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{S}(\mathbf{x}^{\text{input}})$, $\mathcal{L}_{\mathbf{x}} = \mathcal{L}_{\mathbf{x}'}$

(2) The set contains the observed sequence, i.e., $\mathbf{x}^{\text{input}} \in \mathcal{S}(\mathbf{x}^{\text{input}})$ and

(3) The same set would be derived from any sequence which it contains, i.e., $\forall \mathbf{x} \in \mathcal{S}(\mathbf{x}^{\text{input}})$, $\mathcal{S}(\mathbf{x}) = \mathcal{S}(\mathbf{x}^{\text{input}})$.

These conditions ensure that samples generated under the hypothesized process and their constrained surrogates have the same marginal distribution [see Fig. 1(a), circular arrow].

In addition to being parameter agnostic, other statistical advantages of constrained surrogates include unbiased estimation of *any* test statistic, reduced variance, and exact hypothesis tests for arbitrary test statistics or sample lengths [4,16,49,50].

Here we are particularly interested in constrained surrogates representing the hypothesis that observations arose (i.i.d.) under some family of distributions $\{p_\xi(x)\}$. An approach allowing us to remain completely agnostic to the family of distributions would be to generate samples without replacement, i.e., shuffling. The above definition of a constrained surrogate is satisfied in this illustrative case because (0) the sequence is chosen uniformly at random from among all sequences reachable via shuffling; (1) the likelihood of a sequence of i.i.d. observations does not depend on the order in which its elements appear; (2) the originally observed sequence can be generated by shuffling; and (3) the set of sequences reachable by randomizing the order of a previously shuffled sequence is the same as the set reachable from the sequence originally observed. Shuffling avoids assumptions about $p(x)$, which entails strong restrictions on realizable sequences. Specifically, shuffling preserves all order-insensitive properties of the original time series, including, e.g., a sample's minimum, maximum, median, quantiles, mean, and variance. Therefore, while shuffling is useful to test the i.i.d. hypothesis, it is of limited interest in other applications. In particular: as shuffling does not generate any sample x_i that has not been previously observed, it cannot be used to estimate the probability of unobserved future events; and because shuffling simultaneously represents all families of distributions, it cannot be used to reach conclusions specific to a particular distributional family. In the next section, we introduce alternative constrained surrogate approaches that utilize assumptions about the functional form of $p(x)$ (though not its parameters) to overcome these limitations of shuffling as well as the bias and additional parametric assumptions of the typical surrogate approach. Thereby we achieve a balance between (1) shuffling, which avoids bias and assumptions about distribution or parameters, but restricts realizable sequences such that no new observation can be generated, and (2) typical surrogates, which can provide a nonzero chance of realizing any sequence, but restrict both distribution and parameters, and incur bias.

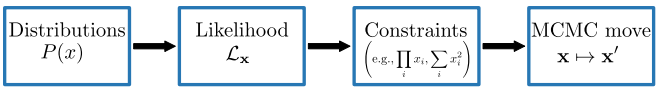


FIG. 2. The steps involved in the generation of constrained surrogates.

III. GENERATING THE SURROGATES

A. General approach

Goal. Given an input sequence $\mathbf{x}^{\text{input}}$, our goal is to generate surrogate sequences $\mathbf{x}$ that preserve the likelihood $\mathcal{L}_{\mathbf{x}}(\xi)$ that the dataset was generated independently from a distribution within a family $\{p_\xi(x)\}$ spanning all parameter values ξ. More precisely, we wish to sample with uniform probability all sequences $\mathbf{x} = x_1, x_2, \ldots, x_N$ satisfying

$$\mathcal{L}_{\mathbf{x}}(\xi) := \prod_{i=0}^{N} p_\xi(x_i) = \mathcal{L}_{\mathbf{x}}^{\text{input}}(\xi), \tag{1}$$

independently of ξ.

Approach. In Fig. 2 we illustrate our general idea to achieve the goal stated above. First, we identify distribution-specific constraints that preserve the likelihood in Eq. (1). Then we design a Markov chain Monte Carlo (MCMC) approach enforcing these constraints via simple local moves. More specifically, at each iteration $r \mapsto r+1$, we randomize a k-subset $x_1, x_2, \ldots, x_k$ while satisfying the identified constraints. The symmetry of these transitions, and the possibility of reaching any sequence in $\mathcal{S}(\mathbf{x}^{\text{input}})$ (ergodicity), guarantees asymptotic ($r \to \infty$) uniform sampling from the space of sequences which satisfy the required constraints [51].

B. From distributions to constraints

The key observation underlying our approach is that the likelihood in Eq. (1) can be kept constant by imposing a few specific constraints on the sequence $\{x_i\}$. For instance, constraining the likelihood of observations under the assumption of an i.i.d. truncated Gaussian distribution corresponds to fixing the sums $\sum_i x_i$ ($\propto$ first moment) and $\sum_i {x_i}^2$ ($\propto$ second moment). Our MCMC implementations discussed below succeed in fixing the following constraints:

(1) Thresholds: $x \geqslant x_{\min}$ or $x \leqslant x_{\max}$
(2) Geometric mean $(\prod_{i=1}^{N} x_i)^{1/N}$ or product $\prod_{i=1}^{N} x_i$
(3) Mean $\frac{1}{N}\sum_{i=1}^{N} x_i$ or sum $\sum_{i=1}^{N} x_i$
(4) Second moment $\sum_{i=1}^{N} {x_i}^2$.

We also combine two or more of these constraints and consider transformation of variables $y_i = f(x)$, e.g., $y_i = \log x_i$. Table I shows how these allow us to tackle some of the most common distributions used in statistical physics and time-series analysis.

C. From constraints to MCMC methods

We now construct MCMC methods that sample from sequences $\mathbf{x}$ that satisfy the constraints of Sec. III B, targeting uniform (i.e, constant probability) sampling of the $(N-K)$-dimensional manifold generically formed by imposing K constraints over N data points. For instance, in the Gaussian case, this corresponds to sampling uniformly from the set of

TABLE I. The likelihood function of important probabilistic models $p(x)$ is specified by one or more properties. Surrogates representing the model can be generated by randomizing the sequence while constraining these properties. The stated likelihood assumes the observed sequence sequence $x_1, x_2, \ldots, x_N$ is i.i.d. In cases for which the constrained properties are "$x_1, \ldots, x_N$," the only possible constrained surrogates are reachable by shuffling. In each case the constant C is defined such that the total probability is unity.

Name	Probability of x	Support	Likelihood	Constrained properties
Power-law	$Cx^{-\gamma}$	$[x_{\min}, \infty)$	$C^N\big(\prod_i x_i\big)^{-\gamma}$	$\prod_i x_i$
Lognormal	$C\exp\big[-\frac{1}{2}\big(\frac{\log x-\mu}{\sigma}\big)^2\big]/x$	$[x_{\min}, \infty)$	$C^N\big(\prod_i x_i\big)^{-1}\exp\big\{-\frac{1}{2\sigma^2}\big[\sum_i(\log x_i)^2 - 2\mu\log\prod_i x_i + N\mu^2\big]\big\}$	$\prod_i x_i$, $\sum_i(\log x_i)^2$
Exponential	$C\exp(-\lambda x)$	$[x_{\min}, \infty)$	$C^N\exp\big(-\lambda\sum_i x_i\big)$	$\sum_i x_i$
Gaussian	$C\exp\big[-\frac{1}{2}\big(\frac{x-\mu}{\sigma}\big)^2\big]$	$[x_{\min}, \infty)$	$C^N\exp\big[-\frac{1}{2\sigma^2}\big(\sum_i x_i^2 - 2\mu\sum_i x_i + N\mu^2\big)\big]$	$\sum_i x_i$, $\sum_i x_i^2$
Uniform	$(b-x_{\min})^{-1}$	$[x_{\min}, b]$	$(b-x_{\min})^{-N}$	$\max_{i\in\{1,\ldots,N\}} x_i$
Power-law, exp. cutoff	$Cx^{-\gamma}e^{-\lambda x}$	$[x_{\min}, \infty)$	$C^N\big(\prod_i x_i\big)^{-\gamma}\exp\big(-\lambda\sum_i x_i\big)$	$\sum_i x_i$, $\prod_i x_i$
Weibull, unknown k	$\frac{k}{\lambda}\big(\frac{x}{\lambda}\big)^{k-1}\exp[-(x/\lambda)^k]$	$[x_{\min}, \infty)$	$(\lambda^{-1}k)^N\big(\lambda^{-N}\prod_i x_i\big)^{k-1}\exp\big(-\lambda^{-k}\sum_i x_i^k\big)$	$x_1, \ldots, x_N$
Weibull, known $k \notin \{0, 1\}$				$\prod_i x_i$, $\sum_i x_i^k$
Shifted power-law, unknown δ	$C(x+\delta)^{-\gamma}$	$[x_{\min}, \infty)$	$C^N\big[\prod_i(x_i+\delta)\big]^{-\gamma}$	$x_1, \ldots, x_N$
Shifted power-law, known δ				$\prod_i(x_i+\delta)$

sequences that have the same first and second moment as the input sequence. We achieve this with localized transitions where in each step $r \mapsto r+1$ we propose a move $\mathbf{x} \mapsto \mathbf{x}'$. We use symmetric transitions, $P(\mathbf{x} \to \mathbf{x}') = P(\mathbf{x}' \to \mathbf{x})$, and ensure that all entries of $\mathbf{x}$ can be modified, leading to a Markov chain that (for $r \to \infty$) ergodically explores the constraint manifold with uniform probability [51].

In each iteration we randomize $K+1$ of the N values of $\mathbf{x}$,

$$\mathbf{x}_{K+1} := x_{i_1}, \ldots, x_{i_{K+1}} \mapsto x'_{i_1}, \ldots, x'_{i_{K+1}} := \mathbf{x}'_{K+1}, \qquad (2)$$

where the integers $i_1, \ldots, i_{K+1} \in [1, N]$ are chosen uniformly at random without replacement. Our aim is for the move to satisfy all the global constraints so that our exploration is restricted to the allowable sequences. We thus compute the constraints in $x_{i_1}, \ldots, x_{i_{K+1}}$ and impose the same constraints into $x'_{i_1}, \ldots, x'_{i_{K+1}}$. In the Gaussian example, $K=2$, and we will therefore choose three random points from $\mathbf{x}$ to modify without altering their sum and sum of squares. Generically, the choice about how to modify them is a one-dimensional problem because the K constraints will be satisfied in some subset of a curve in $\mathbb{R}^{K+1}$ parametrized by $\gamma(s)$, for $s \in [a, b]$. The choice of the move (new values of the points) is thus reduced to the random choice of $s \in [a, b]$ that determines $\mathbf{x}'$.

The distribution $P(s)$ of $s \in [a, b]$ implies the probability with which sequences $\mathbf{x}$ are sampled. To guarantee symmetric transition probabilities, we choose destinations uniformly at random along a 1-manifold (made up of one or more curves) of points which preserve the K constrained properties. To see how the desired uniform sampling can be achieved, we transform to coordinates $\boldsymbol{\beta}, s$, where $\boldsymbol{\beta}$ is the K vector listing the K constrained properties. In these coordinates, the probability of choosing a point $\mathbf{x}_{K+1}$ is

$$P(\mathbf{x}_{K+1}) = P(\boldsymbol{\beta})P(s)\left|\det\frac{\partial\boldsymbol{\beta}, s}{\partial\mathbf{x}_{K+1}}\right|,$$

where $\frac{\partial\boldsymbol{\beta},s}{\partial\mathbf{x}_{K+1}}$ is the Jacobian matrix of the transformation from coordinates $\mathbf{x}_{K+1}$ to coordinates $\boldsymbol{\beta}, s$. Uniform sampling, i.e., constant $P(\mathbf{x}_{K+1})$, is therefore equivalent to a product $P(s)|\det\frac{\partial\boldsymbol{\beta},s}{\partial\mathbf{x}_{K+1}}|$ which is independent of s, or to

$$P(s) = C\left|\det\frac{\partial\boldsymbol{\beta}, s}{\partial\mathbf{x}_{K+1}}\right|^{-1}, \qquad (3)$$

where C is a normalization constant.

An illustration of one MCMC step in the case of truncated Gaussians is shown in Fig. 3. All the cases in which we implement this general framework are summarized in Table II (for their derivations, see Appendix A). Constrained uniform surrogates can be generated without iteration. An important practical point is to determine the number of MCMC transitions to be used in the sampling. In our case, we can show that (see Appendix B)

$$\tilde{r} = \frac{N}{2}(\log_2 N - \log_2\epsilon) \qquad (4)$$

transitions are sufficient to achieve a relative error in the expected value (of either a sequence of length N or its log-transform) of no more than ϵ. In our simulations, unless stated otherwise, we use $2\tilde{r}$ (with $\epsilon = 2^{-10}$) transitions between samples surrogate samples and multiple samples are generated starting from the previous one.

IV. PERFORMANCE

In this section we use synthetic data to illustrate the properties of constrained surrogates and explore their suitability

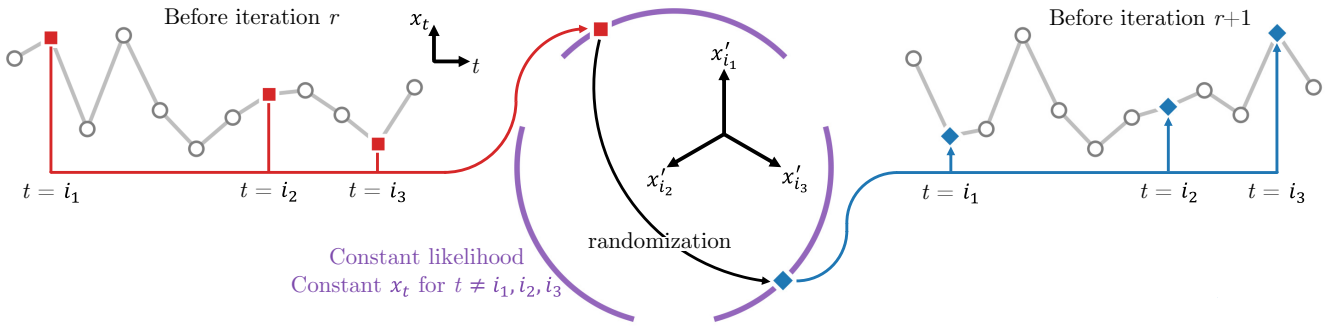


FIG. 3. Illustration of the proposed algorithm for generating constrained likelihood surrogates. The figure concentrates on one MCMC step for the case of a truncated ($x \geqslant x_{\min}$) Gaussian distribution. In this case the likelihood is specified by $K = 2$ continuous properties which we must preserve to produce constrained surrogates (the same general approach applies to other distributions). In each iteration r we randomly choose $K + 1$ elements $x_{i_1}, \ldots, x_{i_{K+1}}$ (red squares) from the sequence. These correspond (red line) to a point (red square) on the curve (purple arcs) giving the collection of points in $x'_{i_1} \cdots x'_{i_{K+1}}$ space (black coordinate axes) which can replace $x_{i_1}, \ldots, x_{i_{K+1}}$ without altering the likelihood. The next location (blue diamond) is selected uniformly at random (black arrow), and the sequence is updated according to $x_{i_1} = x'_{i_1}, \ldots, x_{i_{K+1}} = x'_{i_{K+1}}$ (blue arrows).

for different tasks. Using synthetic data of the types listed in Table III, we consider surrogates designed to represent the hypothesis that data arose i.i.d. under five different families of distribution $\{p_\xi(x)\}$: power law, truncated lognormal, exponential, truncated Gaussian, and uniform. The surrogates we apply include constrained surrogates, as defined in Sec. III, and typical surrogates, generated by first determining the parameters $\hat{\xi}$ that maximize the likelihood of the hypothesized model and then generating i.i.d. sequences from the distribution $p_{\hat{\xi}}(x)$ with parameter $\hat{\xi}$.

A. Hypothesis testing

First we consider using surrogates to test the validity of a hypothesized null model. As usual in hypothesis testing, this null hypothesis is rejected if the p value—defined as the probability under the null hypothesis of observing a value of a discriminating statistic at least as extreme as the value observed in the input sequence $\mathbf{x}^{\text{input}}$—falls below a predetermined threshold parameter termed the *nominal size* (frequently 0.05 or 0.1). Surrogates enable computationally efficient hypothesis testing because they allow this probability to be estimated by evaluating the discriminating statistic across an ensemble of surrogates $\{\mathbf{x}_n\}$. The standard method for assessing goodness-of-fit involves hypothesis tests that use as discriminating statistic the Kolmogorov-Smirnov (KS) distance, computed relative to the maximum likelihood parameters of the portion of the dataset above the lower cutoff [8]. In this section, to illustrate the capacity of constrained surrogates to provide theoretical support for arbitrary discriminating statistics and also because of its physical relevance (e.g., for investigating extreme events), we instead employ as a test statistic the sample maximum, a quantity of interest in the study of extreme events. We focus on lower-directed hypothesis tests and quantify the performance of surrogate methods via two key metrics:

(1) The *size* of a test is the rate of incorrect rejection of the null hypothesis, i.e., rejection of the null hypothesis when it is actually true. Ideally a test is *exact* or *accurate*, which occurs when its size equals the predetermined nominal size.

TABLE II. Key ingredients for constrained surrogates representing five important families of distributions. In each iteration of the algorithm to generate a constrained surrogate representing a given distribution, we randomly choose $K + 1$ elements $x_{i_1}, \ldots, x_{i_{K+1}}$ and from these determine the values of the K properties $\beta_1, \ldots, \beta_K$ to be constrained. We then randomize the values of the elements $x_{i_1}, \ldots, x_{i_{K+1}}$ by picking a point along the curve $\gamma : s \mapsto \gamma(s)$ by choosing s with probability $P(s)$ from the specified domain. R is a fixed 3×3 rotation matrix (see Appendix A), $\rho(\boldsymbol{\beta}) = \frac{1}{\sqrt{3}}\sqrt{3\beta_2 - \beta_1^2}$, and the exponential function exp is applied element-wise.

Name	Constrained properties $\boldsymbol{\beta}$	Curve $\gamma : \mathbb{R} \to \mathbb{R}^{K+1}$, $s \mapsto (x_{i_1}, \ldots, x_{i_{K+1}})$	Probability $P(s)$	Domain of s
Exponential	$\beta_1 = x_{i_1} + x_{i_2}$	$(s, \beta_1 - s)$	β_1^{-1}	$[0, \beta_1]$
Gaussian	$\beta_1 = \sum_{j=1}^3 x_{i_j}$, $\beta_2 = \sum_{j=1}^3 x_{i_j}^2$	$r(\boldsymbol{\beta})R\,(\cos t, \sin t, 0)^T + (1,1,1)^T\beta_1/3$	$1/2\pi$	$[0, 2\pi)$
Powerlaw	$\beta_1 = x_{i_1}x_{i_2}$	$(s, \beta_1/s)$	$[s\log(\beta_1)]^{-1}$	$[1, \beta_1]$
		$(\exp s, \beta_1/\exp s)$	$\log(\beta_1)^{-1}$	$[0, \log(\beta_1)]$
Lognormal	$\beta_1 = \prod_{j=1}^3 x_{i_j}$, $\beta_2 = \sum_{j=1}^3 (\log x_{i_j})^2$	$\exp[\rho(\boldsymbol{\beta})R\,(\cos s, \sin s, 0)^T + (1,1,1)^T(\log\beta_1)/3]$	$1/2\pi$	$[0, 2\pi)$
Uniform[a]	$\beta_1 = \max_{i\in\{1,\ldots,N\}} x_i$	–	–	–

[a]To generate constrained surrogates representing the uniform distribution we must preserve the maximum value but do not need other components of our framework (see Appendix A).

TABLE III. The probability density functions used to generate our synthetic data with lower cutoff $x_{\min} = 1$. In each case the constant C is defined such that the total probability is unity.

Name	Probability	Support	Parameter(s)
Exponential	$p(x) = Cx^{-\lambda}$	$[x_{\min}, \infty)$	$\lambda = 1$
Gaussian (truncated)	$p(x) = C\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)^2\right]$	$[x_{\min}, \infty)$	$\mu = -1, \sigma = 1$
Power-law	$p(x) = Cx^{-\gamma}$	$[x_{\min}, \infty)$	$\gamma = 2.5$
Lognormal (truncated)	$p(x) = C\frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\left(\frac{\log x-\mu}{\sqrt{2}\sigma}\right)^2\right]$	$[x_{\min}, \infty)$	$\mu = -1, \sigma = 1$
Uniform	$p(x) = C$	$[x_{\min}, b)$	$b = 9$

(2) The *power* of a test is the rate of correct rejection of the null hypothesis, i.e., rejection of the null hypothesis when it is incorrect. This rate depends not only on surrogate method and discriminating statistic, but also on the true generating process which underlies the input sequence.

In Fig. 4 we consider both accuracy and power. To assess size (in panels with red borders) we generate $\mathbf{x}^{\text{input}}$ as an i.i.d. sample from a given family of distributions and test whether different types of surrogates reject the hypothesis that the sample arose under this distribution or a generalization (power-law and exponential distributions are a limiting case of truncated lognormal and Gaussian distributions). We compare how well-constrained and typical surrogates allow us to match the hypothesis test's true size with a chosen nominal size. We see that typical surrogates consistently undershoot nominal size for small sample lengths N and that, for sequences arising under lognormal, Gaussian, or uniform processes, this deviation remains evident even for the largest sample lengths considered. This underestimation can be interpreted as advantageous because it reduces the probability of incorrectly rejecting a true exponential, Gaussian, or lognormal process, but also entails inaccuracy, because size should be a parameter which can be specified before testing [4]. Furthermore, a mismatch between the true and nominal size is inherently disturbing, because it implies we are not correctly representing the null hypothesis. Turning to power (in panels with purple borders), we generate $\mathbf{x}^{\text{input}}$ from a distribution and test whether different surrogate methods can reject the incorrect hypothesis that the data arose from a distinct probability functions. We see that constrained surrogates frequently provide higher power to reject an incorrect hypotheses. When the null model is uniform, this higher power equals the nominal size. This occurs because constrained surrogates preserve the maximum-likelihood estimate of the model parameter, which in this case is also the discriminating statistic. Other tests show different patterns: in upper-directed tests also using sample maximum or using KS distance, typical surrogates can offer higher power (see Supplemental Material Figs. S1 and S2 [52]); and for two-tailed tests involving log-kurtosis or coefficient of variation the more powerful surrogate depends on both the true underlying process and the null hypothesis (see Figs. S3 and S4 [52]). As a rule of thumb, we found that for common descriptive statistics, constrained (typical) surrogates tended to be better for lower-directed (upper-directed) tests (see Figs. S5 and S6 [52]). The generality of this rule is surprising given that a lower-directed test using a given test statistic is equivalent to an upper-directed test using the negation of that test statistic.

Finally, we test whether our prescription of using $\tilde{r}$ MCMC transitions [see Eq. (4)], which was obtained by bounding the relative error of a sequence, is also appropriate for hypothesis testing. To do so, we examine how size and power vary with the number of transitions used for fixed data length $N = 64$. We found that size remains at or below nominal size and stabilizes to the nominal size by or just after $\tilde{r}$ transitions (see Fig. S7 [52]). In almost all cases, power has also plateaued by $\tilde{r}$ transitions, although when the true process is power-law or lognormal, the power provided by constrained Gaussian surrogates can continue to increase beyond this number of transitions.

B. Estimation

Now we utilize surrogates in estimating properties of an underlying system, with a focus on its extreme events. For constrained surrogates, these estimates are based both on the observed time series and the model (family of distributions) assumed to describe the underlying process, but not the parameter value defining the precise process (e.g., assuming a power-law distribution but not the exponent γ). Specifically, we take as $\mathbf{x}^{\text{input}}$ synthetically generated sequences from an underlying process, generate surrogates with the same size N as $\mathbf{x}^{\text{input}}$, and for each surrogate compute the sample maximum (i.e., the maximum of the N values). The first row of Fig. 5 reveals that expectation calculated using constrained surrogates coincides with expectation under the true process, implying that—as expected theoretically—constrained surrogates provide an unbiased estimate of the sample maximum. This contrasts with the typical approach, which leads to expectations that can either underestimate or overestimate, especially for lower sample length N. The second row shows that constrained surrogates also reduce the variance in this estimate relative to the original observed maximum and, for the heavy-tailed power-law and lognormal processes, also lead to lower variance than the biased estimate based on typical surrogates. Finally, from the third row we see that constrained surrogates consistently reduce error in estimates of sample expectation, and that in three of the five cases the improvement is greater than under typical surrogates. In some cases, particularly for low sample length N, typical surrogates can increase error in this estimate. Constrained surrogates also exhibit advantages for other statistics, especially when sample length N is small (see Fig. S8 [52]).

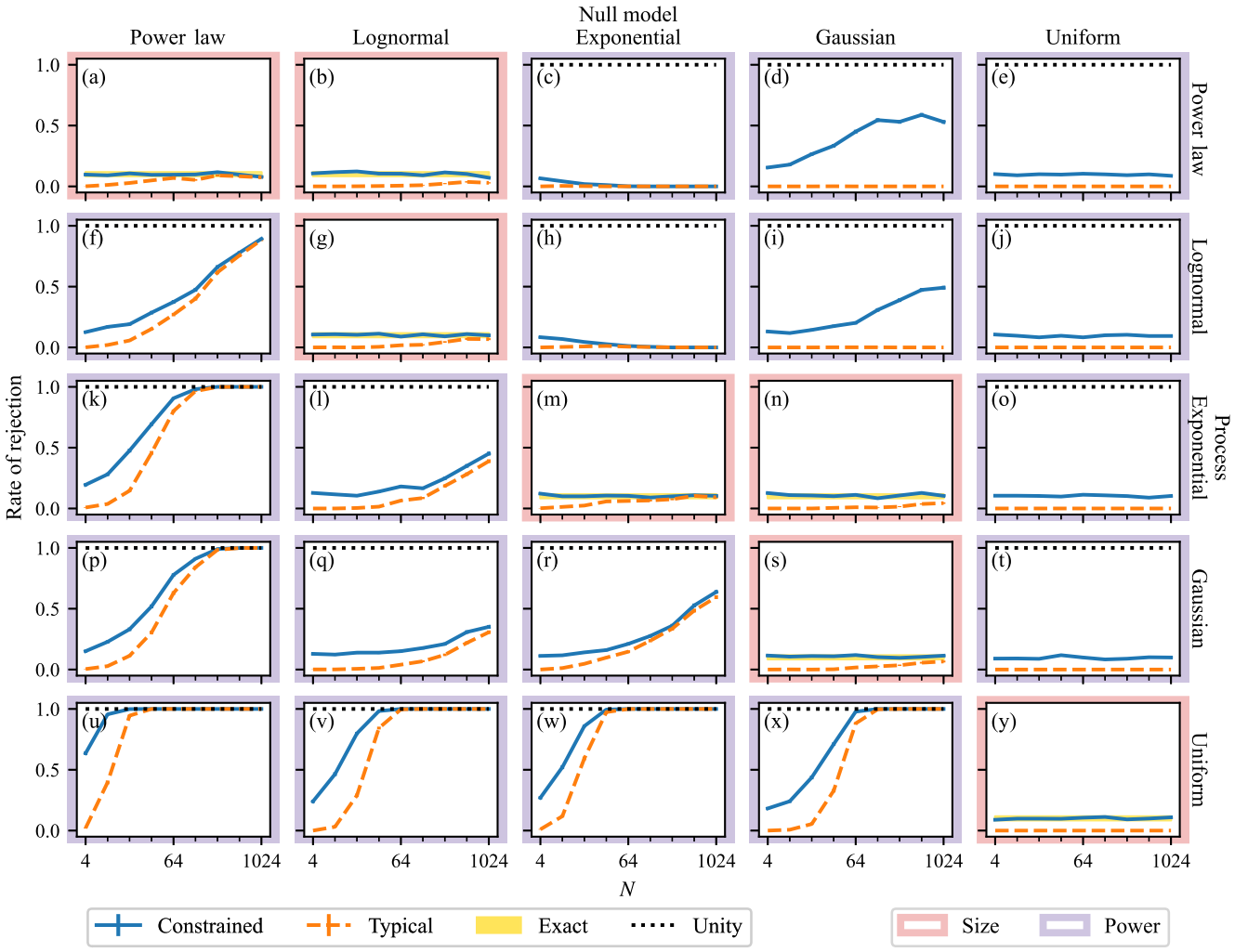


FIG. 4. Constrained surrogates provide exact hypothesis tests for arbitrary discriminating statistics and can increase power to reject false hypotheses. The power (size) of the test of null hypothesis was estimated from 1000 tests, each using N i.i.d. samples and 19 surrogates. Results are shown for different combinations of underlying distribution (from top to bottom) and null hypothesis (from left to right): power-law; lognormal; exponential; truncated Gaussian; and uniform distribution. Tests are lower-directed, have nominal size 10%, and use as the sample maximum as discriminating statistic. Panels showing tests for which the null hypothesis is correct (i.e., for which the rate of rejection is the size of the test), comprising all panels on the diagonal and two off-diagonal panels, have red borders. These six panels feature a gold band which spans the 90% confidence interval for a Bernoulli process, but the band is thin and is often obscured by the blue line showing the size achieved by constrained surrogates. Panels showing tests for which the null hypothesis is incorrect (i.e., for which the rate of rejection is the test's power to reject an incorrect hypothesis), comprising the remaining panels, have purple borders and feature a dotted black line corresponding to the ideal (100%) rejection rate. Error bars correspond to standard error but are often indistinguishable from the line width.

V. APPLICATIONS

In this section we apply surrogate methods to estimation and hypothesis testing in data representing diverse types of empirical complex system.

A. Bushfires

First, we consider a time series describing the magnitudes of bushfires that occurred in Australia from July 2019 until June 2020, representing one of Australia's worst recorded bushfire years. Bushfires were inferred from satellite imaging data as described in Ref. [39] and their sizes, in units of $500\,\mathrm{m}^2$, range from 1 to 12 446. The severity of bushfires in the considered period has been linked to a critical transition associated with power-law scaling of bushfire magnitude [39], in line with a long tradition of linking fires, natural disasters, extreme events, critical phenomena, and power-law distributions [32,34,35,37,53].

In Fig. 6 we investigate the hypothesis that the distribution of bushfire magnitudes is lognormal. This distribution contains power laws as a particular case so that we are also testing the most popular distribution. Visually, the log-binned empirical probability density convincingly follows a straight line over several orders of magnitude. The tails are also well described by the lognormal distribution up to the maximum observed bushfire magnitude, with the maximum likelihood fit deviating only imperceptibly from a straight line corresponding to a power-law distribution. In our quantitative analysis of the tail of the distribution we used the lower cutoff

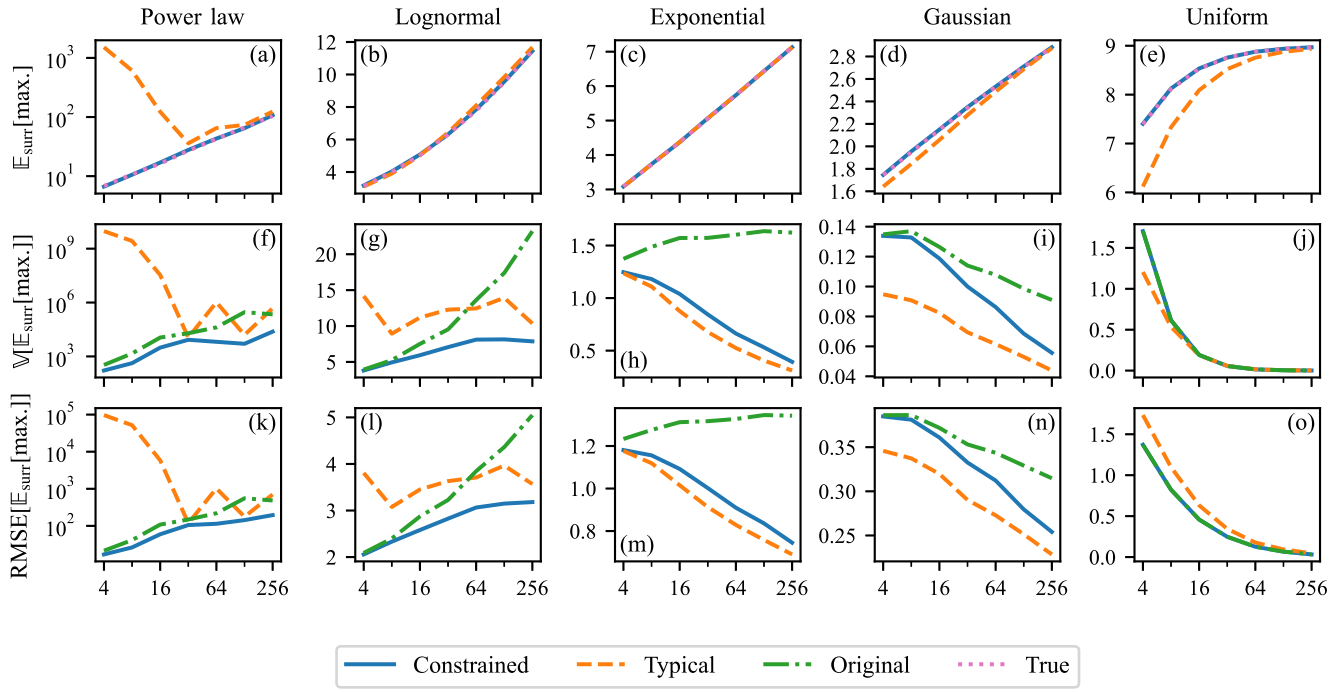


FIG. 5. Constrained surrogates do not bias sample statistics, and can reduce variance and error. Expectation (first row), variance $\mathbb{V}$ (second row), and root mean square error (RMSE, third row) of estimated expectation $\mathbb{E}_{\text{surr}}$ of sample maximum. The results are plotted against the length N of the input sequence $\mathbf{x}^{\text{input}}$, which is equal to the length of the surrogate sequences $\mathbf{x}$. $\mathbb{V}$ is calculated using 10^4 independent input sequences $\mathbf{x}^{\text{input}}$, each comprising N values drawn i.i.d. from a (from left to right): power-law, lognormal, exponential, truncated Gaussian, and uniform distribution. $\mathbb{E}_{\text{surr}}$ is estimated using 100 surrogates, each from the same $\mathbf{x}^{\text{input}}$. The surrogates considered are either constrained (blue solid), typical (orange dashed), unchanged from the original observation (green dot-dashed), or generated independently from the true underlying process (pink dotted).

$x_{\text{min}} = 32.5$ chosen via the widely used method of minimizing the KS distance relative to the maximum likelihood parameters [8]. Repeating the analysis using $x_{\text{min}} = 16.5$ and 64.5 produced the same qualitative results. Using the KS distance as discriminating statistic, as recommended in Ref. [8], to test the hypothesis that the dataset arose under a lognormal distribution, we observe rates of rejection of an i.i.d. lognormal hypothesis not much larger than the nominal size of the test [across a range of values of N; see Fig. 6(b)]. All this evidence—obtained following standard statistical recipes in

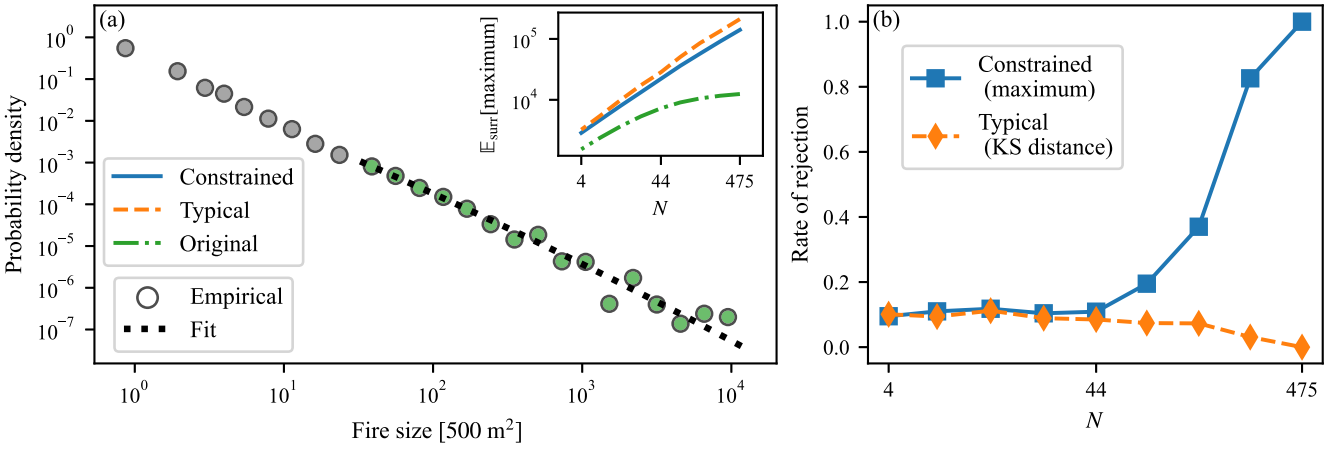


FIG. 6. Constrained surrogates improve conclusions about the distribution of bushfire magnitudes. (a) Log-binned empirical distribution of fire magnitude (circles) and the maximum likelihood fit to a lognormal model (dotted black line) of values above the lower cutoff. Inset: Expectations of sample maximum estimated as the mean over 10^5 independent samples of length N downsampled (i.e., drawn i.i.d. without replacement) either from the original data or surrogates produced from the full dataset. (b) Rates of rejection estimated from 1000 hypothesis tests, each using a sample of length N from the observed dataset, 99 surrogates, and either the KS distance together with typical surrogates or maximum with constrained surrogates. Only values of the original data which exceed the fitted lower cutoff $x_{\text{min}} = 32.5$ [500 m^2] are considered.

the field [47]—corroborates the connection between bushfires magnitudes and power-law or criticality models.

However, the conclusions above are in conflict with the strong deviations between the frequency of extremely large bushfires and the power-law or lognormal predictions. This is evident when comparing the observation and predictions of maximum fire magnitude [see inset of Fig. 6(a)]. Across the range of considered data lengths, the model predictions substantially exceed those based on subsampling the original dataset, although the use of constrained surrogates does slightly ameliorate these overestimates. Fortunately, constrained surrogates enable us to use arbitrary test statistics in hypothesis tests, and we therefore turn to a statistic of socioeconomic significance: the sample maximum. Using this statistic and constrained surrogates in lower-directed hypothesis tests, rates of rejection increase to unity as the sample length approaches that of the original dataset. For lower sample lengths, as expected from benchmarking, constrained surrogates can provide substantially higher rates of rejection than can typical surrogates (see Fig. S9 [52]). Similar results are reached for the power-law model (Fig. S10 [52]). Altogether, constrained surrogates allowed us to reach the nuanced yet significant conclusions that, while power-law and lognormal distributions provide a good description—spanning four orders of magnitude—of the bulk of the Australian bushfire distribution, these models overestimate the probability of extremely large fires.

B. Words

Next, we investigate the distribution of repeated appearances of the same word in written English. We quantify this by counting the number x of word tokens (waiting times) between word repetitions within a text. Deviations of the waiting times from an exponential distribution, expected from a Poissonian (i.e., memoryless) null hypothesis, have been used to characterize the burstiness and connected to different (syntactic and semantic) properties of the words [32,55–57].

In Fig. 7 we consider using constrained surrogates to test the hypothesis that each waiting time sequence arose under an i.i.d. exponential process, employing as discriminating statistic the coefficient of variation σ_x/μ_x given by the ratio of the standard deviation to the mean. The coefficient of variation quantifies burstiness: for an i.i.d. exponential process and large sample sizes N, it approaches one. We compare our surrogate-based approach with the simpler strategy of directly inspecting the coefficient of variation. For words with $N = 30$ (main panel in Fig. 7), we see the direct connection between the p value of our test, σ_x/μ_x, and the role played by different words in the text: function words have values with $\sigma_x/\mu_x \approx 1$ (e.g., "interest"), topical words have $\sigma_x/\mu_x \gg 1$ are (e.g., "strait"), and the rare words with regular appearance have $\sigma_x/\mu_x \ll 1$ (e.g., the word "chapter": chapters of Darwin's text are of similar size and open with "Chapter n"). In testing an i.i.d. exponential hypothesis, as expected for this low sample length, we see high p values when coefficient of variation is close to unity. However, the observed p values rapidly decrease as we move away from this value of coefficient of variation. Extending our scope to the words for which $1 \leqslant \sigma_x/\mu_x < 1.5$, we still observe p values

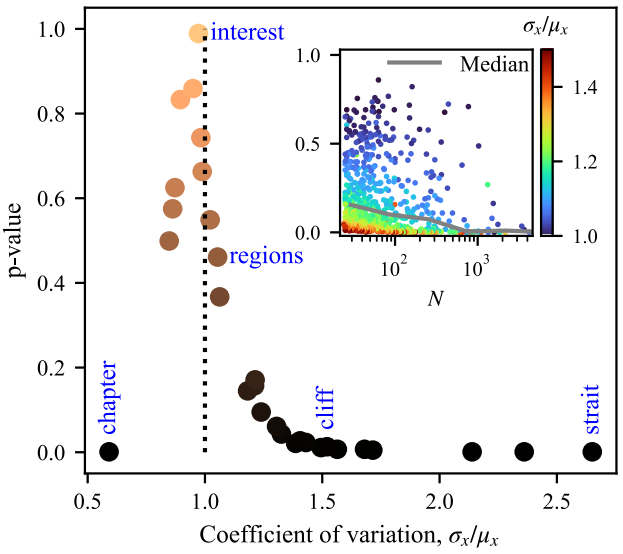


FIG. 7. Constrained surrogates to test the distribution of words in texts. Reported is the outcome of the tests of an exponential distribution of waiting times x between repetitions of the same word in the book *The Voyage of the Beagle* by Charles Darwin. The main panel shows the relation between the p value of the test and the coefficient of variation σ_x/μ_x for words with $N = 30$ waiting times, with five words highlighted and marker color indicating p value. Inset: p value versus N. Each marker corresponds to a word with $1 \leqslant \sigma_x/\mu_x < 1.5$, and its color indicates its corresponding coefficient of variation σ_x/μ_x. The median p value is shown as a function of N (gray line). Hypothesis tests use 999 constrained surrogates and employ σ_x/μ_x as discriminating statistic. We consider the text's 1000 most frequent words and a lower cutoff $x_{\min} = 9.5$, which approximates the typical length of a clause in written English [54]; repeating the analysis using $x_{\min} = 4.5$ and $x_{\min} = 19.5$ does not qualitatively change results.

tending to decrease as coefficient of variation departs from a value of 1. For a given coefficient of variation, p values also decrease as N increases, with fluctuations for the largest values of N, where data are scarce. Other ranges of σ_x/μ_x exhibit analogous patterns (see Fig. S11 [52]). This reflects how, for a sufficiently long sample, even a small departure from a hypothesized family of distributions becomes statistically significant. Our constrained surrogate-based approach to testing for an exponential distribution thus accounts both for the burstiness (quantified by σ_x/μ_x) and also the sample size N.

C. Brain data

Now we turn our attention to neuron connectivity. Specifically, we examine the sequence listing each neuron's degree (number of other neurons with which a synaptic connection is shared) and strength (number of synaptic connections with other neurons), both of which have been described as lognormally distributed across all neurons in the brain of the nematode *C. elegans* [58]. Of particular interest in debates about distributions in complex systems is the fatness of their tails, which can be assessed with a range of statistics [59]. In our case an especially pertinent measure is the excess kurtosis of the log transformed data (hereafter called "log-kurtosis"), because this is precisely zero for a nontruncated lognormal

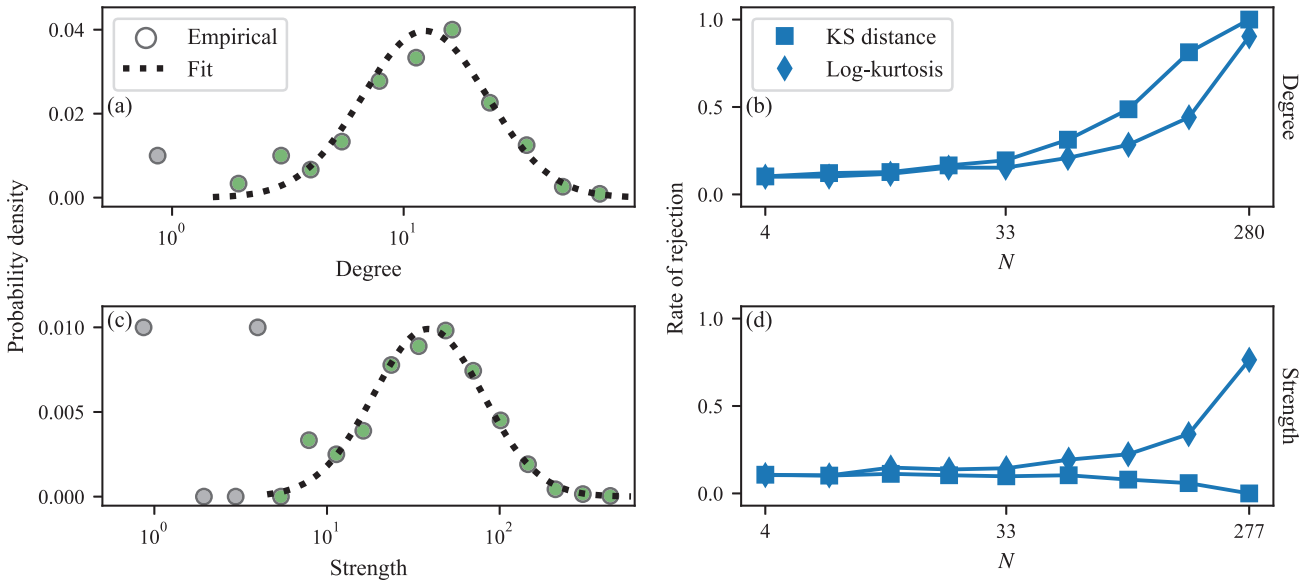


FIG. 8. Constrained surrogates lead to improved characterization of neuron connectivity. Conclusions about the validity of lognormal models depend on whether we focus on goodness of fit or fatness of tails. (a), (b) Log-binned empirical distribution (circles) of (a) degree (number of other neurons with which a synaptic connection is shared) and (b) strength (number of synaptic connections with other neurons) in the neuron-resolution connectome of *C. elegans*, together with the maximum likelihood fit to a lognormal model (dotted black line). (c), (d) Rates of rejection estimated from 1000 hypothesis tests, each using a sample of length N from the sequence of observed (c) degrees and (d) strengths above the lower cutoff, 99 constrained surrogates and either the KS distance (measuring goodness of fit) or log-kurtosis (measuring fatness of tails). Only values which equal or exceed the lower cutoffs $x_{\min} = 1.5$ (for degree) and $x_{\min} = 4.5$ (for strength) are considered.

distribution and positive (negative) for a distribution with fatter (thinner) tails.

In Fig. 8 we investigate whether observed goodness-of-fit and fatness of tails are consistent with i.i.d. lognormal distributions in degree and strength, choosing the lower cutoff $x_{\min}$ for each property as the lowest half-integer which excludes the most substantial deviations from lognormality. Visual inspection reveals good fits over a substantial range, but, for more quantitative insights, we consider upper-directed hypothesis tests based on the KS distance from the maximum likelihood fit to an i.i.d. lognormal process. For the degree (strength) dataset, as sample length increases, rate of rejection increases towards unity (remains around or below the test's nominal size). Therefore, according to goodness of fit as captured by KS distance, the degree (strength) distribution is consistent (inconsistent) with an i.i.d. lognormal distribution. However, turning to two-sided hypothesis tests employing the log-kurtosis, for both the degree and strength datasets we see rates of rejection increasing to well above nominal size as sample length increases. Therefore, in terms of the physically relevant log-kurtosis property, neither degree nor strength is consistent with an i.i.d. lognormal process. This example shows the importance of the choice of the statistic used in a statistical test and therefore highlights the key advantage of constrained surrogates, which can be used to test an entire family of distributions based on arbitrary test statistics.

D. City populations

Finally, we consider the paradigmatic problem of describing the distribution of the population x of the N largest cities within a country. The functional form of such city-size distributions has been discussed for more than a century, with the power-law and lognormal models as leading candidates [10,32]. We apply our surrogate approaches to datasets from five countries: Australia, Brazil, Germany, the United Kingdom, and the United States. For the two largest datasets (Brazil and the United States), all five models are rejected at the 0.1 level of significance; for the three other countries, only the lognormal model avoids rejection in all cases (see Supplemental Material Table S1 and Fig. S12 [52]). These results corroborate earlier work that favored the lognormal over the power-law for the full distribution of U.S. city populations [10].

VI. DISCUSSION AND CONCLUSION

Surrogate methods connect empirical observations with theoretical models and thus permit quantitative investigation of complex systems where analytic approaches are intractable. In this work we introduced a general methodology for generating constrained surrogates representing entire families of continuous distributions, including some central to statistical physics and complex systems: Gaussian, lognormal, power-law, and exponential distributions. The resulting algorithms enable surrogate generation for arbitrary datasets, without specification or estimation of model parameters. Code to generate surrogates for arbitrary time series is available in our repository [60].

The significance of our constrained surrogates is that they lead to improved statistical methods to

(i) Test the validity of an entire family of distributions $\{p_\xi(x)\}$ using arbitrary test statistics and

(ii) Obtain unbiased estimates of arbitrary statistics under the hypothesized distribution family.

These improvements derive from the fact that constrained surrogates sidestep the problem of parameter estimation to simultaneously represent a whole family of processes, thereby providing accurate hypothesis tests and unbiased estimates for arbitrary, physically meaningful quantities. We showed the relevance of these theoretical advantages in tests with synthetic data (Sec. IV) and four paradigmatic applications of statistical laws in complex systems (Sec. V). For instance, we found that only through the use of test statistics associated with extreme events are we able to reject the hypothesis of power-law fire sizes in Australia, and thereby allow a more accurate assessment of the probability of extreme fires.

Which type of surrogate is preferable in practice depends on the choice of (test) statistic and, for hypothesis tests, whether the test is lower or upper directed. The main practical advantage of constrained surrogates, which follows from its theoretical properties, is that they guarantee exact test size without statistic-specific benchmarking. A limitation is that constrained surrogates are necessarily uninformative for statistics fixed by the constrained properties themselves, such as maximum-likelihood estimates of model parameters. In our numerical analysis of a wide range of descriptive statistics, constrained surrogates tended to provide higher power for lower-directed tests, whereas typical surrogates tended to provide higher power for upper-directed tests (see Figs. S5 and S6 [52]). The computational cost to generate constrained surrogates is also higher and could restrict their use at large data size N. On a desktop computer, we estimate that a surrogate for $N = 1300$ can be generated within one second and for $N = 67\,000$ within 1 min. The computational time is roughly proportional to the number of MCMC transitions $\tilde{r} = O(N \log N)$ [see Eq. (4)] with the constant of proportionality depending mainly on the null hypothesis and, to a lesser extent, on the input data (see Fig. S13 for the computational cost needed in different settings [52]).

More broadly, our constrained surrogates provide a parameter-agnostic framework for the analysis of statistical laws in complex systems that moves beyond traditional goodness-of-fit testing. By enabling hypothesis tests and estimation based on arbitrary statistics, this approach offers a principled way to distinguish agreement in some aspects of a distribution from systematic discrepancies in others. As such, constrained surrogates contribute directly to ongoing debates about the interpretation and applicability of statistical laws in empirical systems [32,61,62]. Important directions for future work include applying the present framework to additional distribution families, extending it from point to interval estimation, and relaxing the i.i.d. assumption to incorporate correlations and temporal dependencies [16,63].

ACKNOWLEDGMENTS

We gratefully acknowledge the support and hospitality of the Sydney Mathematical Research Institute (SMRI). J.M.M. is supported by the National Natural Science Foundation of China (Grants No. T2541015, No. W2632005, and No. 12375036).

DATA AVAILABILITY

The data that support the findings of this article are openly available [60].

APPENDIX A: SPECIFIC CASES

We now present methods for randomizing $x_{i_1}, \ldots, x_{i_{K+1}}$ to $x'_{i_1}, \ldots, x'_{i_{K+1}}$ leading to uniform sampling among all sequences that have equal likelihood under the considered family of distributions.

1. Sum

Each iteration we randomize x_{i_1}, x_{i_2} while maintaining the value of $\beta_1 = x_{i_1} + x_{i_2}$. We consider the change of coordinates

$$\beta_1 = x_{i_1} + x_{i_2},$$
$$s = x_{i_1},$$

which corresponds to Jacobian

$$\left|\det \frac{\partial \boldsymbol{\beta}, s}{\partial \mathbf{x}_{K+1}}\right| = \left|\det \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix}\right| = 1.$$

It follows from Eq. (3) that we can achieve symmetric transition probabilities by choosing s with constant probability density, i.e., uniformly on $[0, \beta_1]$.

With s chosen uniformly at random, transition probabilities are symmetric. Because a transition which does not change the current state is always possible, the Markov chain defined by the algorithm is aperiodic. Therefore, the algorithm will sample uniformly at random from some set satisfying the constraint. To show that our algorithm has full range, i.e., that it samples uniformly at random from the *entire* collection of sequences with the same likelihood, we show that it is possible, using the algorithm, to reach a chosen sequence $\mathbf{x}''$ starting from any sequence $\mathbf{x}$ such that $\sum x_i = \sum x''_i$. Let x''_i be the maximum value of $\mathbf{x}''$, and let x_j be the largest value of $\mathbf{x}$ except, possibly, for x_i. If $x_i + x_j \geqslant x''_i$ then perform the transition $x_i, x_j \to x''_i, x_i + x_j - x''_i$, otherwise, change $\mathbf{x}$ according to $x_i, x_j \to x_i + x_j, 0$. After several repeats of this operation, we will have $x_i = x''_i$. The sequences $\mathbf{x}$ and $\mathbf{x}''$ now fail to match in at most $N - 1$ places and, repeating the argument, we are able to iteratively reduce the number of mismatches until $\mathbf{x}$ and $\mathbf{x}''$ coincide at every point.

2. Sum and second moment

Each iteration we randomize $x_{i_1}, x_{i_2}, x_{i_3}$ while maintaining the values of $\beta_1 = x_{i_1} + x_{i_2} + x_{i_3}$ and $\beta_2 = x_{i_1}^2 + x_{i_2}^2 + x_{i_3}^2$. The second condition describes a sphere in $\mathbb{R}^3$ with center at the origin and radius $\sqrt{\beta_2}$, while the first represents a plane with normal $\hat{\mathbf{n}} = \frac{1}{\sqrt{3}}(1, 1, 1)^T$ and passing through the point $\mathbf{c} = \frac{\beta_1}{3}(1, 1, 1)^T$. The set of allowable points lie on the intersection of these surfaces, which is a circle with center $\mathbf{c}$ and radius

which can be identified as

$$\rho(\boldsymbol{\beta}) = \frac{1}{\sqrt{3}}\sqrt{3\beta_2 - {\beta_1}^2}.$$

We can transform from a coordinate system (x, y, z) in which this circle is centered at the origin and lies in the xy plane via

$$\begin{pmatrix} x_{i_1} \\ x_{i_2} \\ x_{i_3} \end{pmatrix} = R\begin{pmatrix} x \\ y \\ 0 \end{pmatrix} + \frac{1}{\sqrt{3}}\begin{pmatrix} z \\ z \\ z \end{pmatrix},$$

where

$$R = \begin{pmatrix} 0 & \sqrt{\frac{2}{3}} & \frac{1}{\sqrt{3}} \\ -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{3}} \end{pmatrix}$$

is a rotation matrix. The rotation matrix R has been identified by choosing its first two columns, to which R maps unit vectors in the x and y directions, as orthonormal vectors in the plane normal to $\hat{\mathbf{n}}$, and choosing its third column as $\hat{\mathbf{n}}$ itself. The first two columns are ordered such that the cross product of the first with the second is $\hat{\mathbf{n}}$ (rather than $-\hat{\mathbf{n}}$), which leads to a determinant of $+1$ (rather than -1). Moving to coordinates $\boldsymbol{\beta}, s$ such that

$$\begin{pmatrix} u \\ v \\ z \end{pmatrix} = \begin{pmatrix} \rho(\boldsymbol{\beta})\cos s \\ \rho(\boldsymbol{\beta})\sin s \\ \beta_1/\sqrt{3} \end{pmatrix},$$

we reduce randomization of $(x_{i_1}, x_{i_2}, x_{i_3})^T$ to a choice of $s \in [0, 2\pi)$.

To determine the appropriate distribution for s, we need to know how the Jacobian $|\det \frac{\partial \boldsymbol{\beta}, s}{\partial \mathbf{x}_{K+1}}|$ depends on s. To determine this, we note that the Jacobian of the inverse transformation can be written

$$\left|\det \frac{\partial \mathbf{x}_{K+1}}{\partial \boldsymbol{\beta}, s}\right| = \left|\det \frac{\partial \mathbf{x}_{K+1}}{\partial x, y, z}\right|\left|\det \frac{\partial x, y, z}{\partial \boldsymbol{\beta}, s}\right|.$$

The first factor is

$$\left|\det \frac{\partial \mathbf{x}_{K+1}}{\partial x, y, z}\right| = |\det R| = 1,$$

while the second factor is

$$\begin{aligned}\left|\det \frac{\partial x, y, z}{\partial \boldsymbol{\beta}, s}\right| &= \left|\det \begin{pmatrix} \frac{\partial \rho}{\partial \beta_1}\cos s & \frac{\partial \rho}{\partial \beta_2}\cos s & -\rho(\boldsymbol{\beta})\sin s \\ \frac{\partial \rho}{\partial \beta_1}\sin s & \frac{\partial \rho}{\partial \beta_2}\sin s & \rho(\boldsymbol{\beta})\cos s \\ \frac{1}{\sqrt{3}} & 0 & 0 \end{pmatrix}\right| \\ &= \frac{1}{\sqrt{3}}\rho(\boldsymbol{\beta})\frac{\partial \rho}{\partial \beta_2}.\end{aligned}$$

Their product is therefore $|\det \frac{\partial \mathbf{x}_{K+1}}{\partial \boldsymbol{\beta}, s}| = \frac{1}{\sqrt{3}}\rho(\boldsymbol{\beta})\frac{\partial \rho}{\partial \beta_2}$. Since this is independent of s, so is its inverse, the Jacobian $|\det \frac{\partial \boldsymbol{\beta}, s}{\partial \mathbf{x}_{K+1}}|$. It follows from Eq. (3) that we can achieve symmetric transition probabilities by choosing s uniformly on $[0, 2\pi)$.

Arguments similar to those for constrained exponential surrogates (see Appendix A 1) imply aperiodicity and full range.

3. Product

To generate constrained power-law surrogates we must randomize a sequence $\mathbf{x}$ while fixing the product $\prod_i x_i$. We consider two theoretically equivalent approaches, the first of which is more direct. Each iteration we randomize x_{i_1}, x_{i_2} while maintaining the value of $\beta_1 = x_{i_1}x_{i_2}$. We consider the change of coordinates

$$\begin{aligned}\beta_1 &= x_{i_1}x_{i_2}, \\ s &= x_{i_1},\end{aligned}$$

which corresponds to Jacobian

$$\left|\det \frac{\partial \boldsymbol{\beta}, s}{\partial \mathbf{x}_{K+1}}\right| = \left|\det \begin{pmatrix} x_{i_2} & x_{i_1} \\ 1 & 0 \end{pmatrix}\right| = x_{i_1} = s.$$

It follows from Eq. (3) that we can achieve symmetric transition probabilities by choosing $s \in [1, \beta_1]$ with probability density $P(s) = C/s$. A change of variables shows that this is equivalent to choosing $\log s$ uniformly on $[0, \log \beta_1]$, which suggests our second approach to generating constrained power-law surrogates.

This alternative method for generating constrained power-law surrogates allows us to utilize the existing solution for generating constrained exponential surrogates. Since fixing the product $\prod_i x_i$ is equivalent to fixing $\sum_i \log x_i$, we apply our constrained exponential surrogate algorithm to the log-transformed sequence $\mathbf{y} = (y_1, \ldots, y_N)$ where, for $i = 1, \ldots, N$, $y_i = \log x_i$, then exponentiate the resulting sequence. Our constrained power-law surrogate algorithm inherits the full range of the constrained exponential surrogate algorithm. To show uniform sampling we need to show that the probability $P_{\mathbf{X}}(\mathbf{x})$ of observing a sequence $\mathbf{x}$ after applying our constrained power-law surrogate algorithm satisfies

$$P_{\mathbf{X}}(\mathbf{x}) = P_{\mathbf{X}}(\mathbf{x}') \tag{A1}$$

for any pair of sequences $\mathbf{x}, \mathbf{x}'$ which have the same product. We already know that the analogous result

$$P_{\mathbf{Y}}(\mathbf{y}) = P_{\mathbf{Y}}(\mathbf{y}') \tag{A2}$$

holds for $\mathbf{y}, \mathbf{y}'$. For the change of variables $\mathbf{y} = \mathbf{f}(\mathbf{x})$ we have Jacobian $|\mathbf{f}'(\mathbf{x})| = 1/\prod_i x_i$, and so Eq. (A2) transforms to $P_{\mathbf{X}}(\mathbf{x})\prod_i x_i = P_{\mathbf{X}}(\mathbf{x}')\prod_i x_i'$. Equation (A1), which implies uniform sampling, now follows from preservation of the product $\prod_i x_i$.

From this argument we infer a slightly more general strategy. To produce a constrained surrogate for $x_1, \ldots, x_N$ while fixing $\beta_1(x_1, \ldots, x_N) = \prod_{i=1}^N x_i, \beta_2(x_1, \ldots, x_N), \ldots, \beta_K(x_1, \ldots, x_N)$, we can take a surrogate of its log-transform $\mathbf{y}$ which preserves $\beta_1(\exp y_1, \ldots, \exp y_N), \ldots, \beta_K(\exp y_1, \ldots, \exp y_N)$, then perform an inverse log transform on the result. The next example is of this type.

4. Product and second log-moment

To generate constrained lognormal surrogates we must randomize a sequence $\mathbf{x}$ while fixing the first and second log-moments $\sum_i \log x_i$ and $\sum_i (\log x_i)^2$. To accommodate these constraints, we apply our constrained truncated Gaussian surrogate algorithm to the log-transformed sequence $\mathbf{y}$,

then exponentiate the resulting sequence. Our constrained lognormal surrogate algorithm inherits from our constrained truncated Gaussian surrogate algorithm full range and, by preservation of the product $\prod_i x_i$, uniform sampling (for analogous reasoning, see Appendix A 3).

5. Maximum

To randomize the sequence $\mathbf{x}$ while retaining the maximum $\beta_1 = \max_{i\in\{1,\ldots,N\}} x_i$, we independently choose each element x_i uniformly at random from the interval $[x_{\min}, \beta_1)$. We then uniformly at randomly choose a single element x_i and set its value to preserve the originally observed maximum of the sequence, i.e., as $x_i = \beta_1$. Note that we have constrained the property which determines the likelihood of a uniform distribution without applying the majority of our framework.

APPENDIX B: NUMBER OF TRANSITIONS

Here we estimate the number of transitions required to approximate the final distribution. Our main finding is that for any $p \in [1, \infty]$ and tolerance $\epsilon > 0$, the relative error in expectation after r iterations will satisfy

$$\frac{\|\mathbb{E}[\mathbf{x}^{(r)}] - \mathbb{E}[\mathbf{x}^{(\infty)}]\|_p}{\|\mathbb{E}[\mathbf{x}^{(\infty)}]\|_p} < \epsilon \tag{B1}$$

as long as $r \geqslant \frac{N}{b}[\log_b N - \log_b(\epsilon)]$, where $b = 2$ ($b = 3$) for constrained exponential (truncated Gaussian) surrogates and, for constrained power-law (lognormal) surrogates, the same result holds for the continuous transformation $\log x^{(0)}{}_1, \ldots, \log x^{(0)}{}_N$ of the observed time series $\mathbf{x}^{(0)} = x^{(0)}{}_1, \ldots, x^{(0)}{}_N$. In this expression, $\mathbf{x}^{(r)} = x^{(r)}{}_1, \ldots, x^{(r)}{}_N$ represents the result of r iterations of a constrained surrogate algorithm, and $\mathbb{E}[\cdot]$ is the expectation over many independent trials. In particular, $\mathbb{E}[\mathbf{x}^{(0)}] = \mathbf{x}^{(0)}$, which is the originally observed time series, and $\mathbb{E}[\mathbf{x}^{(\infty)}] := \lim_{r\to\infty} \mathbb{E}[\mathbf{x}^{(r)}] = m, \ldots, m$ comprises N instances of the mean value m of the limiting expectation $\mathbb{E}[\mathbf{x}^{(\infty)}]$.

To demonstrate this, we first consider the exponential surrogate algorithm and the case $x_{\min} = 0$. We investigate the observed sequence $\mathbf{x}^{(0)} = Nm, 0, 0, \ldots, 0$, which represents the most extreme deviation from the limiting expectation $\mathbb{E}[\mathbf{x}^{(\infty)}]$, and assume that any other observed sequence will achieve condition (B1) at least as quickly. We also assume that for $r \geqslant 0$ and $t \geqslant 2$,

$$\mathbb{E}\big[x_1^{(r)}\big] \geqslant m \geqslant \mathbb{E}\big[x_t^{(r)}\big]. \tag{B2}$$

For an iteration of the constrained exponential surrogate algorithm which involves $x^{(r)}{}_1$ and $x^{(r)}{}_t$, we have

$$\mathbb{E}\big[x_1^{(r+1)}\big] = \frac{1}{2}\big(\mathbb{E}\big[x_1^{(r)}\big] + \mathbb{E}\big[x_t^{(r)}\big]\big) \leqslant \frac{1}{2}\mathbb{E}\big[x_1^{(r)}\big] + \frac{m}{2},$$

where the inequality follows from the right-hand inequality in relationship (B2). Hence the error $\mathbb{E}[x^{(r)}{}_1] - \mathbb{E}[x^{(\infty)}{}_1] = \mathbb{E}[x^{(r)}{}_1] - m$ in the expectation of $x^{(r)}{}_1$ satisfies

$$\big|\mathbb{E}\big[x_1^{(r+1)}\big] - m\big| \leqslant \tfrac{1}{2}\big|\mathbb{E}\big[x_1^{(r)}\big] - m\big|,$$

where the absolute value can be used because of the left-hand inequality in relationship (B2). This implies that the error in $\mathbb{E}[x^{(r)}{}_1]$ more than halves in each iteration which involves $x^{(r)}{}_t$. Since, on average, $x^{(r)}{}_t$ will be involved in 2 in N iterations, we can expect error to shrink as

$$\begin{aligned}\big|\mathbb{E}\big[x_1^{(r)}\big] - m\big| &\leqslant 2^{-2r/N}\big|\mathbb{E}\big[x_1^{(0)}\big] - m\big| \\ &= 2^{-2r/N}|Nm - m| \leqslant 2^{-2r/N}Nm.\end{aligned}$$

The same argument holds for other coordinates, so we have that for $t \geqslant 1$,

$$\big|\mathbb{E}\big[x_t^{(r)}\big] - m\big| \leqslant 2^{-2r/N}Nm.$$

The relative error in $\mathbb{E}[\mathbf{x}^{(r)}]$ must then satisfy

$$\frac{\|\mathbb{E}[\mathbf{x}^{(r)}] - \mathbb{E}[\mathbf{x}^{(\infty)}]\|_p}{\|\mathbb{E}[\mathbf{x}^{(\infty)}]\|_p} \leqslant \frac{N^{1/p}2^{-2r/N}Nm}{N^{1/p}m} = 2^{-2r/N}N.$$

This holds also for $x_{\min} > 0$, as does the tighter bound

$$\frac{\|\mathbb{E}[\mathbf{x}^{(r)}] - \mathbb{E}[\mathbf{x}^{(\infty)}]\|_p}{\|\mathbb{E}[\mathbf{x}^{(\infty)}]\|_p} \leqslant 2^{-2r/N}N\frac{m - x_{\min}}{m}.$$

Therefore, for constrained exponential surrogates, relative error less than ϵ is expected for a number of iterations $r \geqslant \frac{N}{2}(\log_2 N - \log_2 \epsilon)$.

For an iteration of the constrained truncated Gaussian surrogate algorithm which involves $x^{(r)}{}_1$, $x^{(r)}{}_t$ and $x^{(r)}{}_u$, we instead have

$$\mathbb{E}\big[x_1^{(r+1)}\big] = \tfrac{1}{3}\big(\mathbb{E}\big[x_1^{(r)}\big] + \mathbb{E}\big[x_t^{(r)}\big] + \mathbb{E}\big[x_u^{(r)}\big]\big)$$

and expect $x^{(r)}{}_t$ to be involved in 3 out of N iterations. The same arguments as for constrained exponential surrogates indicate that, for the constrained truncated Gaussian surrogate algorithm, we can achieve relative error less than ϵ for a number of iterations $r \geqslant \frac{N}{3}(\log_3 N - \log_3 \epsilon)$.

[1] U. Rüde, K. Willcox, L. C. McInnes, and H. D. Sterck, Research and education in computational science and engineering, SIAM Rev. **60**, 707 (2018).

[2] S. V. Kovalchuk, V. V. Krzhizhanovskaya, and P. M. A. Sloot, Twenty years of computational science, J. Comput. Sci. **46**, 101187 (2020).

[3] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. Doyne Farmer, Testing for nonlinearity in time series: The method of surrogate data, Physica D **58**, 77 (1992).

[4] J. Theiler and D. Prichard, Constrained-realization Monte-Carlo method for hypothesis testing, Physica D **94**, 221 (1996).

[5] G. Lancaster, D. Iatsenko, A. Pidde, V. Ticcinelli, and A. Stefanovska, Surrogate data for hypothesis testing of physical systems, Phys. Rep. **748**, 1 (2018).

[6] A. Clauset and R. Woodard, Estimating the historical and future probabilities of large terrorist events, Ann. Appl. Stat. **7**, 1838 (2013).

[7] B. Pasquini, P. Pedroni, and S. Sconfietti, First extraction of the scalar proton dynamical polarizabilities from real Compton scattering data, Phys. Rev. C **98**, 015204 (2018).

[8] A. Clauset, C. R. Shalizi, and M. E. J. Newman, Power-law distributions in empirical data, SIAM Rev. **51**, 661 (2009).

[9] A. Deluca and Á. Corral, Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions, Acta Geophys. **61**, 1351 (2013).

[10] Y. Malevergne, V. Pisarenko, and D. Sornette, Testing the Pareto against the lognormal distributions with the uniformly most powerful unbiased test applied to the distribution of cities, Phys. Rev. E **83**, 036111 (2011).

[11] A. Papana, E. Siggiridou, and D. Kugiumtzis, Detecting direct causality in multivariate time series: A comparative study, Commun. Nonlinear Sci. Numer. Simul. **99**, 105797 (2021).

[12] Y. Antonacci, C. Barà, L. Sparacino, I. Pirovano, A. Mastropietro, G. Rizzo, and L. Faes, Spectral information dynamics of cortical signals uncover the hierarchical organization of the human brain's motor network, IEEE Trans. Biomed. Eng. **72**, 1655 (2025).

[13] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, The backbone of the climate network, Europhys. Lett. **87**, 48007 (2009).

[14] S. V. Scarpino and G. Petri, On the predictability of infectious disease outbreaks, Nat. Commun. **10**, 898 (2019).

[15] X.-J. Zhang, J. M. Moore, T.-T. Gao, X. Zhang, and G. Yan, Brain-inspired wiring economics for artificial neural networks, PNAS Nexus **4**, pgae580 (2024).

[16] J. M. Moore, G. Yan, and E. G. Altmann, Nonparametric power-law surrogates, Phys. Rev. X **12**, 021056 (2022).

[17] D. Kugiumtzis, Test your surrogate data before you test for nonlinearity, Phys. Rev. E **60**, 2808 (1999).

[18] J. H. Lucio, R. Valdés, and L. R. Rodríguez, Improvements to surrogate data methods for nonstationary time series, Phys. Rev. E **85**, 056202 (2012).

[19] S. D. Pethel and D. W. Hahs, Exact significance test for Markov order, Physica D **269**, 42 (2014).

[20] E. Siggiridou, C. Koutlis, A. Tsimpiris, and D. Kugiumtzis, Evaluation of Granger causality measures for constructing networks from multivariate time series, Entropy **21**, 1080 (2019).

[21] D. C. Corrêa, J. M. Moore, T. Jüngling, and M. Small, Constrained Markov order surrogates, Physica D **406**, 132437 (2020).

[22] S. Stramaglia, T. Scagliarini, Y. Antonacci, and L. Faes, Local Granger causality, Phys. Rev. E **103**, L020102 (2021).

[23] R. D. Markello, J. Y. Hansen, Z.-Q. Liu, V. Bazinet, G. Shafiei, L. E. Suárez, N. Blostein, J. Seidlitz, S. Baillet, T. D. Satterthwaite, M. M. Chakravarty, A. Raznahan, and B. Misic, Neuromaps: Structural and functional interpretation of brain maps, Nat. Methods **19**, 1472 (2022).

[24] F. Váša and B. Mišić, Null models in network neuroscience, Nat. Rev. Neurosci. **23**, 493 (2022).

[25] E. Ceolini, K. R. Ridderinkhof, and A. Ghosh, Age-related behavioral resilience in smartphone touchscreen interaction dynamics, Proc. Natl. Acad. Sci. USA **121**, e2311865121 (2024).

[26] D. Li, A. Zalesky, Y. Wang, H. Wang, L. Ma, L. Cheng, T. Banaschewski, G. J. Barker, A. L. W. Bokde, R. Brühl, S. Desrivières, H. Flor, H. Garavan, P. Gowland, A. Grigis, A. Heinz, H. Lemaître, J.-L. Martinot, M.-L. P. Martinot, E. Artiges, *et al.*, Mapping the coupling between tract reachability and cortical geometry of the human brain, Nat. Commun. **16**, 7489 (2025).

[27] Z. Ma, C. Zeng, Y.-C. Zhang, and T. M. Bury, Predicting critical transitions with machine learning trained on surrogates of historical data, Commun. Phys. **8**, 258 (2025).

[28] M. E. J. Newman, Power laws, Pareto distributions and Zipf's law, Contemp. Phys. **46**, 323 (2005).

[29] A. D. Broido and A. Clauset, Scale-free networks are rare, Nat. Commun. **10**, 1017 (2019).

[30] Á. Corral, I. Serra, and R. Ferrer-i-Cancho, Distinct flavors of Zipf's law and its maximum likelihood fitting: Rank-size and size-distribution representations, Phys. Rev. E **102**, 052113 (2020).

[31] M. Serafino, G. Cimini, A. Maritan, A. Rinaldo, S. Suweis, J. R. Banavar, and G. Caldarelli, True scale-free networks hidden by finite size effects, Proc. Natl. Acad. Sci. USA **118**, e2013825118 (2021).

[32] E. G. Altmann, *Statistical Laws in Complex Systems* (Springer Nature, Cham, 2025).

[33] Á. Corral and Á. González, Power law size distributions in geoscience revisited, Earth Space Sci. **6**, 673 (2019).

[34] P. Bak, C. Tang, and K. Wiesenfeld, Self-organized criticality: An explanation of the $1/f$ noise, Phys. Rev. Lett. **59**, 381 (1987).

[35] D. Sornette, *Critical Phenomena in Natural Sciences: Chaos, Fractals, Self-Organization and Disorder: Concepts and Tools* (Springer Science & Business Media, Berlin, 2006).

[36] V. Frette, K. Christensen, A. Malthe-Sørenssen, J. Feder, T. Jøssang, and P. Meakin, Avalanche dynamics in a pile of rice, Nature (London) **379**, 49 (1996).

[37] Á. Corral, L. Telesca, and R. Lasaponara, Scaling and correlations in the dynamics of forest-fire occurrence, Phys. Rev. E **77**, 016101 (2008).

[38] G. Nicoletti, S. Suweis, and A. Maritan, Scaling and criticality in a phenomenological renormalization group, Phys. Rev. Res. **2**, 023144 (2020).

[39] G. Nicoletti, L. Saravia, F. Momo, A. Maritan, and S. Suweis, The emergence of scale-free fires in Australia, iScience **26**, 106181 (2023).

[40] S. Wang, J. Meng, S. Fang, T. Liu, K. Christensen, J. Kurths, and J. Fan, Self-organized criticality in atmospheric rivers, Phys. Rev. Lett. **136**, 094201 (2026).

[41] B. Gutenberg and C. F. Richter, Frequency of earthquakes in California, Bull. Seismol. Soc. Am. **34**, 185 (1944).

[42] E. G. Altmann, G. Cristadoro, and M. D. Esposti, On the origin of long-range correlations in texts, Proc. Natl. Acad. Sci. USA **109**, 11582 (2012).

[43] L. D. Landau and E. M. Lifshitz, *Statistical Physics, Part 1*, Course of Theoretical Physics Vol. 5 (Pergamon Press, Oxford, 1980).

[44] H. C. Berg, *Random Walks in Biology* (Princeton University Press, Princeton, 2025).

[45] M. Mitzenmacher, A brief history of generative models for power law and lognormal distributions, Internet Math. **1**, 226 (2004).

[46] B. Pearson and B. Fox-Kemper, Log-normal turbulence dissipation in global ocean models, Phys. Rev. Lett. **120**, 094501 (2018).

[47] A. Clauset, M. Young, and K. S. Gleditsch, On the frequency of severe terrorist events, J. Conflict Resolut. **51**, 58 (2007).

[48] M. J. van der Heyden, C. G. C. Diks, B. P. T. Hoekstra, and J. DeGoede, Testing the order of discrete Markov chains using surrogate data, Physica D **117**, 299 (1998).

[49] J. Theiler and D. Prichard, Using "surrogate surrogate data" to calibrate the actual rate of false positives in tests for nonlinearity in time series, in *Nonlinear Dynamics and Time Series: Building a Bridge Between the Natural and Statistical Sciences*, edited by C. D. Cutler and D. T. Kaplan, Fields Institute Communications Vol. 11 (American Mathematical Society, Providence, RI, 1997), p. 99.

[50] J. Besag and D. Mondal, Exact goodness-of-fit tests for Markov chains, Biometrics **69**, 488 (2013).

[51] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods* (Springer, New York, 2004).

[52] See Supplemental Material at http://link.aps.org/supplemental/10.1103/fs71-p5y8 for additional figures and an additional table supporting our results.

[53] B. D. Malamud, G. Morein, and D. L. Turcotte, Forest fires: An example of self-organized critical behavior, Science **281**, 1840 (1998).

[54] J. Lyu, M. I. Chishti, and Z. Peng, Marked distinctions in syntactic complexity: A case of second language university learners' and native speakers' syntactic constructions, Front. Psychol. **13**, 1048286 (2022).

[55] M. Ortuño, P. Carpena, P. Bernaola-Galván, E. Muñoz, and A. M Somoza, Keyword detection in natural languages and DNA, Europhys. Lett. **57**, 759 (2002).

[56] E. G. Altmann, J. B. Pierrehumbert, and A. E. Motter, Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words, PLoS ONE **4**, e7678 (2009).

[57] A. Corral, R. Ferrer-i Cancho, G. Boleda, and A. Diaz-Guilera, Universal complex structures in written language, arXiv:0901.2924.

[58] B. Piazza, D. L. Barabási, A. F. Castro, G. Menichetti, and A.-L. Barabási, Physical network constraints define the lognormal architecture of the brain's connectome, bioRxiv (2025).

[59] J. Voit, *The Statistical Mechanics of Financial Markets* (Springer, Berlin, 2005).

[60] J. M. Moore, Power-law, GitHub, 2026, https://github.com/JackMurdochMoore/surrogates-for-distributions.git.

[61] M. P. H. Stumpf and M. A. Porter, Critical truths about power laws, Science **335**, 665 (2012).

[62] P. Holme, Rare and everywhere: Perspectives on scale-free networks, Nat. Commun. **10**, 1016 (2019).

[63] M. Gerlach and E. G. Altmann, Testing statistical laws in complex systems, Phys. Rev. Lett. **122**, 168301 (2019).