Penalized KS method to fit data sets with power law distribution over a bounded subinterval

ARTICLE HISTORY
Compiled November 19, 2020

Word count: 15000

ABSTRACT
We develop a variation of a Kolmogorov-Smirnov (KS) method for estimating a power law region, including its lower and upper bounds, of the probability density in a set of data which can be modeled as a continuous random sample. Our main innovation is to stabilize the estimation of the bounds of the power law region by introducing an adaptive penalization term involving the logarithmic length of the interval when minimizing the Kolmogorov-Smirnov distance between the random sample and the power law fit over various candidate intervals. We show through simulation studies that an adaptively penalized Kolmogorov-Smirnov (apKS) method improves the estimation of the power law interval on random samples from various theoretical probability distributions. Variability in the estimation of the bounds can be further reduced when the apKS method is applied to subsamples of the original random sample, and the subsample estimates are averaged to yield a final estimate.

KEYWORDS
power law, truncated power law, bounded power law, power law interval, KS, MLE, validation, penalization

1. Introduction

A common question in analyzing data sets for a scalar variable is whether the distribution of the data is consistent with a power law tail [1–4]. Suppose we observe data $\{X_j\}_{j=1}^n$ which we assume to be independent and identically distributed with common cumulative distribution function (CDF) $F(x)$. In statistical language, then, we assume the data set can be modeled as a random sample. Then, the question of a "power law tail" is whether the unknown CDF $F(x)$ has a power law property for large values of x:

$$1 - F(x) = \Pr(X > x) = C_F x^{1-\alpha} \text{ for } x \geq x_*$$

with power law exponent $\alpha > 1$ and constants $C_F, x_* > 0$. Note that we are not defining the CDF of the random sample here, but rather describing a feature of the CDF for state space values above x_*. Alternatively, if we specialize to absolutely continuous random variables with density function $f(x)$, we are asking whether the density has a power law relationship for large values of x:

$$f(x) = Cx^{-\alpha} \text{ for } x \geq x_* \tag{1}$$

where the constant $C = C_F(1-\alpha)$. One reason power law tails are of interest is they indicate that large values are realized with substantially higher probabilities; some illustrations of this point can be found in [2]. A second reason power law tails are of interest is that they indicate self-similarity of the variable over some range of values, which could point to some simplified, low-parameter generative model as an adequate explanation of the distribution of the data values [5].

Fitting power laws to data by linear regression on log-log plots, often using the empirical tail probability as the response variable to reduce the noisiness in histograms [6], can give misleading conclusions [2,4,7,8], for statistical reasons spelled out in [9]. More principled methods have been proposed in the statistical literature for assessing whether a power law tail can be supported by the data in a random sample and estimating the parameters α, C, and x_* of the power law fit. The *KS method* described by [2] has found great popularity recently in the applied mathematics and scientific communities, presumably because of its relative simplicity, sound foundations on the asymptotic convergence properties of the underlying maximum likelihood estimation (MLE) of the power law exponent, and the subsequent validation step that checks whether the power law fit is statistically meaningful. Some concrete illustrations are provided in [10] for how a maximum likelihood approach produces higher quality estimates for the power law exponent than some other statistical processing approaches used in the applied science literature. The KS method optimizes the Kolmogorov-Smirnov (KS) distance between putative power law fits and the random sample, over proposed tail regions, to estimate the lower bound x_* of the tail region.

The problem of systematically identifying power law regimes bounded on both ends, the goal of this paper, has received more limited attention relative to the case of power law tails bounded below. A variety of applications have upper bounds on the power law scaling due to finite size effects [6,11] or limits of resolution of the measuring device [12,13]. As some examples, stock price fluctuations can have cutoffs due to automatic mechanisms for controlling extreme price changes [12] and aphid movement data appears to have an exponential tail over large times but a power law region at intermediate times [14]. Geophysical and astrophysical examples with limited power law regions are described in [1] and [15], respectively. Using power law tail fits to data sets without accounting for upper cutoffs have been found to lead to biased estimates of the power law exponent [13,15].

We discuss next a few available approaches to estimating power law scaling over a region bounded above and below. In [13], the maximum likelihood estimation of the power law exponents is extended to the case where the putative power law region is bounded both from above and below. The robustness of estimated power law exponents under various choices of upper and lower cutoffs is studied via maximum likelihood maps in [16], but neither this work nor the one in [13] directly address the issue of specifying which choice of these upper and lower bounds is most appropriate for defining the extent of the power law region. A direct application of MLE on a purely power law distribution which vanishes outside the power law region $[x_*, x^*]$, which is typically called an upper-truncated Pareto distribution in the literature but labeled EPL1 here (Appendix D.2, see also Figure 8), results in the estimation of the lower and upper bounds of the power law region to be the minimum and maximum value of the random sample, respectively [12,15,17]. This application of MLE was shown by [18] to lead to biased estimators of the bounds; they propose instead to consider statistics of the maximum datum value, together with the MLE exponent α, to ascertain whether the power law region has an upper bound, and if so, what the upper bound should be. Somewhat different bias-reducing estimators are proposed by [15] and [19] for the upper

bound of the upper-truncated Pareto distribution, and are shown through numerical examples to be successful in reducing bias in some cases, but not as successful in reducing variation. Some non-MLE estimators for the upper bound of the power law region are also compared in [15] and are shown to be less effective than the MLE-based estimators.

Typically, though, just as many probability models relevant in applications have power law tails without simply being a Pareto distribution, so do probability models with power law behavior over a finite region typically have non-power law components outside this region. Some applied works, for example [14], apply MLE to fully parameterized probability models with power law regions turning over to other specified (i.e., exponential) behavior, but here we wish to eschew hypothesizing some particular parametric model outside the power law region. On the other hand, purely non-parametric treatments of the probability distribution outside the power-law region become cumbersome when it is supported over a wide range of values [4]. Some studies use *ad hoc* procedures to set the bounds of the power law region. For example, in [20], a power law distribution was argued to be a good fit to a citation distribution in the interval $[85, 500]$ by discarding data beyond the value 500 on the grounds that the scant data in the tail contributes excessive fluctuations to the estimator. More general considerations are given in [12] and [9] for how to apply an MLE approach for a power law over a bounded subinterval within the support of a more general, unspecified probability distribution, when the bounds are not known a priori. They primarily focus on heuristic procedures for how to choose the lower bound x_*, for example by choosing some largest r data points to exclude a non-power-law core (below x_*), with the parameter r chosen based on visual inspection of the complementary CDF [12], or by choosing the bounds of the power law region by considering where the plot of the MLE power law exponent versus bound parameter remains flat, as well as using a χ^2 goodness-of-fit test for the quality of the power law fit over a proposed interval based on binning the data within that interval into a small number of subsets [9]. While giving reasonable success in estimating the lower bounds on some example distributions with non-power-law core, neither of these approaches were fully implemented on any examples for estimating the upper bound. More importantly, neither approach appears comparably systematic to the KS method of [2] for estimating the lower bound of a power law tail. Finally, visual inspection of the maximum likelihood maps can certainly give insight into the extent of a power law region [16], but we are unaware of a systematic procedure for producing a specific estimator for the upper and lower bounds of a power law region based on these maximum likelihood maps.

We therefore focus our efforts on extending the KS method to estimating both the lower and upper bounds of a power law interval, rather than a power law tail. While the KS method does have a fairly direct extension to this setting, we find by tests on various distributions that the estimators for the upper and lower bounds are highly variable. This variability in the estimated bounds does not seriously affect the quality of the estimator for the power law exponent $\hat{\alpha}$, which is generally the statistical quantity of primary interest. Consequently, the roughness of the estimates of the power law region may be of secondary concern, and possibly thought of as mere tuning parameters for finding the power law exponent. In particular, if a power law actually exists over a region $[x_*, \infty)$, and x_* is too conservatively estimated as $\hat{x}_* \geq x_*$ (with $\hat{x}_*$ being the estimator for x_*), the main cost is in the exclusion of relevant data points in the estimator of the exponent of the power law tail [21]. The fitted power law is still valid over the claimed region $[\hat{x}_*, \infty) \subseteq [x_*, \infty)$. Similarly, if the underlying distribution truly exhibits power law behavior over the bounded interval $[x_*, x^*]$, again the power

law can be validly claimed over any estimated interval satisfying $[\hat{x}_*, \hat{x}^*] \subseteq [x_*, x^*]$. One reason the KS method produces highly variable estimators $[\hat{x}_*, \hat{x}^*]$ for the bounds is precisely because the power law is indeed a good fit over any subset of the true power law region.

But this introduces problems in interpretation. For example, many consider the range of a power law fit as a key measure of its quality [3]. So a random sample for which the power law range of the distribution is significantly smaller than the true range may needlessly undermine the case for the meaningfulness of the power law fit. More acutely, numerical simulations and laboratory experiments may not have the resources to produce data which has a convincing power law fit over many decades. But, particularly in testing specific theoretical hypotheses suggesting a power law in the underlying probability distribution, one would like to infer from the data whether there is some evidence for a nascent power law regime, possibly only spanning somewhat more than a decade, rather than simply give up and say the data does not have sufficient quality to convincingly support a power law. Our concern with the direct extension of the KS method to a power law region is that the estimators might be highly variable and that they may be too conservative, needlessly restricting the inferential power of the algorithm.

To address these issues, we introduce a penalty term in the estimation scheme that yield estimators $\hat{x}_*$ and $\hat{x}^*$ that favor intervals $[\hat{x}_*, \hat{x}^*]$ that are wider, with larger values of $\hat{x}^*/\hat{x}_*$. The resulting optimization then trades off the quality of the fit, as measured by the KS distance, with the logarithmic width of the interval supporting the power law fit. The idea is that a less conservative, less variable, but still defensible power law inference procedure could be obtained by choosing the widest possible interval that nearly, but not necessarily exactly, minimizes the goodness-of-fit to a power law as measured by the KS distance. An obvious issue is the arbitrary choice of the penalty coefficient; we handle this through an adaptive procedure which makes the penalization as aggressive as possible while still producing a power law fit over an interval validated by the semiparametric bootstrap step. We call the resulting method the adaptively-penalized KS (apKS) method.

We remark that a similar variation of the KS method has been briefly described in [22] and studied in [1]. In their method, the interval with the maximum number of data points (or the largest log-range) that is validated by a parametric bootstrap is reported as the bounded (truncated) power law fit. A related idea, with a different validation test, is presented in [23]. As we discuss in Appendix C, our approach will fall in between a direct extension of the KS method of [2] and the method of [1,22], with considerable computational cost savings relative to the exhaustive optimization approach of the latter method.

We begin in Section 2 by revisiting the case in which a power law is sought over a tail region $x \geq x_*$, with a view toward analyzing the selection of the value for the (single) bound $\hat{x}_*$ under the original KS method of [2]. We then motivate and develop our proposed apKS approach for estimating x_* in Section 3. While this does lead to an improvement in the estimation of the lower bound for distributions with power law tails, as noted above, the real motivation of our work is to improve the handling of the upper bound of a distribution with a bounded power law region. The reason we spend a great deal of space on the power law tail case is to illustrate the technical issue and proposed resolution in this simpler setting. The same apKS approach is then extended directly to the inference of a power law regime over a bounded interval $[x_*, x^*]$ in Section 4, where we also present a direct extension of the KS method of [2] to simultaneously select lower and upper bound estimators $\hat{x}_*$ and $\hat{x}^*$. In Section 5,

we compare the performance of the KS and apKS algorithms on a variety of simulated random samples from probability distributions with exact power law regions bounded on one or both sides (described in Appendix D).

Generally speaking, we find that the apKS method seems to give significant improvements for distributions which have an abrupt, non-smooth transition between the power law region and non-power law region of the probability distribution. For distributions that smoothly interpolate between a power law region and a non-power law region, we find that the penalization approach does not improve as significantly on the original algorithm. Further statistical analysis reveals that, for these distributions, the variability of the estimators for the bounds of the power law region is not substantially reduced (for either the KS or apKS method) as the random sample sizes increase from 10^4 to 10^6 data points.

We continue the analysis of the statistical methods in Section 6 with particular attention to this apparent lack of improvement in the estimators as the size of the random sample increases. We argue that this appears to be a broad problem for power-law fitting algorithms based on the KS distance and propose a "batching" remedy to use the additional data to improve the statistical quality of the estimators for the power law bounds. The conclusions from our numerical experiments are discussed in Section 7. MATLAB codes together with a manual for using the apKS code can be found at `https://github.com/folmez/apKS-matlab`.

2. Power Law Tail Fit

To illustrate the technical concerns and develop our proposed algorithm in the simplest setting, we will first address the inference of a power law region over a semi-infinite tail region $[x_*, \infty)$ for a probability distribution underlying a random sample. An example of a PDF with a strict power law region with exponent $\alpha = 1.5$ and $x_* = 10$ is

$$f(x) = \begin{cases} Ae^{-\beta x} & : 0 \leq x < x_* \\ Cx^{-\alpha} & : x_* \leq x, \end{cases} \tag{2}$$

illustrated in Figure 1, with constants A, C, and β chosen so $\int_0^\infty f(x)dx = 1$ and $f(x)$ is continuously differentiable.

After specifying some basic terminology in Subsection 2.1, we recapitulate in Subsection 2.2 the KS method of [2] in its original formulation. We show in Subsection 2.3 that the estimator for the lower bound x_* has a large standard error for a simple family of distributions with power law tails, which raises a central issue we wish to address in later sections.

2.1. Setup

For a random sample $X_{1:n} \equiv \{X_i\}_{i=1}^n$ we can construct the *empirical CDF*

$$\hat{F}(x) \equiv \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}},$$

the fraction of data values not exceeding x, which is the CDF for a probability model that selects uniformly from the data points. The *Kolmogorov-Smirnov distance* (or

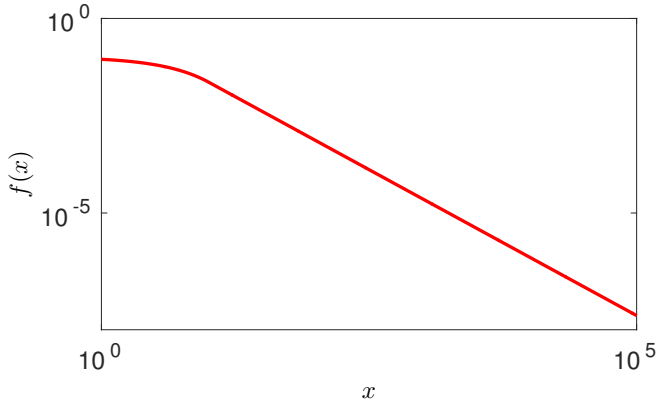

Figure 1. Log-log plot of probability density function given by Eq. (2) with a strict power law tail for $x \geq x_* = 10$ with exponent $\alpha = 1.5$.

statistic) is a widely used measure for the discrepancy of a random sample from a theoretical distribution. It is defined as the greatest distance between the CDF F for the theoretical probability distribution and the empirical CDF $\hat{F}$ [24, Sec. III.1]:

$$\rho_{\mathrm{KS}} = \sup_{y \in \mathbb{R}} |\hat{F}(y) - F(y)| \tag{3}$$

We specialize the KS distance for the purpose of assessing the quality of a power law fit to the tail of the probability distribution of a random sample by comparing the data and theoretical power law model only over a semi-infinite interval $[x_*, \infty)$

$$\rho_{\mathrm{KS}}(x_*) = \sup_{y \geq x_*} |\hat{F}^{x_*}(y) - F^{x_*, \hat{\alpha}(x_*)}(y)| \tag{4}$$

where $\hat{\alpha}(x_*)$ is the estimated exponent of the power law fit over the semi-infinite interval $[x_*, \infty)$, and $\hat{F}^{x_*}(\cdot)$, $F^{x_*, \hat{\alpha}(x_*)}(\cdot)$ are cumulative distribution functions for the restrictions, respectively, of the random sample and the theoretical power law fit, to the semi-infinite interval $[x_*, \infty)$. That is,

$$\hat{F}^{x_*}(y) = \begin{cases} 0 & \text{for } y < x_*, \\ \frac{\hat{F}(y) - \hat{F}(x_*)}{1 - \hat{F}(x_*)} & \text{for } y \geq x_* \end{cases}$$

and

$$F^{x_*, \alpha}(y) = \begin{cases} 0 & \text{for } y < x_*, \\ 1 - (y/x_*)^{1-\alpha} & \text{for } y \geq x_*. \end{cases}$$

2.2. KS method and validation for power law tail fitting

Here we briefly summarize the three-step procedure from [2] for fitting and validating a theoretical power law tail (1) to a given random sample. The first step is to fit a power law tail to the sample, inferring an estimated power law exponent $\hat{\alpha}$ and lower bound $\hat{x}_*$. Following [2], we call the fitting procedure the "KS method" since it is based

on minimizing the Kolmogorov-Smirnov distance. The second step is to validate the power law tail fit through a semiparametric bootstrap. The third step is to compare the power law fit to fits to alternative distributions. The purpose of the present work is to propose a variation for the first two steps, so we will leave the third step out of our subsequent considerations, since it could be employed in the same way for our modified method as described for the original method in [2].

2.2.1. Step 1: Fitting (KS method)

The first step in the estimation is similar to a profile likelihood method in considering the two parameters α and x_* sequentially. For a hypothesized value of the lower bound x_* of the power law tail, the power law scaling parameter is estimated through the following simple MLE formula applied to the portion of the random sample above this lower bound:

$$\hat{\alpha}(x_*) = 1 + N(x_*) \left[\sum_{i=1}^{n} \log \left(\frac{X_i}{x_*} \right) I_{\{X_i \geq x_*\}} \right]^{-1}, \tag{5}$$

where

$$N(x_*) = \sum_{i=1}^{n} I_{\{X_i \geq x_*\}}$$

denotes the number of data points beyond the hypothetical lower bound. The lower bound in the power law tail model is selected by minimizing the KS distance in (4) over all hypothetical lower bounds chosen from the data points:

$$\hat{x}_* = \underset{x_* \in X_{1:n}}{\arg\min}\ \rho_{\mathrm{KS}}(x_*). \tag{6}$$

This fitting procedure described above will be referred to as the *KS method*. The algorithm for the KS method is as follows:

Input: Random sample: $X_{1:n}$
Loop: For every $k = 1, 2, \ldots, n$:
(1) Compute the exponent $\hat{\alpha}(X_k)$ of the power law fit with lower bound X_k using (5).
(2) Compute the KS distance between the power law fit, with parameters $\hat{\alpha}(X_k)$ and X_k, and the random sample using the KS distance formula (4).
Output: The estimated power law exponent $\hat{\alpha}$ of the power law fit with corresponding lower bound $\hat{x}_*$, selected by the minimum KS distance among all n fits.

The number of operations performed at each step of the loop in this algorithm is $O(n)$, and hence, the total number of operations performed for the KS method is $O(n^2)$. When the random sample size is $n = 10^3$, execution of this algorithm takes about 20 milliseconds on a Macbook Pro with a 2.9 GHz Intel Core i7 processor.

2.2.2. Step 2: Validation

In order to validate a power law fit obtained by the KS method, semiparametric bootstrap samples are generated and a power law is fitted to each of these samples using the exact fitting procedure described above. Each of these power law fits has a KS distance value, and one computes the fraction of these KS distance values that are greater than the KS distance value of the fit from the first step. This fraction is the p-value estimated with a null hypothesis of the purported power law interval; we declare a failure to reject if $p > 0.1$. We will say that such a failure to reject will "validate" the presence of a power law interval. As discussed in [2], this should not be confused with the more typical use of p-values to reject a null hypothesis, where a positive result corresponds to a low p-value which makes the null hypothesis less tenable and therefore rejected at some level of confidence [25]. This is a standard application of hypothesis testing in goodness-of-fit tests, especially the classical Kolmogorov-Smirnov test of normality, even if it does not conform to the classical theory of testing. We generate 2500 bootstrap samples in order to estimate the p-value with an absolute error of 0.01. Our parameter choices for validation follow those of [2].

The validation of the power law fits is, due to the generation and fitting of many bootstrap samples, considerably more costly than the fitting step. With a sample size $n = 10^3$ and 2500 semiparametric bootstrap samples, the p-value is estimated in somewhat less than a minute on a Macbook Pro with 2.9 GHz Intel Core i7 processor.

2.2.3. Step 3: Model comparison

The *likelihood ratio test* (appendix C of [2]) is used to compare the power law distribution to alternative distributions that may be a better fit to the given data.

2.3. Quality of estimation of x_* with the KS method

In a forthcoming paper (**citation suppressed for anonymous version**), we define what it means for a probability distribution to have an *asymptotic* power law tail. In that case, the lower bound of the power law tail does not itself have a precise meaning, so its algorithmic selection has an obvious potential to be problematic. Here we show that, in fact, the selection of $\hat{x}_*$ by the KS method is problematic even for probability distributions that have an exact power law tail in the more restricted sense of Eq. (1).

We take the simple family of models with an exact power law tail defined by Eq. (2) and generate 100 simulated random samples from the models with exponents $\alpha = 1.25$, $\alpha = 1.5$, and $\alpha = 2.5$, with lower bound $x_* = 1$ in each case. Figure 2 reports the empirical CDF of the selections $\hat{x}_*$ obtained from the KS method applied to each of these random samples. We see moderate variability in the estimator $\hat{x}_*$ for $\alpha = 2.5$, and more severe variability for power law tails with $\alpha = 1.25$ and $\alpha = 1.5$. Not surprisingly, the variability is mostly in the direction of overestimation, for an $x_* > x_*^0$ where x_*^0 is the true value of the parameter, still gives a valid power law fit. This is not true for $x_* < x_*^0$. Nonetheless, for reasons discussed in Section 1, a less variable estimator would be preferable particularly when the KS method is extended to power laws on bounded intervals. One might object that the exponents $\alpha = 1.25$ and $\alpha = 1.5$ giving rise to the high variability of $\hat{x}_*$ correspond to unusual probability distributions with such slowly decaying tails as to have an infinite mean. In fact, [2] reviews several examples such as solar flare intensity, religious followers, intensity of wars, and word use count, whose probability distribution exhibit an apparent power law tail with exponents $1 < \alpha < 2$

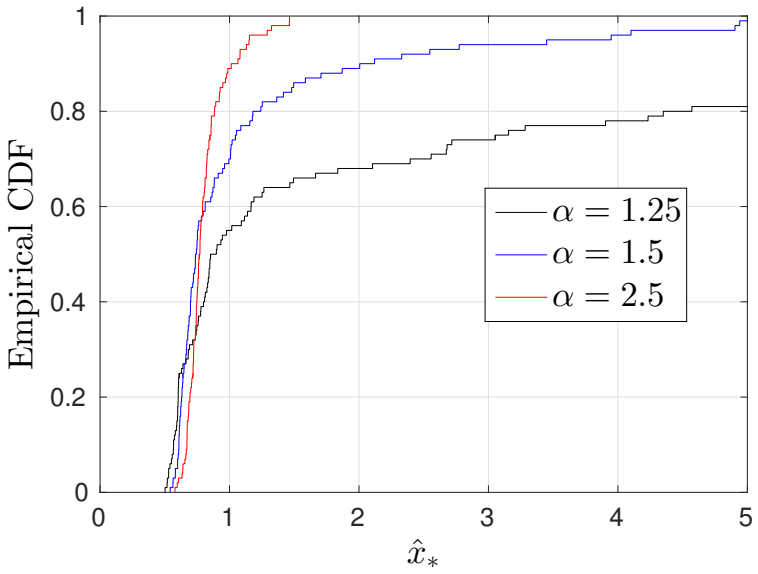

Figure 2. Empirical CDF of $\hat{x}_*$ as estimated by the KS method for 100 random samples from the probability distributions (2) with exponents $\alpha = 1.25$, 1.5 and 2.5. In each case, the true value of x_* is $x_*^0 = 1$.

and therefore an infinite mean.

Therefore, we pause to examine why the estimator $\hat{x}_*$ produced by the KS method has such a large variability even on such an idealized probability model as (2). We plot in Figure 3 the KS distance function $\rho_{\mathrm{KS}}(x_*)$ as a function of possible lower bound x_* for a representative random sample for each of the three parameter choices considered in Figure 2. These clearly reveal a problem with the estimator $\hat{x}_*$ arising from the KS method, which yields $\hat{x}_*$ as the value of x_* which minimizes $\rho_{\mathrm{KS}}(x_*)$. For $\alpha = 1.5$ and $\alpha = 1.25$ we see a very flat region for $x_* \geq 1$, with apparently noisy fluctuations. The KS distance rises rapidly as x decreases below the true value $x_*^0 = 1$ because the quality of the power law fit deteriorates as data from the non-power law region is incorporated. But the KS distance appears relatively indifferent to restricting the sample to shrinking subintervals (x_*, ∞) as x_* increases beyond the true value x_*^0. One might have imagined that the KS distance should increase with x_* because the decreasing number of data points should make the empirical CDF appear "rougher" and therefore a less good fit to the smooth power law, and this seems to be borne out for $\alpha = 2.5$, but not in any noticeable way for $\alpha = 1.25$ or $\alpha = 1.5$. We can now envision that the functions $\rho_{\mathrm{KS}}(x_*)$ produced for each random sample from $\alpha = 1.25$ or $\alpha = 1.5$ will look roughly similar, but have different "noise" in the flat region, so that the estimator $\hat{x}_*$ could be thought of as essentially random choice of values in the flat region of $\rho_{\mathrm{KS}}(x_*)$. This explains the variability seen in Figure 2. An obvious remedy to this unnecessary fluctuation in $\hat{x}_*$ produced by the KS method is, rather than to prescribe the estimator $\hat{x}_*$ to minimize $\rho_{\mathrm{KS}}(x_*)$, to choose the smallest value of x_* for which $\rho_{\mathrm{KS}}(x_*)$ is in some sense close to its minimum value. That is, we'd like the estimator $\hat{x}_*$ to be at the left edge of the flat region of $\rho_{\mathrm{KS}}(x_*)$ to reduce its variability; we see from Figure 3 that this would appear to give a quite good estimate of the true value of x_*^0. This is similar to the selection of many types of tuning parameters, such as the selection of the tuning parameter in ridge regression; one often stops at a value when the marginal improvement is trivial. In the next section, we implement this idea in an effort to improve the behavior of the lower bound estimator $\hat{x}_*$.

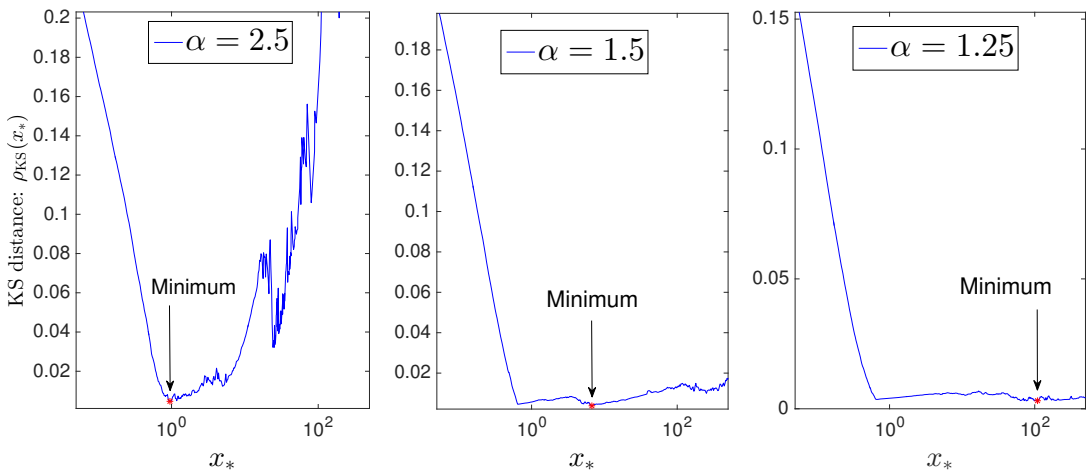

Figure 3. Estimation of the lower bound parameter of the power law fit using the KS distance metric for random samples from the probability distributions given by Eq (2) with exponents $\alpha = 1.25$, 1.5 and 2.5 for true lower bound value $x_*^0 = 1$. The horizontal axis represents the lower bound candidates x_*.

3. A modified estimator for the lower bound x_* of the power law tail

We can conceptually summarize our findings from Subsection 2.3 concerning the original KS method estimator $\hat{x}_*$ in terms of the following tension: Excessively small choices of x_* will lead to a large KS distance between the data on $[x_*, \infty)$ and a power law model because data is included from a region beyond which the power law fit should be valid. On the other hand, excessively large choices of x_* will exclude too much data and lead to a large KS distance to the power law fit because of the inferior estimation of the power law model parameters from the reduced data. Put another way, if we think for the moment of x_* as a tuning parameter for estimating the power law model fit, then decreasing x_* too far increases bias in the estimation of the power law fit parameters from the data, while increasing x_* too far increases the variation in the estimation of the power law fit parameters. Our results from Subsection 2.3 appear to show that encoding this bias-variation tradeoff through the KS distance gives rise to a strong resistance against choosing excessively small values of x_*, but a weak resistance to choosing excessively large values of x_*, and this latter fact gives rise to the high variability of the estimator $\hat{x}_*$. Our proposed resolution is to develop a new estimator that will give a boost to the resistance against large values of x_* by introducing a penalty term to the KS distance used to compute $\hat{x}_*$. We will introduce an explicit tuning parameter d as the coefficient of a penalty term; it plays essentially the role of setting the weight of variation relative to bias in estimating x_*. The objective is to try to arrive at an estimator $\hat{x}_*$ which tends to the lower end of the range of possible bounding values which give a small KS distance to the data. That is, we seek to devise an estimator $\hat{x}_*$ so that the power law tail is, ideally, claimed over the maximally defensible semi-infinite interval $[\hat{x}_*, \infty)$. While it can be helpful to think of x_* as simply a tuning parameter in order to understand the role of the penalty coefficient d, we actually do care about the value of x_*. The coefficient d of the penalty term, however, has no transparent meaning, so in a sense we are transferring the tuning aspect from the parameter of interest x_* to the parameter d.

We describe the proposed penalized optimization approach in Subsection 3.1, then turn to the question of how to choose the strength of the penalty term in Subsection 3.2.

We comment at the end of Subsection 3.2 how the validation by semiparameteric bootstrap should be implemented under our modified approach. The resulting *adaptively* penalized KS method is then presented as an algorithmic sketch in Subsection 3.3.

3.1. Penalized KS distance

The obvious way to fix an overestimated lower bound is to push it back where it should be by gently penalizing the KS distance values with an increasing penalty function of x. This can be done by adding a logarithmic penalty function to the KS distance

$$\rho_{\mathrm{pKS}}(x_*) = \rho_{\mathrm{KS}}(x_*) + d \log\left(\frac{x_*}{x_c}\right), \tag{7}$$

where $d > 0$ is a fixed penalty coefficient and x_c is chosen as the smallest data point in a given random sample, simply to ensure that the penalty term is positive. Of course x_c plays no role in the optimization since it simply offsets ρ_{pKS} by an additive constant. Notice that the penalized KS distance with the penalty coefficient $d = 0$ is the same as the regular KS distance. The benefit of using this metric is that it addresses the problem of lower bound overestimation because it favors smaller candidates. The reason we adopt a logarithmic structure to the penalty term is to reflect the presumed scale-invariance of the random sample in the power law region. The logarithmic structure will be carried over in Subsection 4.3 to estimating upper and lower bounds of a bounded power law region, where it has a further motivation. The width of a power law region is often expressed in terms of decades, or more generally, as the log-ratio of the upper and lower cutoffs. The logarithmic penalty term most naturally expresses a preference to increase this log-ratio width of the power law region, when consistent with the data.

Suppose for now an appropriate penalty coefficient d is known. Then the exact same procedure, which is described for the KS method in Section 2.2.1, is used except that, now, the KS distance in equation (6) is replaced with the penalized KS distance described from equation (7). This method will be referred to as the *penalized KS (pKS) method* (with penalty coefficient d).

3.2. Tuning of penalty coefficient

Suppose the penalized KS method is used with an arbitrarily chosen penalty coefficient d. If this penalty coefficient d is chosen to be too large, then there is a risk of pushing $\hat{x}_*$ too far to the left, incorporating data that is not really in the power law tail and producing a bad power law fit, which will likely be rejected in the validation step. If d is chosen to be too small, then the penalty term may be ineffective in improving the estimation of x_* relative to the KS method. Therefore, an adaptive algorithm is proposed which pushes $\hat{x}_*$ using the penalized KS distance as much as possible while making sure the resulting power law fit stays validated. Put another way, the penalty term is chosen as large as possible without producing unacceptable bias in $\hat{x}_*$ and the resulting power law fit, as measured by the validation step. So, bias and variation in the estimation of x_* is balanced is by trying to effectively minimize the variation in the construction of $\hat{x}_*$, subject to the constraint of incurring sufficiently small bias that the validation step is still satisfied. The algorithm selects d in a way that is similar to finding the zero of a function by bisection. In the algorithm, a very small $d = d_*$ is

initially chosen so that it produces a power law fit still accepted by the validation step and a very large $d = d^*$ is chosen so that it produces a power law fit rejected by the validation step. It is seen in the simulations that the choices $d_* = 10^{-10}$ and $d^* = 1$ are sufficient to achieve these results.

Next, a power law fit is obtained by the penalized KS method with a penalty coefficient that is equal to the geometric mean of these two coefficients which is $\sqrt{d_* d^*}$. If the resulting fit is accepted by the validation step, d_* is replaced with $\sqrt{d_* d^*}$, and if it is rejected, then d^* is replaced with $\sqrt{d_* d^*}$. This procedure is iterated until d_* and d^* become very close, and in the end d_* is taken to be the appropriate penalty coefficient for the random sample and denoted as $\hat{d}$.

Notice that the resulting power law fit is obtained by the penalized KS method with the penalty coefficient d_* and therefore has already been validated. We refer to this procedure as the *adaptively penalized KS (apKS)* method.

In the above procedure, the question arises whether the validation of a power law fit obtained by the penalized KS method via the semiparametric bootstrap should be based on the penalized or unpenalized KS distance. We choose to follow exactly the same validation procedure as in the original work [2], that is, using the unpenalized KS distance (4) to measure the distances between both the original random sample and the bootstrapped samples to the power law fits over the proposed semi-infinite intervals $[\hat{x}_*, \infty)$. Not only does this give a more conservative p-value (since the power law fit optimizes the penalized KS distance rather than the unpenalized KS distance), but the alternative approach would seem to require that a different adapted penalized KS distance should be used to measure the distance between each bootstrapped sample and its corresponding power law fit by the apKS method. Following [2], we reject the power law fit if the p-value obtained from the semiparametric bootstrap is less than 0.1, and otherwise accept it as validated.

3.3. Algorithm for adaptively penalized KS method

Input: Random sample $X_{1:n}$
KS Estimate: Apply the KS method and obtain a power law fit with lower bound $\hat{x}_*$ and exponent $\hat{\alpha}$. Continue only if the power law fit is validated by a semiparametric bootstrap, else terminate.
Initialize Penalization: Try $d = d_* = 10^{-10}$ and check the penalized KS method produces a validated power law fit, and try $d = d^* = 1$ and check the penalized KS method produces a power law fit which is rejected by the validation step. (These checks have never failed in in our simulations, but the algorithm terminates with an error if they would).
Loop:

(1) If $\frac{|d^* - d_*|}{d_*} < \epsilon_d = 10^{-2}$, then break out of the loop. Otherwise, continue.
(2) Take $d = \sqrt{d_* d^*}$.
(3) Fit a power law to $X_{1:n}$ by using penalized KS method with penalty coefficient d.
(4) If the resulting power law fit is validated by semiparametric bootstrap, update $\hat{x}_*, \hat{\alpha}$ and set $d_* := d$; otherwise set $d^* := d$.
(5) Go to step 1.

Output: A validated power law fit with $\hat{x}_*$ and $\hat{\alpha}$ and an empirically determined $\hat{d}$.

Table 1. Sample run of the apKS method on a random sample of size $n = 10^4$ that is power law distributed according to the density given in equation (2) with power law exponent $\alpha = 1.5$, lower bound $x_* = 1$, and initial penalty coefficients $d_* = 10^{-10}$ and $d^* = 1$. The relative tolerance $\epsilon_d = 10^{-2}$. The number of iterations required is 14.

d	$\hat{\alpha}$	$\hat{x}_*$	$\rho_{\mathrm{KS}}(\hat{x}_*)$	p-value	Validation
10^{-10}	1.54	5.23	0.0259	0.83	Accepted
1.00000	1.15	0.00	0.3876	0.00	Rejected
0.00001	1.54	5.23	0.0259	0.83	Accepted
0.00316	1.51	0.55	0.0266	0.28	Accepted
0.05623	1.50	0.48	0.0271	0.14	Accepted
0.23714	1.15	0.00	0.3876	0.00	Rejected
0.11548	1.15	0.00	0.3876	0.00	Rejected
0.08058	1.15	0.00	0.3876	0.00	Rejected
0.06732	1.15	0.00	0.3876	0.00	Rejected
0.06153	1.15	0.00	0.3876	0.00	Rejected
0.05882	1.50	0.48	0.0271	0.14	Accepted
0.06016	1.50	0.48	0.0271	0.14	Accepted
0.06084	1.15	0.00	0.3876	0.00	Rejected
0.06050	1.50	0.48	0.0271	0.14	Accepted

The computer time required for the above algorithm is less than or equal to the product of the number of iterations and the computer time required for the KS method and the validation procedure combined. The number of iterations required for the apKS method depends only on the choice of d_*, d^* and ϵ_d. To be precise, the number of iterations is $n_{\mathrm{iter}} + 2$ (including the two steps in the initiation) where n_{iter} is the smallest number such that $|(d^*/d_*)^{1/2^{n_{\mathrm{iter}}}} - 1| < \epsilon_d$, which can be well approximated for small ϵ_d by

$$n_{\mathrm{iter}} = \left\lceil \frac{\left[\ln \epsilon_d^{-1} + \ln \ln(d^*/d_*)\right]}{\ln 2} \right\rceil,$$

where d_* and d^* here refer to the initial choices of these parameters. For our recommended initial values $d_* = 10^{-10}$ and $d^* = 1$, $n_{\mathrm{iter}} = 14$ iterations are required in the apKS method.

A sample run of the algorithm for the apKS method is outlined in the Table 1. As seen in the table, the power law fit produced by the algorithm appears quite insensitive to the penalty coefficients d tried by the adaptive algorithm, whenever those values are sufficiently small. Also notice in the table that although there are 14 different choices of the penalty coefficient tested, the number of distinct power law fits obtained is only four. Therefore, this particular run of the above algorithm took only four times the computer time required to run the KS method and the validation procedure combined, since repeating power law fits need not be validated more than once.

4. Power Law Fit over Bounded Interval

We turn now to our primary focus, the search for a power law fit over an interval bounded both above and below. We begin in Subsection 4.1 with some mathematical discussion of probability distributions that can be said to possess a power law over a bounded interval. In Subsection 4.2, we present a direct extension of the KS method to finding bounded intervals of power law behavior in probability distributions underlying a finite random sample.

In Subsection 4.3, we carry over the ideas behind the adaptively penalized KS

method for fitting power law tails from Section 3 to the fitting of power law regions over bounded intervals.

4.1. Meaning of Power Law over Bounded Interval

We say that a probability distribution has a power law over a bounded interval $I = [x_*, x^*]$ when its CDF F satisfies the following property, for some real constants $C > 0$ and α:

$$F(x') - F(x) = \frac{C}{\alpha - 1}(x^{1-\alpha} - x'^{1-\alpha}) \text{ for } 0 < x_* < x \leq x' < x^*,$$

with the standard logarithmic modification $F(x') - F(x) = C \ln(x'/x)$ for the case when $\alpha = 1$. For continuous random variables, this criterion can be expressed more directly in terms of the probability density function:

$$f(x) = Cx^{-\alpha} \text{ for } x_* < x < x^*. \tag{8}$$

Contrary to power law tails discussed in Section 1, the condition $\alpha > 1$ for the exponent can be relaxed when the power law region is bounded from above and below.

4.2. KS method for bounded power law fitting

4.2.1. Estimating the exponent α of a bounded power law fit on interval I

Following the approach of [2], we estimate the exponent α of a putative power law model fit over a proposed interval $[x_*, x^*]$ by only referring to the portion of data that falls within this interval. The probability density for a power law distribution over this interval has the form $f_{x_*,x^*,\alpha}(x) = C_\alpha x^{-\alpha}$ on $[x_*, x^*]$ where $C_\alpha = \frac{-\alpha+1}{(x^*)^{-\alpha+1}-(x_*)^{-\alpha+1}}$. As before, we proceed with a profile likelihood function for α on a given proposed interval $[x_*, x^*]$, which is

$$\mathcal{L}_X(\alpha) = \prod_{i=1}^{N} C_\alpha \exp\left(-\alpha \ln X_i I_{\{X_i \in [x_*, x^*]\}}\right).$$

The structure of the likelihood function does not yield an explicit expression for the value of α which maximizes $\mathcal{L}$, so $\mathcal{L}$ is optimized numerically over a set of values in $[0.1, 3.5]$ with two significant digits. The maximizing value is then the MLE estimate for the power law exponent, written $\hat{\alpha}(x_*, x^*)$.

4.2.2. Interval estimation for bounded power law fitting

The KS method for unbounded power law fitting can easily be generalized to estimate a bounded power law interval as well. The *KS distance* for a bounded power law fit to the random sample $X_{1:n}$ on the interval $[x_*, x^*]$ is defined as

$$\rho_{\mathrm{KS}}(x_*, x^*) := \sup_{y \in [x_*, x^*]} |\hat{F}^{x_*,x^*}(y) - F^{x_*,x^*,\hat{\alpha}(x_*,x^*)}(y)|, \tag{9}$$

where $\hat{\alpha}(x_*, x^*)$ is the estimated exponent of the power law fit to the data on interval $[x_*, x^*]$. In this equation, $\hat{F}^{x_*,x^*}(\cdot)$ is the empirical CDF for the data values in the interval $[x_*, x^*]$ while $F^{x_*,x^*,\hat{\alpha}(x_*,x^*)}$ is the theoretical CDF of a power law distribution on the interval $[x_*, x^*]$ with the exponent $\hat{\alpha}(x_*, x^*)$ estimated from data.

The KS distance described in equation (9) is minimized over all possible intervals in order to find a bounded power law interval:

$$(\hat{x}_*, \hat{x}^*) = \underset{x_*, x^* \in X_{1:n}}{\arg\min}\, \rho_{\mathrm{KS}}(x_*, x^*). \tag{10}$$

A first concern is to make sure that a bounded power law fit is being searched on a plausibly long interval. Therefore, smaller intervals should be excluded from the set of candidate intervals. The minimization can be carried out over $\{x_*, x^* \in X_{1:n} : \frac{x^*}{x_*} > l\}$ where l is a reasonable minimum length for the bounded power law interval. We set $l = 10$ in this work, looking for power-laws that extend at least a decade. Equation (10) now becomes

$$(\hat{x}_*, \hat{x}^*) = \underset{x_*, x^* \in X_{1:n} : \frac{x^*}{x_*} > l}{\arg\min}\, \rho_{\mathrm{KS}}(x_*, x^*). \tag{11}$$

A second concern is to make sure that the proposed optimization is computationally feasible. In (11), the KS distance for bounded power law fits is minimized over $O(n^2)$ intervals, where n is the size of the random sample. The number of operations required to compute a single KS distance value is $O(n)$. Therefore, it is not practical to compute a bounded power law fit interval using (11) when $n > 1000$. Fortunately, one can obtain sufficiently accurate results by minimizing over a smaller subset of $X_{1:n}$ with sufficiently many points in every decade of $[\min(X_{1:n}), \max(X_{1:n})]$, which will grow more modestly with respect to the size of the random sample. This can be viewed as a subsample of $X_{1:n}$ that will be denoted by $L_m(X_{1:n})$. The next subsection will give more detail on these sets and on how to choose a suitable m. Replacing $X_{1:n}$ with the smaller subset $L_m(X_{1:n})$ gives

$$(\hat{x}_*, \hat{x}^*) = \underset{x_*, x^* \in L_m(X_{1:n}) : \frac{x^*}{x_*} > l}{\arg\min}\, \rho_{\mathrm{KS}}(x_*, x^*). \tag{12}$$

The above expression (12) for estimating the bounded power law interval $[\hat{x}_*, \hat{x}^*]$ will be referred to as the *KS method (for bounded power law fitting)*. The algorithm for this method is as follows:

Input: Random sample $X_{1:n}$.
Compute, from the random sample $X_{1:n}$, the subset $L_m(X_{1:n})$ where $m = 10$ as described in Subsection 4.2.3.
Loop: For every pair of elements $X_i < X_j$ in $L_m(X_{1:n})$: If the interval $[X_i, X_j]$ is sufficiently long, that is if $\frac{X_j}{X_i} > l = 10$:

(1) Compute the exponent $\hat{\alpha}(X_i, X_j)$ of the bounded power law fit as described in 4.2.1.
(2) Compute the KS distance between the bounded power law fit on $[X_i, X_j]$ with exponent $\hat{\alpha}(X_i, X_j)$ and the random sample using the KS distance formula for bounded power law fits (9).

Output: The bounded power law fit, among all fits, with the minimum KS distance is chosen as the bounded power law fit to the random sample $X_{1:n}$.

The computation time of the above algorithm is dominated by the computations of the KS distance over a number of candidate intervals which is weakly dependent on n, the size of the random sample. This algorithm is executed in about 0.1 seconds for $n = 10^3$ on a Macbook Pro with 2.9 GHz Intel Core i7 processor.

4.2.3. Considerations regarding candidate intervals for power law fit

As discussed briefly above, computational feasibility requires that candidate intervals have endpoints taken from a smaller subset of the given random sample $X_{1:n}$ when minimizing the KS distances of the bounded power law fits. This subset should also be chosen rich enough to allow accurate detection of the endpoints of the power law region.

Let m be the desired number of data points in any decade (an interval whose endpoints differ by a factor of 10) contained in the range of the random sample. Any subset of $X_{1:n}$ which contains at least m points in every of decade of this range must contain at least $m \cdot \left\lfloor \log_{10}\left(\frac{\max(X_{1:n})}{\min(X_{1:n})}\right)\right\rfloor$ points, a number which should increase slowly with respect to data size n. First, this many points are computed that are equally-spaced in the logarithmic scale from $\min(X_{1:n})$ to $\max(X_{1:n})$. Then, the closest data points to each of these logarithmically-equally-spaced points are chosen without repetition. This set contains the desired number of real data points which are roughly equally spaced in the logarithmic scale. This subset of $X_{1:n}$ is denoted by $L_m(X_{1:n})$.

A reasonable choice of m is 10. A larger choice adds considerably to the computation time while gaining little in the accuracy of the fitted power law interval. Notice that random samples that are spread over a wider range of values will require more computation time simply because they contain more decades.

4.2.4. Validation of bounded power law fits (p-value estimation)

As above for unbounded power law fitting, the validation approach from [2] is used to validate the bounded power law fits obtained by the penalized KS method: a bounded power law fit to a random sample is validated by comparing the KS distance of this fit to the KS distances of fits to semiparametric bootstrap samples. The number of semiparametric samples to test and the p-value threshold are chosen exactly as for unbounded power law fitting described in Subsection 2.2.

The bootstrap samples corresponding to a bounded power law fit on an interval are generated partially from the set of actual data points outside of the power law interval and partially from a bounded power law distribution using the exponent as estimated from the original data on the same interval. In effect, the sample is treated as if from a mixture distribution with one component within the power law interval and one without. Suppose the original random sample contains n data points where k of these data points are outside the fitted interval of the bounded power law fit. With probability k/n, a data point is selected with replacement from the set of data points outside the power law interval. With the complementary probability $1 - k/n$, a data point is simulated from the bounded power law distribution using the estimated power law exponent and using the inverse transform method. This process is repeated until the sample has n data points.

The time required for the validation of a bounded power law fit using 2500 semiparameteric bootstrap samples is approximately 5 minutes on a Macbook Pro with 2.9 GHz Intel Core i7 processor for random samples of size $n = 10^3$.

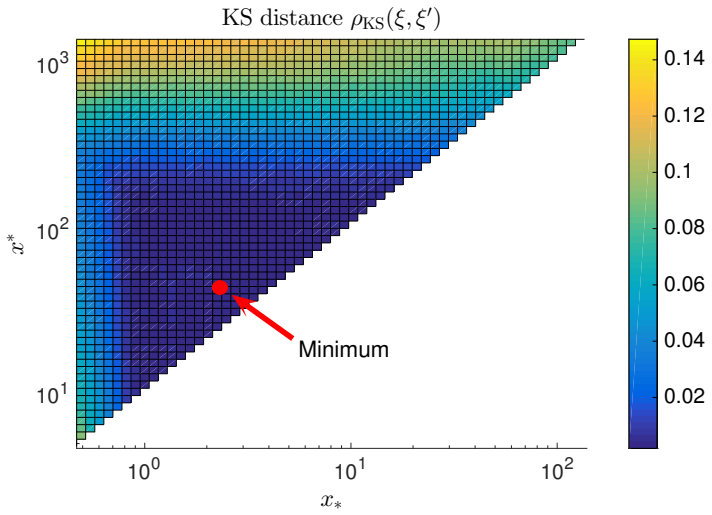

Figure 4. The KS method applied to a random sample generated by 10^4 samples from the EPL3 probability distribution (Appendix D.4) with exponent $\alpha = 1.5$ and nominal power law interval $[x_*, x^*] = [1, 100]$. The interval selected by minimizing the KS distance gives the indicated value $[\hat{x}_*, \hat{x}^*] = [2, 45]$, which is quite far from the desired result. Moreover, the indicated flatness of the minimum KS distance makes the selection of $[x_*, x^*]$ highly variable between samples from the same probability distribution.

4.3. Adaptively penalized KS method for a bounded power law fit

Concerns about the accuracy of estimated bounds are also valid for data with a bounded power law distributed region, especially so when the exponent α of the power law is less than 2. Figure 4 shows a heat map indicating the KS distance values for candidate power law intervals where the horizontal axis represents the lower bound candidate and the vertical axis represents the upper bound candidate. The tested random sample is power law distributed with $\alpha = 1.5$ on $[1, 100]$ according to the density described in Appendix D.4. As expected, smaller KS distance values are concentrated in and around the triangular section bounded by $x_* = 1$ and $x^* = 100$. The KS distance is usually minimized on an interval corresponding to a point in this figure near the point $(1, 100)$ corresponding to the true power law interval $[1, 100]$ for the tested distribution. However, that is not always the case, as seen in the figure.

The KS method produces a bounded power law fit to the data on the interval $[\hat{x}_*, \hat{x}^*]$ with the exponent $\hat{\alpha}(\hat{x}_*, \hat{x}^*)$. As explained above, the bounded power law interval $[\hat{x}_*, \hat{x}^*]$ might have been estimated too conservatively and erratically due to the noisy and roughly flat behavior of KS distances for intervals that lie in the true power law region. In order to address this and to improve the accuracy of the estimated power law interval, we introduce the *penalized KS distance* for bounded power law fits, defined as

$$\rho_{\mathrm{pKS}}(x_*, x^*) := \rho_{\mathrm{KS}}(x_*, x^*) + d \log\left(\frac{x^*}{C x_*}\right), \tag{13}$$

where $d > 0$ is a fixed penalty coefficient and C is a constant which can be chosen as the ratio of the smallest and the largest data points in a given random sample to keep the penalty term nonnegative. Notice that, compared to the KS distance defined in (9), the penalized KS distance favors (i.e. produces smaller values for) intervals with

smaller lower ends x_* and larger upper ends x^* for the bounded power law fits.

We now define an iterative optimization procedure for a given choice of penalty coefficient d, which will serve as a substep in the adaptively penalized scheme we will describe in Subsection 4.3.1. We start with the optimizing interval $[\hat{x}_*(1), \hat{x}^*(1)]$ for the ordinary KS distance (the result of the KS method described in Subsection 4.2). Then we switch to optimizing the penalized KS distance with respect to one endpoint, with the other held fixed, alternating endpoints and iterating until the process converges:

$$\begin{aligned}
[\hat{x}_*(1), \hat{x}^*(1)] &= \operatorname*{arg\,min}_{x_*, x^* \in L_m(X_{1:n}): \frac{x^*}{x_*} > l} \rho_{\mathrm{KS}}(x_*, x^*), \\
\hat{x}_*(k+1) &= \operatorname*{arg\,min}_{L_m(X_{1:n}) \ni x_* < \frac{\hat{x}^*(k)}{l}} \rho_{\mathrm{pKS}}(x_*, \hat{x}^*(k)), \\
\hat{x}^*(k+1) &= \operatorname*{arg\,min}_{L_m(X_{1:n}) \ni x^* > l \cdot \hat{x}_*(k+1)} \rho_{\mathrm{pKS}}(\hat{x}_*(k+1), x^*),
\end{aligned} \tag{14}$$

for $k = 1, 2, \ldots$ where l is the minimum ratio x^*/x_* for intervals considered by the algorithm. For penalty coefficient $d = 0$, this procedure would be identical to the KS method from Subsection 4.2.

In practice, $[\hat{x}_*(k), \hat{x}^*(k)]$ almost always converges to a single limit, which we then identify as the estimated interval $[\hat{x}_*, \hat{x}^*]$. Alternatively, since the iteration is a deterministic procedure on a finite state space, the intervals $[\hat{x}_*(k), \hat{x}^*(k)]$ could become periodic for large k. When such periodic oscillations occur, the interval with the largest value of $\hat{x}^*(k)/\hat{x}_*(k)$ is chosen. In the extremely rare case that there is more than one interval with the same maximum length, the one with the smaller value of the left endpoint $\hat{x}_*(k)$ is chosen, because typically more data points will fall within the power law interval.

The procedure described above for estimating the bounded power law interval using the penalized KS distance will be referred to as the *penalized KS method* with penalty coefficient d (for bounded power law fitting).

4.3.1. Algorithm for adaptively penalized KS method

The method just described requires a choice for the penalty coefficient d. Following the same logic for choosing d iteratively as in Subsection 3.3 for fitting power law tails, we now define the *adaptively penalized KS (apKS) method* (for fitting power laws over a bounded interval).

Input: Random sample $X_{1:n}$
KS Estimate: Apply the KS method, obtain a power law fit on an interval $[\hat{x}_*, \hat{x}^*]$ with exponent $\hat{\alpha}$. Continue only if the power law fit is validated by a semiparametric bootstrap, else terminate.
Initialize Penalization: Try $d = d_* = 10^{-10}$ and check the penalized KS method produces a validated power law fit, and try $d = d^* = 1$ and check the penalized KS method produces a power law fit which is rejected by the validation step. Terminate with an error if these checks fail. (This only seems to happen in unusual circumstances where the entire random sample can be fit by a power law.).
Loop:

(1) If $\frac{|d^* - d_*|}{d_*} < \epsilon_d = 10^{-2}$, then break out of the loop. Otherwise, continue.
(2) Take $d = \sqrt{d_* d^*}$.

Table 2. Sample run of the apKS method for bounded power law fitting on a random sample of size $n = 10^4$ that is power law distributed according to the EPL3 density given in AppendixD.4 with power law exponent $\alpha = 1.5$ on the interval $[1, 100]$, and initial penalty coefficients $d_* = 10^{-10}$ and $d^* = 1$. The relative tolerance $\epsilon_d = 10^{-2}$. The number of iterations required is 14. The row with the chosen power law fit is shown in bold.

d	$\hat{\alpha}$	$\hat{x}_*$	$\hat{x}^*$	$\rho_{\mathrm{KS}}(\hat{x}_*, \hat{x}^*)$	p-value	Validation
10^{-10}	1.48	2.67	138.64	0.0065	0.88	Accepted
1.00000	1.20	0.15	482.24	0.2287	0.00	Rejected
0.00001	1.48	2.67	138.64	0.0065	0.88	Accepted
0.00316	1.50	1.16	168.28	0.0068	0.44	Accepted
0.05623	1.50	0.62	482.24	0.0300	0.00	Rejected
0.01334	1.52	0.77	398.58	0.0160	0.00	Rejected
0.00649	1.49	0.77	210.28	0.0082	0.04	Rejected
0.00453	1.50	0.95	210.28	0.0069	0.20	Accepted
0.00542	1.50	0.95	210.28	0.0069	0.20	Accepted
0.00594	1.50	0.95	210.28	0.0069	0.20	Accepted
0.00621	1.49	0.77	210.28	0.0082	0.04	Rejected
0.00597	**1.50**	**0.95**	**210.28**	**0.0069**	**0.20**	**Accepted**
0.00607	1.49	0.77	210.28	0.0082	0.04	Rejected
0.00600	1.49	0.77	210.28	0.0082	0.04	Rejected

(3) Fit a power law over a bounded interval to $X_{1:n}$ by using the penalized KS method described in Eq. (14) with this penalty coefficient d.
(4) If the resulting bounded power law fit is validated, update $\hat{x}_*, \hat{x}^*, \hat{\alpha}$ and then set $d_* := d$; otherwise set $d^* := d$
(5) Go to step 1.

Output: A validated bounded power law fit on $[\hat{x}_*, \hat{x}^*]$ with the scaling parameter $\hat{\alpha}$.

A sample run of this algorithm is shown in Table 2. As for power law tail fitting, the algorithm is quite insensitive to the choice of the penalty coefficient d as seen in the table.

We remark some similarity in philosophy with the KS method extension of Deluca and colleagues [1,22], where power law fits over various intervals are considered and screened based on the p-value from a parametric bootstrap validation on that interval. The key distinction is that our proposed apKS method selects the interval for the power law fit by combining the value of the actual KS distance between the data and the proposed power law fit, the size (log-range) of the proposed interval, as well as the validation of the fit. The direct extension of the KS method of [2] would select the interval of the power law fit exclusively by minimizing the KS distance of the proposed power law fit from the data, which has some problems as illustrated in Figure 4. On the other hand, the method developed by Deluca and colleagues [1,22] instead makes no use of the actual value of the KS distance of the data from the proposed power law fit in selecting the interval; the quality of the proposed fit over a given interval is assessed only by whether it passes the parametric bootstrap validation criterion. In particular, no preference is given to an interval that has a smaller KS distance between the data and power law fit. Our penalization approach retains the feature of the original KS method in favoring the selection of the power law interval based on the quality of the fit, but rather than strictly optimizing the quality of the fit, trades off the size of the power law interval with the quality of the fit via a conversion factor set by the penalty coefficient d. The adaptive choice of d is an attempt to find the "right balance." We give some further comparison of the apKS method and the variation of the KS method proposed by Deluca and colleagues [1,22] in Appendix C.

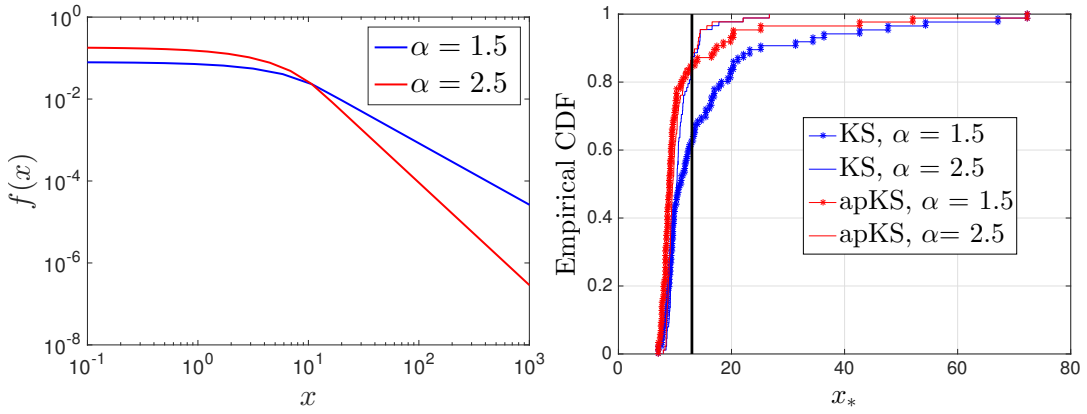

Figure 5. (Left) Log-log plots of probability density (2) with power law tail with exponents $\alpha = 1.5$ (blue) and $\alpha = 2.5$ (red) and lower bound $x_* = 13$. (Right) Empirical CDF of $\hat{x}_*$ obtained by KS (blue) and apKS (red) methods for 100 random samples of size $n = 10^4$ from these probability distributions. The thick vertical line is located at the true value $x_* = 13$ in the horizontal axis.

5. Results of Simulation Studies

The KS and apKS methods presented in the previous sections are tested on simulated random samples and the main results are presented in this section which is divided into two subsections: Subsection 5.1 for fitting power laws over semi-infinite tails, and Subsection 5.2 for power law fitting over a bounded interval. Supporting results from these numerical studies can be found in Appendix A. We remark that our implementation of the KS method differs slightly from the published code [26] by computing the KS distance (3) between a theoretical CDF $F(x)$ and the empirical CDF $\hat{F}(x)$ associated to the random sample $X_{1:n}$ as

$$\max_{1\le j\le n} \{|\hat{F}(X_j) - F(X_j)|, |\hat{F}(X_j) - \frac{1}{n} - F(X_j)|\} \tag{15}$$

rather than, as in [26],

$$\max_{1\le j\le n} \{|\hat{F}(X_j) - F(X_j)|\}.$$

The version (15) would seem to be more precise, since the maximum distance between the theoretical and empirical CDFs could be achieved as either left- or right-handed limits at the data points, though in practice the numerical results are hardly affected.

5.1. Results of Power Law Tail Fitting

We test the methods for power law tail fitting on random samples from a probability distribution that has a power law tail with a lower bound x_* and is exponentially distributed below, and which is defined more precisely in Appendix D.1. We consider two distinct power law exponents: $\alpha = 1.5$ and $\alpha = 2.5$; the PDFs are plotted in the left panel of Figure 5.

Empirical CDFs of the estimated lower bound parameter $\hat{x}_*$ are shown in the right panel of Figure 5. Both methods perform almost identically for the exponent $\alpha = 2.5$.

However, for the exponent $\alpha = 1.5$, the KS method selects a considerably larger and more variable lower bound $\hat{x}_*$ than the apKS method.

We show in Appendix A.1 that the improvements in estimating the lower bound by the apKS method do not come at the expense of the accurate estimation of the primary parameter of a power law model which is the exponent α. We now investigate the tradeoff between bias and variation in the estimation of the lower bound x_*. We saw in Figure 5 that the KS and apKS methods perform comparably well on random samples generated with power law exponent $\alpha = 2.5$ and lower bound $x_* = 13$. Figure A2 in Appendix A.1 confirms more generally that the apKS method does not improve upon the KS method for power law exponent $\alpha = 2.5$, and, in fact, creates a slightly larger negative bias. Thus we do not expect the apKS method to lead to much improvement on lower bound estimation for power law tails with large exponents (say $\alpha > 2$).

By contrast, for $\alpha = 1.5$, we saw in Figure 5 that the apKS method selects the lower bound with smaller bias and variability than the KS method, for reasons indicated by the discussion surrounding Figure 3. The bias and error in the estimation $\hat{x}_*$ of the lower bound is shown for various true lower bounds x_* and power law exponent $\alpha = 1.5$ in the upper half of Figure 6. The variability of the apKS estimate behaves rather erratically across different values of x_*; arguably the apKS estimator is somewhat more accurate than the KS estimator for most values of x_*. Figure 5 seems to suggest that the apKS method is producing a more reliable estimate than the KS method for most random samples, but both methods make terrible overpredictions on a small fraction of random samples. To show this observation from Figure 5 applies also more broadly, we also display in the lower half of Figure 6 the bias-error analysis applied to a trimmed collection of random samples, excluding those which gave the 5% worst estimations $\hat{x}_*$ for each method. This results in a trimming of $5 - 10\%$ of the random samples, depending on the degree of overlap of the random samples producing poor lower bound estimations by the two methods, and both methods are applied to a common trimmed collection of random samples for each true lower bound value x_*. On the trimmed collection of random samples, the apKS method can reduce the errors (typically more than 50%) in $\hat{x}_*$ with a small negative bias for a wide range of lower bound choices when the exponent is $\alpha = 1.5$. That is, based on the perspectives from Figure 5 and Figure 6, a small (5-10%) fraction of the random samples seem to show poor performance for both the KS and apKS methods in estimating the lower bound x_* of the power law tail, but the apKS method performs quite well (and better than the KS method) on most random samples from the probability distribution (2) with power law exponent $\alpha = 1.5$.

When $\alpha \leq 1.5$, the apKS method in fact makes a substantial reduction in bias and also reduces the variation in the estimation of lower bound x_*, as seen in Figure 7. We conclude from this bias-variation analysis that the apKS method can give substantial reduction in variation with minimal or no extra bias for sufficiently shallow (small α) power law tails, and gives rather close results to the ordinary KS method for steeper (larger α) power law tails.

In Appendix A.1.2, we examine the p-values of the KS and apKS methods reported in the validation step for our idealized power law tail model. We show that the KS method generally reports larger p-values than the apKS method, in fact with p-values generally higher than 0.5, and discuss possible reasons.

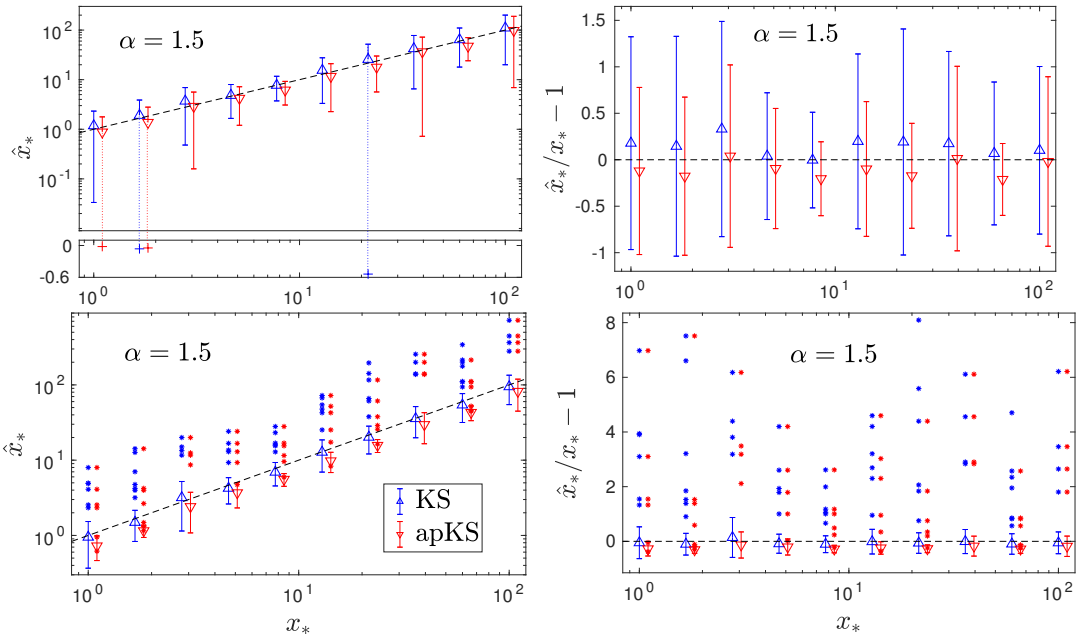

Figure 6. The results in the top half of the figure are obtained from 100 random samples of size $n = 10^4$ from the probability density (2) for power law exponent $\alpha = 1.5$ and the indicated values of lower bound x_*. The features have the same meaning as in Figure A2. The lower half of this figure reports the statistics of the estimators on a trimmed collection of random samples, excluding those with the worst 5% estimates by either the KS or apKS methods, as explained in the text. All the estimates excluded from the statistics are plotted separately as stars. These two panels in the lower half of the figure are the only place where this data trimming procedure was applied. The results for the apKS method are shifted to the right in order to avoid overlap.

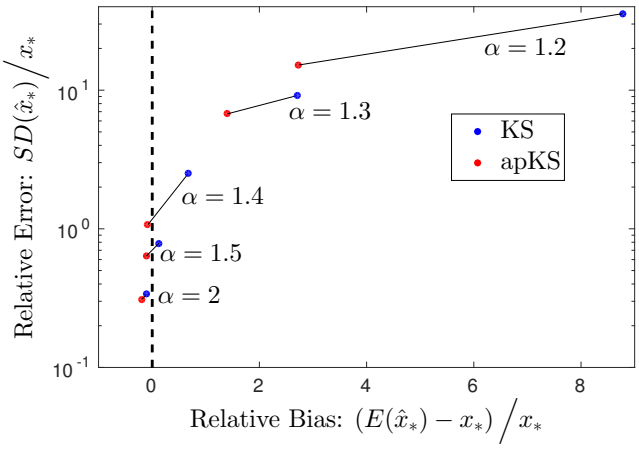

Figure 7. Relative bias and standard errors (KS blue, apKS red) of selected lower bound $\hat{x}_*$, based on 100 random samples from the probability density Eq. (2) for each indicated exponent with $x_* = 10$.

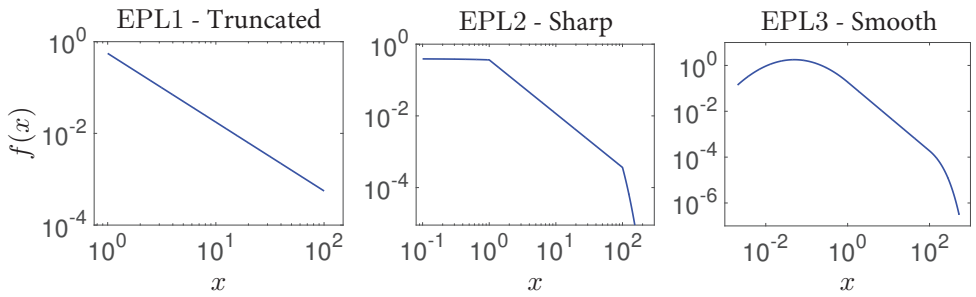

Figure 8. Log-log plots of PDFs of three probability distributions with exact power law distribution with exponent $\alpha = 1.5$ over a bounded interval $[x_*, x^*] = [1, 100]$. From left to right, these three distributions are EPL1 (Appendix D.2) - upper-truncated Pareto, EPL2 (Appendix D.3) - sharp transitions between power law and exponential regions, and EPL3 (Appendix D.4) - smooth transitions from lognormal to power law to exponential tail.

5.2. Results of Power Law Fitting over Bounded Interval

The ability of the KS and apKS methods to predict power-laws with both an upper and lower bound is tested on random samples drawn from three different distributions, defined in detail in Appendix D. These random samples are generated by using the inverse transform sampling method and by rejection sampling [27].

The three PDFs plotted in Figure 8 contain exact power law distributed intervals $[x_*, x^*]$. The left one, the EPL1 distribution, is an exact power law distribution supported on the interval $[x_*, x^*]$, also known as an "upper-truncated Pareto distribution" [12,18]. The middle one, the EPL2 distribution, has sharp transitions between a power law on $[x_*, x^*]$ and exponential behavior elsewhere. The right one, the EPL3 distribution, is like the EPL2 distribution except it has a lognormal core on $[0, x_*]$ and an exponential tail on $[x^*, \infty)$ which make a smoother transition into the power law region on $[x_*, x^*]$ at both ends. Precise formulas for these three probability distributions can be found in Appendices D.2, D.3, and D.4. In Appendix A.2, we show that the apKS method estimator $\hat{\alpha}$ for the power law exponent performs well for a wide range of α.

5.2.1. Estimation of the bounds of the power law interval

The estimated bounded power law intervals for random samples that are distributed as the EPL1, EPL2, and EPL3 distributions are shown, respectively, in Figures 9, 10 and 11. In these figures, we see that while the standard error of the upper bound estimator $\hat{x}^*$ gets worse for both methods as α increases, the standard error of the lower bound estimator $\hat{x}_*$ improves. This is not unexpected. As α increases, it becomes harder to estimate the upper bound because there will be fewer data points in the tail. This also means that there will be more points in the region where the true x_* lies, which is the reason why the standard error of $\hat{x}_*$ improves. A second reason for the improved standard error of $\hat{x}_*$ is that the EPL2 and EPL3 distributions have a sharper transition at x_* between the power law and non-power law regions as α increases. Another common feature seen in all three figures is that the standard errors of the estimated bounds are reduced with the apKS method relative to the KS method.

We also learn from these figures that the estimation of the upper bound produces

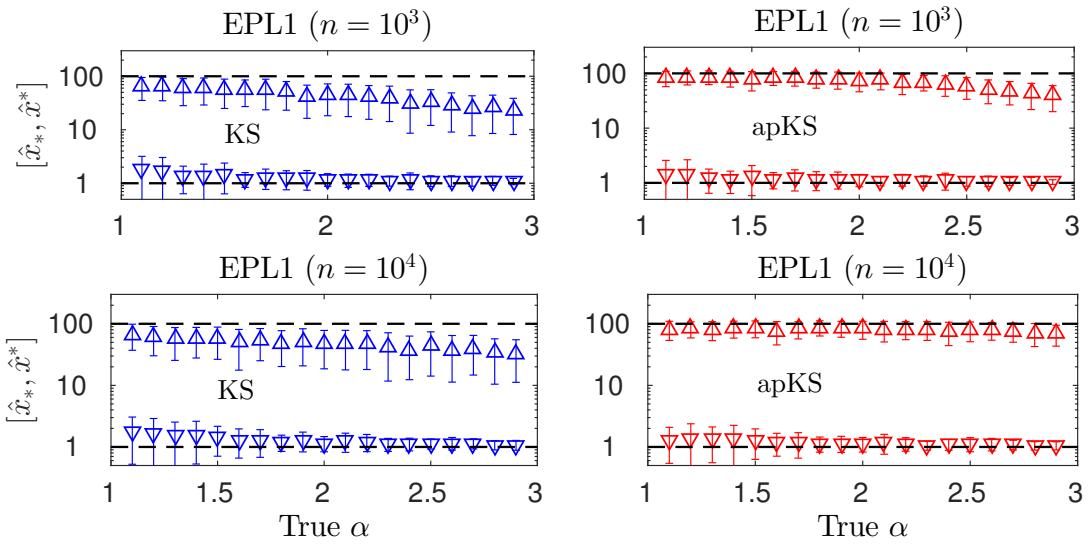

Figure 9. Estimated bounded power law intervals $[\hat{x}_*, \hat{x}^*]$ for those random samples from the EPL1 distribution (left panel of Figure 8) for which the bounded power law obtained by the KS method is validated. Left panels show the results of the KS method, while the right panels show the results of the apKS method. Top panels have sample size $n = 10^3$ while the lower panels have sample size $n = 10^4$. Approximately 10% of the 100 random samples are excluded due to failing validation with the $p-$valued threshold of 0.1. Error bars are centered at the average selected values (denoted by the symbols) and extend one standard error in each direction; the apparent asymmetry is due to the logarithmic scale of the vertical axis. The lower bars correspond to $\hat{x}_*$ and the upper bars correspond to $\hat{x}^*$. Thick black dashed lines correspond to the true bounds $[x_*, x^*] = [1, 100]$.

very large relative errors compared to the errors of the lower bound. Even for the EPL1 distribution where the entire random sample follows a power-law, the standard error of $\hat{x}^*$ can be comparable to the expected value of $\hat{x}^*$. The variability in the upper bound $\hat{x}^*$ estimated by a direct extension of the KS method was also noted in [1], and we see it still persists to some extent in the apKS method. The problem can essentially be blamed on the relative paucity of data points in the vicinity of x^*. Two positive results, however, are that: 1) the variability in the estimates of the bounds $\hat{x}_*$ and $\hat{x}^*$ don't appear to affect significantly the quality of the estimate of the power law exponent $\hat{\alpha}$ (Figure A4), and 2) when an ensemble of random samples from a common probability distribution is available (for example in simulation), then the average of the estimated bounds can be reasonably accurate. Though caution would appear to be indicated in interpreting the estimate for the upper bound x^* with either the KS or apKS method, we will discuss in Subsection 6.1 how the ensemble averaging idea can be applied even for a single random sample to improve the quality of the estimators for the bounds, $\hat{x}_*$ and $\hat{x}^*$.

6. Variance of Estimated Interval Bounds

We see from Figure A4 that the standard deviation of the estimated power law exponent decreases approximately proportionally to $n^{-1/2}$, as expected from an MLE with standard asymptotic behavior [28, Sec. 1.2]. By contrast, we see from Figures 9, 10 and 11 that the estimated bounds of a power law region in our distributions all have substantial variation which does not seem to improve noticeably when the random sample size is increased from $n = 10^3$ to $n = 10^4$. The first half of Table 3

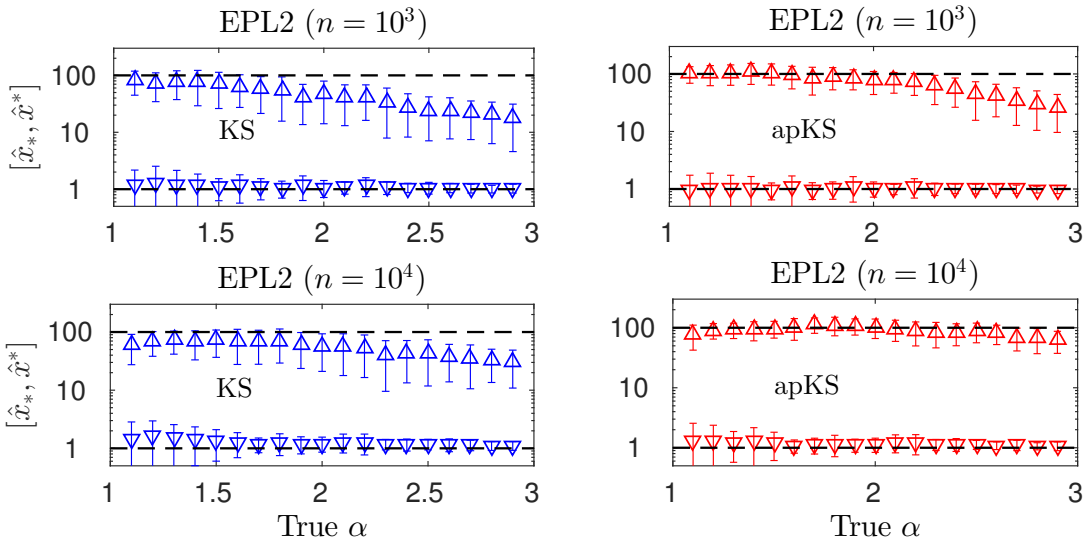

Figure 10. Estimated bounded power law intervals $[\hat{x}_*, \hat{x}^*]$ for the random samples from the EPL2 distribution (center panel of Figure 8) with true power law interval bounds $[x_*, x^*] = [1, 100]$. Other features of this figure are similar to Figure 9.

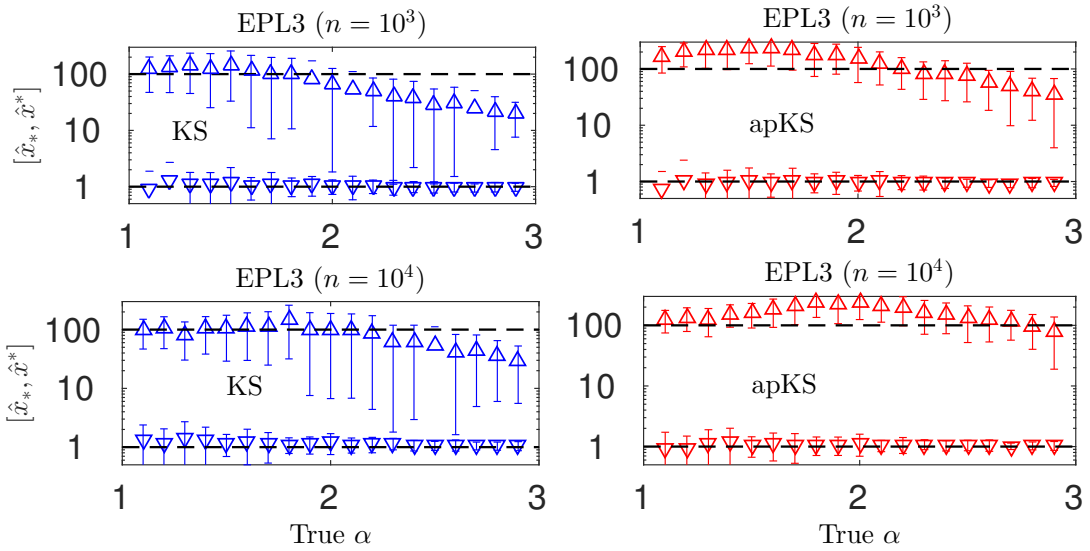

Figure 11. Estimated bounded power law intervals $[\hat{x}_*, \hat{x}^*]$ for the random samples from the EPL3 distribution (right panel of Figure 8) with true power law interval bounds $[x_*, x^*] = [1, 100]$. Other features of this figure are similar to Figure 9.

shows quantitatively that the bias and standard error of the power law interval bound estimates do not really improve as the random sample size is increased from $n = 10^3$ to $n = 10^4$ for the EPL3 distribution.

One way one might try to understand the apparently inherent variability and relative incompatibility of the power law intervals $[\hat{x}_*, \hat{x}^*]$ produced by the algorithms for different random samples from a common probability distribution is if we temporarily think of the bounds x_* and x^* as tuning parameters selected for the purpose of estimating the power law exponent α. This task might be viewed as analogous to kernel density estimation for the nonparametric fitting (or more precisely smoothing) of data to a probability density [28, Ch. 4]. Kernel density estimation requires a bandwidth (or smoothing parameter) h, which is generally adapted to the data. There appears to be no expectation that the bandwidth used for different random samples from the same probability distribution should be the same, nor that the probability density obtained from smoothing one random sample should necessarily be consistent with the other random samples. In other words, we have no sound theoretical basis for expecting good statistical properties for the estimators of the bounds, as we do for the MLE estimator $\hat{\alpha}$ for the power law exponent. A natural question is whether we can devise an MLE framework which jointly estimates the power law exponent and bounds of the power law interval. The implementation of an MLE estimate for the bounds, however, seems rather unclear without embedding the bounds in a probability model which also explicitly models the non-power law regions, and such an approach would seem undesirable for general applications.

Still, we might hope for better, since the bounds x_* and x^* should have some meaning in the context of the probability model, apart from the size of the random sample. To explore whether one could hope that another variation of the KS method might be capable of reducing the variability of the bound estimation further, we conduct in Appendix B a "compatibility of bounds" study for random samples generated from the EPL3 model, for which the power law bounds x_* and x^* have a precise meaning.

The general conclusion from this study is that when we determine the interval over which a power law can be fit to a random sample which is drawn from a distribution such as EPL3 which truly does have a bounded power law interval, then the inferred power law interval $[\hat{x}_*, \hat{x}^*]$ does not reliably support power law fits to other random samples drawn from the same distribution. This would seem to indicate that a substantial component of the variability in the inferred power law bounds $[\hat{x}_*, \hat{x}^*]$ is due to the variability in the quality of fit to an MLE power law fit, as measured by the KS metric, across random samples drawn from the same distribution, even if the underlying distribution does truly support a bounded power law interval. Therefore, further tweaks to the KS method (even more radical changes such as that proposed by [1], that only uses the KS metric to validate candidate intervals) are unlikely to remove the bulk of the variability in the estimates of the bounds of the power law interval to a given random sample.

6.1. Batching: A possible remedy

Combining this conclusion with our previous observation that the variability of the estimated power law interval bounds does not improve substantially with random sample size (Table 3) while the sample means are reasonably good at moderately large sample sizes, we propose subdividing the sample into batches, estimating bounds and power law exponents for each batch, and then obtaining statistical estimates for the

whole random sample by averaging the results from each batch. More precisely, the *batching* approach can be summarized as follows:

(1) Split a sample into b disjoint subsamples (i.e. *batches*) whose union is the original random sample.
(2) Run the KS (or apKS) method on each batch and validate each bounded power law fit by estimating a p-value obtained by using semiparametric bootstrap samples. By collecting the estimates from each batch, we obtain power law exponents $\hat{\alpha}^{(1)}, \ldots, \hat{\alpha}^{(b)}$, bounded power law intervals $[\hat{x}_*^{(1)}, \hat{x}^{*(1)}], \ldots, [\hat{x}_*^{(b)}, \hat{x}^{*(b)}]$, along with the $p-$values.
(3) If all of the bounded power law fits are validated (sufficiently large p-value), report the average of the estimated exponents and bounds as the bounded power law parameters. The bounded power law hypothesis is deemed not valid otherwise.

The basic idea is that if the quality of the results isn't improving for the size of the random sample, then such batching is a natural way to make use of the data to average out model fitting errors. More precisely, if every data point is independently sampled from a common underlying probability distribution, then the estimators from each batch will be independent, so averaging over batches will produce at least a reduction in the standard error as more batches are included. Indeed, such a batching approach is used for the same reason in one class of methods for estimating the spectral density of a statistically stationary signal by breaking it into subintervals, computing the periodogram on each, and then averaging the periodograms [29]. We have no guarantee, of course, that this batching approach would lead to an asymptotically unbiased estimate for the bounds of the power law region. Results of the KS and apKS methods with batching are shown in the lower half of Table 3. We see indeed that the variability of the bounds across different random samples is noticeably reduced by batching. The bias in $\hat{x}_*$ appears somewhat lessened, while the bias in $\hat{x}^*$ shows no improvement by batching. We note that, as in all previous tables, the reported errors of the estimators are simply standard errors computed from the collection of estimates for each of the (here 50) random samples considered which passed the validation step on each batch. The batching approach in fact gives a way to obtain some sort of error estimate for both the power law exponent and power law interval bounds from a single sample by computing a sample standard deviation across batches.

We have only attempted here to give some preliminary results with the batching idea. Many issues merit more in-depth analysis in future work. For instance, the validation of a power law interval in a random sample shouldn't necessarily require validation on each batch, since even probability distributions with a true power law interval (as in the EPL3 model we have been using) will fail the validation step on each batch with some probability set by the critical p-value. Another key issue is determining how many batches should be used for a given sample. Perhaps this could be determined by studying at what random sample size the quality of the parameter estimates stops increasing, and using that as a guide to batch size. ([16] found a sample size of 10^4 to be adequate for accurate power law exponent estimation for their method of maximum likelihood maps; our preliminary results suggest a batch size of 10^3 might be workable for the apKS method, especially given the averaging across batches.) Alternatively, one might try an adaptive approach which tries processing the random sample with different numbers of batches.

Table 3. Averages and standard errors of validated bounded power law fit parameters on 50 random samples from the EPL3 distribution containing an exact power law with exponent $\alpha = 1.5$ over the interval $[x_*, x^*] = [1, 100]$ and for various sample sizes n. Results corresponding to 29 (out of 50) validated power law fits are shown in the table for $n = 10^3$, 28 validated fits for $n = 10^4$, and 18 validated fits for $n = 10^5$. Here a validated power law fit to a random sample means that the power law fit was validated on each batch obtained from that sample. The rows denoted "w/b" employ the batching approach described in Subsection 6.1, with 10 batches used in each case.

Method	n	$\hat{\alpha}$	$\hat{x}_*$	$\hat{x}^*$
KS	10^3	1.49 ± 0.06	$1.04 \pm \mathbf{0.58}$	$104.72 \pm \mathbf{82.20}$
KS	10^4	1.50 ± 0.02	$1.20 \pm \mathbf{0.54}$	$99.67 \pm \mathbf{78.69}$
KS	10^5	1.50 ± 0.01	$1.39 \pm \mathbf{0.51}$	$77.45 \pm \mathbf{56.63}$
apKS	10^3	1.50 ± 0.06	$0.88 \pm \mathbf{0.24}$	$250.28 \pm \mathbf{136.99}$
apKS	10^4	1.49 ± 0.02	$1.11 \pm \mathbf{0.21}$	$181.87 \pm \mathbf{60.77}$
apKS	10^5	1.50 ± 0.00	$1.11 \pm \mathbf{0.42}$	$115.00 \pm \mathbf{54.57}$
KSw/b	10^3	1.46 ± 0.07	$0.88 \pm \mathbf{0.21}$	$87.60 \pm \mathbf{37.26}$
KSw/b	10^4	1.50 ± 0.02	$1.08 \pm \mathbf{0.38}$	$131.38 \pm \mathbf{39.43}$
KSw/b	10^5	1.50 ± 0.01	$1.25 \pm \mathbf{0.20}$	$101.75 \pm \mathbf{23.16}$
apKSw/b	10^3	1.46 ± 0.06	$0.78 \pm \mathbf{0.21}$	$166.45 \pm \mathbf{35.67}$
apKSw/b	10^4	1.51 ± 0.02	$0.97 \pm \mathbf{0.14}$	$247.30 \pm \mathbf{28.35}$
apKSw/b	10^5	1.50 ± 0.01	$1.07 \pm \mathbf{0.13}$	$176.82 \pm \mathbf{15.00}$

7. Conclusions

We have developed a generalization of the KS method described in [2] for determining and validating power law tail fits in data sets representable as random samples to determining and validating power law fits over bounded intervals. The motivation for this extension is that several theoretical models for power law behavior can make explicit reference to mechanisms which provide upper and lower cutoffs to the power law behavior. Statistical methods for fitting power law tails can be expected to stumble on data which are generated from or might be fit to such theoretical models, since the tail region would include data from a region which is not supposed to even theoretically approximate a power law behavior.

Because the size of the interval of a putative power law fit is often taken as one measure of the significance of the power law model to the data set [3], we focused specific attention to how the lower bound to the power law interval is selected by the KS method and our extension of it to bounded power law intervals. We found that even on idealized distributions with exact power law regions and sharp transitions to non-power-law behavior, the KS method produces highly variable values for the bounds of the power law region. The reason was traced to the fact that these bounds are selected by minimizing a KS metric over a set of candidate bounds, and the KS metric function being minimized can have a very shallow and broad minimum. Consequently, due to finite sampling fluctuations, the minimizing value would jump around within a broad region for different random samples from the same probability distribution. For these idealized probability distributions, the introduction of a penalty term in the optimization step gave substantial improvement to the quality of the bounds selected, reducing their variability across random samples at a small bias cost. Integral to this success is a data-adapted, iterative approach to choosing the coefficient of the penalty term. The improvement of our penalized version of the KS method relative to the original KS

method was more pronounced for shallower power laws (smaller values of α). A probability distribution with a finite mean and a power law tail must have the exponent of the power law tail satisfy $\alpha > 2$, but no such restriction applies to the exponent characterizing a bounded power law region. Consequently, the improvements from the adaptively penalized optimization version of the KS method for smaller exponents α are likely more relevant for the fitting of bounded power law regions relative to power law tails.

The variations in the estimated bounds, though reduced by the apKS method, are still seen to be substantial. By examining the extent to which bounds selected for the power law region on one random sample could be consistently applied to another random sample from the same distribution, we found that the variability in the bounds across random samples seems to be an inherent finite sampling issue, at least for methods which use the KS metric to assess the quality of fit in choosing the bounds of the power law region. This would appear to cast doubt on the ability to obtain meaningful estimates for the bounds of a power law region, but we offer a few more encouraging observations.

First of all, this variability in the bound estimates is an issue primarily when analyzing a single sample obtained from experiment, observation, or survey. For samples generated by simulation and/or repeatable experiments in the same environment, the consideration of the statistics of the bounds of the power law region selected for each sample can more reliably be related to a possible underlying theoretical probability model with a power law tail or bounded power law interval. For then, just as in the results reported in the present work, error bars can be reported and the averages of the power law interval bounds converge to meaningful values with at worst moderate bias. In future work (**citation suppressed in anonymous version**), we will apply in this way the apKS method to stochastic simulations of a certain neuronal network model [30,31] which appears on log-log plots to produce power law behavior over only an intermediate range (not a full tail). Secondly, even if only a single random sample is available, we provided some preliminary results in Section 6 showing improved statistics by breaking the random sample into smaller batches, applying the KS or apKS method on these, and then averaging the batch estimates for the power law exponent and bounds.

Our purpose in the present work has been to develop and examine extensions of the KS method for estimating the exponent and bounds of a power law region for idealized probability distributions with unambiguous power law behavior over a finite interval. In future work (**citation suppressed in anonymous version**), we shall explore the behavior of the apKS and KS methods when applied to random samples generated by probability distributions that have bounded power law regions in an intermediate asymptotic, rather than exact sense, meaning that the transition between the power law interval and elsewhere is smoothly distributed rather than an abrupt change at precise values x_* and x^*. Such power law regions are closer to what one expects for probability distributions of quantities obtained from experimental and simulated data, when one seeks to claim such power law behavior is supported. But the corresponding lack of precision of the values of x_* and x^* produces some inherent ambiguity as to what a good estimation scheme should be achieving. In (**citation suppressed in anonymous version**), we will show how the apKS method's estimate of bounds even in these circumstances yield useful quantitative information, with a measurable improvement in quality from the bounds estimated by the KS method.

References

[1] Deluca A, Corral Á. Fitting and goodness-of-fit test of non-truncated and truncated power-law distributions. Acta Geophysica. 2013;61(6):1351–1394.

[2] Clauset A, Shalizi CR, Newman MEJ. Power-Law Distributions in Empirical Data. SIAM Review. 2009;51(4):661–703. Available from: http://epubs.siam.org/doi/10.1137/070710111.

[3] Stumpf MP, Porter MA. Critical truths about power laws. Science. 2012;335(6069):665–666.

[4] Handcock MS, Jones JH. Likelihood-based inference for stochastic models of sexual network formation. Theoretical Population Biology. 2004;65(4):413–422.

[5] Mitzenmacher M. A brief history of generative models for power law and lognormal distributions. Internet Mathematics. 2004;1(2):226–251.

[6] Marković D, Gros C. Power laws and self-organized criticality in theory and nature. Physics Reports. 2014 Mar;536(2):41–74. Available from: http://www.sciencedirect.com/science/article/pii/S0370157313004298.

[7] White EP, Enquist BJ, Green JL. On estimating the exponent of power-law frequency distributions. Ecology. 2008;89(4):905–912.

[8] Goldstein ML, Morris SA, Yen GG. Problems with fitting to the power-law distribution. European Physical Journal B. 2004;41(2):255–258.

[9] Bauke H. Parameter estimation for power-law distributions by maximum likelihood methods. European Physical Journal B. 2007;58(2):167–173.

[10] Edwards AM. Using likelihood to test for Lévy flight search patterns and for general power-law distributions in nature. Journal of Animal Ecology. 2008;77(6):1212–1222.

[11] Laurson L, Illa X, Alava MJ. The effect of thresholding on temporal avalanche statistics. Journal of Statistical Mechanics: Theory and Experiment. 2009;2009(1):P01019.

[12] Aban IB, Meerschaert MM, Panorska AK. Parameter estimation for the truncated Pareto distribution. Journal of the American Statistical Association. 2006;101(473):270–277.

[13] Langlois D, Cousineau D, Thivierge JP. Maximum likelihood estimators for truncated and censored power-law distributions show how neuronal avalanches may be misevaluated. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics. 2014;89(1):12709.

[14] Mashanova A, Oliver TH, Jansen VAA. Evidence for intermittency and a truncated power law from highly resolved aphid movement data. Journal of The Royal Society Interface. 2010 Jan;7(42):199–208. Available from: http://rsif.royalsocietypublishing.org.libproxy.rpi.edu/content/7/42/199.

[15] Maschberger T, Kroupa P. Estimators for the exponent and upper limit, and goodness-of-fit tests for (truncated) power-law distributions. Monthly Notices of the Royal Astronomical Society. 2009;395(2):931–942.

[16] Baró J, Vives E. Analysis of power-law exponents by maximum-likelihood maps. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics. 2012;85(6):66121.

[17] Edwards AM, Freeman MP, Breed GA, et al. Incorrect Likelihood Methods Were Used to Infer Scaling Laws of Marine Predator Search Behaviour. PLoS ONE. 2012;7(10):e45174.

[18] Clark DR. A Note on the Upper-Truncated Pareto Distribution. In: Casualty Actuarial Society E-Forum, Winter; April; 2013. p. 1–22. Available from: http://www.soa.org/Library/Monographs/Other-Monographs/2013/April/mono-2013-as13-1-clark.pdf.

[19] Zhang J. Reducing bias of the maximum-likelihood estimation for the truncated Pareto distribution. Statistics. 2013 Aug;47(4):792–799.

[20] Redner S. How popular is your paper? an empirical study of the citation distribution. The European Physical Journal B-Condensed Matter and Complex Systems. 1998;4(2):131–134.

[21] Clauset A, Young M, Gleditsch KS. On the Frequency of Severe Terrorist Events. Journal of Conflict Resolution. 2007;51(1):58–87. Available from: http://arxiv.org/abs/physics/0606007.

[22] Peters O, Deluca A, Corral A, et al. Universality of rain event size distributions. Journal of Statistical Mechanics: Theory and Experiment. 2010;2010(11):P11030.
[23] Mansfield ML. Numerical tools for obtaining power-law representations of heavy-tailed datasets. The European Physical Journal B. 2016 Jan;89(1):1–13. Available from: http://link.springer.com.libproxy.rpi.edu/article/10.1140/epjb/e2015-60452-3.
[24] Feller W. An Introduction to Probability Theory and Its Applications, Vol. 1, 3rd Edition. WSE; 2008.
[25] Mayo DG, Cox DR. Frequentist statistics as a theory of inductive inference. Lecture Notes-Monograph Series. 2006;:77–97.
[26] Clauset A. Power-law distributions in empirical data ; 2007 [cited March 4, 2018]; Available from: http://tuvalu.santafe.edu/ aaronc/powerlaws.
[27] Press WH, Teukolsky SA, Vetterling WT, et al. Numerical recipes in FORTRAN. 2nd ed. Cambridge: Cambridge University Press; 1992. Chapter 7; p. 266–319.
[28] Wasserman L. All of Nonparametric Statistics. New York: Springer; 2007.
[29] Yaglom AM. Correlation theory of stationary and related random functions. Volume I: Basic results. New York: Springer-Verlag; 1987.
[30] Schmidt D, Best J, Blumberg MS. Random graph and stochastic process contributions to network dynamics. In: Feng W, Feng Z, Grasselli M, et al., editors. Dynamical Systems and Differential Equations, DCDS Supplement 2011 Proceedings of the 8th AIMS International Conference (Dresden , Germany); September. American Institute of Mathematical Sciences; 2011. p. 1279–1288.
[31] Best J, Behn CD, Poe GR, et al. Neuronal Models for Sleep-Wake Regulation and Synaptic Reorganization in the Sleeping Hippocampus. Journal of Biological Rhythms. 2007 Jun; 22(3):220–232. Available from: http://jbr.sagepub.com/content/22/3/220.

Appendix A. Supplementary Results from Simulation Studies

A.1. Supplementary Results for Power Law Tail Fitting

A.1.1. Supplementary Bias-Error Analysis of Estimators

The results of the bias-error analysis of the exponent are shown in Figure A1. The exponents estimated by the apKS method are consistently more biased compared to the KS method but the size of the excess bias is not more than 0.01, which is quite small relative to the target value $\alpha = 1.5$ being estimated, as well as the standard error of either method. The reason for this extra bias is that the apKS method pushes the lower bound of the power law tail further left than the KS method, so the quality of the data in the tail will have somewhat less of the ideal power law quality. The standard errors of the estimate $\hat{\alpha}$ are comparable in both the KS and apKS methods.

A.1.2. p-values reported in validation step

By definition, the KS distances for power law fits obtained by the apKS method are slightly larger than those for the KS method. However, the validation process is identical for the two methods. Therefore, the power law fits obtained by the apKS methods are expected to have smaller p-values than the KS method. Recall from Subsection 2.2 that the p-values are defined as the fraction of the semiparametric bootstrap samples from the power law fit which have larger KS distances from their power law fits than the original random sample had to its power law fit. This is precisely what we mean by p-value in the following discussion; recall this quantity does not in our context quite have the usual meaning of a criterion for rejecting a null

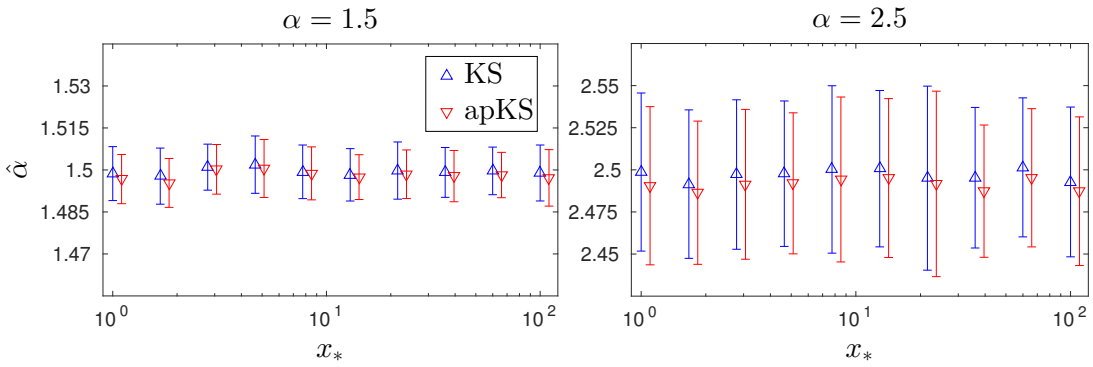

Figure A1. Estimates $\hat{\alpha}$ for the power law exponent obtained from 100 random samples of size $n = 10^4$ from the probability density (2) for each indicated value of power law exponent α and lower bound x_*. The error bars are centered at the average of $\hat{\alpha}$ and extend by one standard error in each direction. Left panel corresponds to $\alpha = 1.5$ and the right panel corresponds to $\alpha = 2.5$. The results for the apKS method are shifted to the right in order to avoid overlap.

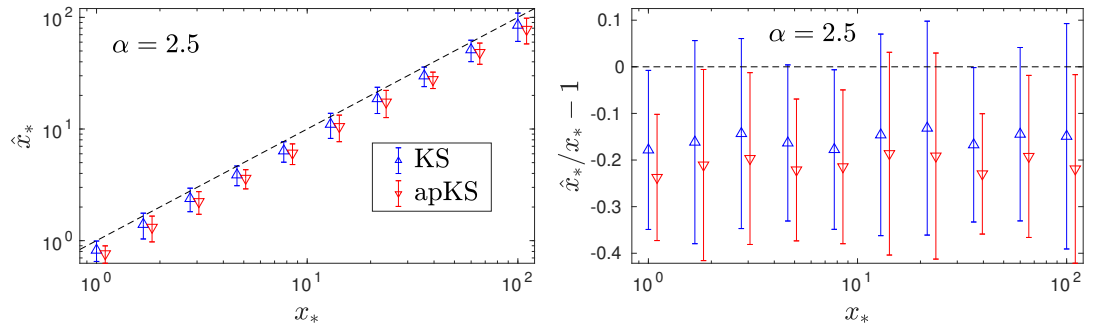

Figure A2. Estimates $\hat{x}_*$ for the lower bound of power law tail obtained from 100 random samples of size $n = 10^4$ from the probability density (2) for power law exponent $\alpha = 2.5$ and the indicated values of actual lower bound of the tail, x_*. Ten logarithmically-equally spaced choices from 1 to 100 are considered for x_* in this model. The error bars on the left side are located at the average estimated value of $\hat{x}_*$ and extend up and down by one standard error. The right panel plots the same results in terms of the relative size of the estimated values of $\hat{x}_*$ with respect to the ideal target values x_*. The results for the apKS method are shifted to the right in order to avoid overlap.

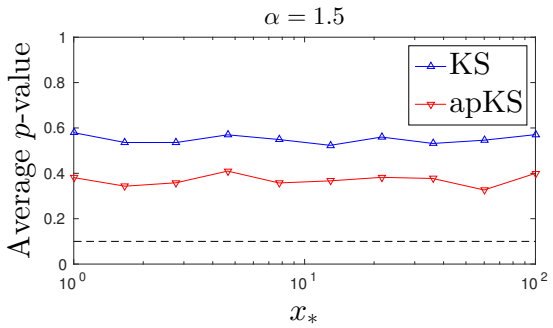

Figure A3. Average p-values of power law fits plotted in the left panel of Figure A1.

hypothesis. In order to get a sense of how much smaller the p-values are, following [2], *average* p-values are plotted in Figure A3 for the ensemble of 100 random samples of size $n = 10^4$ from the probability density (2) with power law exponent $\alpha = 1.5$ and various choices of lower bound. As expected, the power law fits obtained by the apKS method have an average p-value that is smaller than those for the KS method but also significantly larger than the threshold 0.1 for acceptance of a power law fit. Consistent with the choice of 0.1 as a threshold for the p-value in the validation step, the power law fits by the KS method are validated 91% of the time when $\alpha = 1.5$, and 90.5% of the time when $\alpha = 2.5$ for the model (2), which truly has a power law tail.

What might be somewhat surprising is to see p-values have an average greater than 0.5 for the KS method. Indeed, [2] shows that when the KS method is tested on a Pareto distribution, which is a pure power law distribution, the average p-values obtained at the validation step are about 0.5. But we can understand how a distribution such as (2), with a power law tail but non-power law core, could produce p-values at the validation step greater than 0.5 on average. First of all, the distribution does have a perfect power law tail, so the fit quality of the sample over that region should be comparable in both the original sample and the bootstrapped samples. Next we note that the KS method would estimate the lower bound of the power law region as $\hat{x}_* < x_*$ in 71% of the random samples simulated to generate Figure A3. That is, even the KS method is somewhat (though mildly) aggressive in its estimation of $\hat{x}_*$. Since the KS method is self-consistently also applied to the bootstrapped samples, which have a perfect power law at least for $x \geq \hat{x}_*$, the lower bounds $\hat{x}_*$ estimated in the validation step will more often than not be more aggressive than the lower bound estimated for the original random sample: $\hat{x}_*^B \lesssim \hat{x}_*$. Since the KS distances for the power law fits for the original and bootstrapped random samples will be comparable over the true power law region $x \geq x_*$, the bootstrapped sample will tend to have a larger KS distance from the power law fit over its claimed power law region $(\hat{x}_*^B, \infty)$ than the KS distance for the original power law fit whenever $\hat{x}_*^B < x_* < \hat{x}_*$, which is reasonably often. In fact, even when $\hat{x}_*^B < \hat{x}_* < x_*$, both the power law fit to the original and bootstrapped sample extends beyond the bound of the power law tail in the generating distribution, so the KS distance for the bootstrapped power law fit can still be often greater than that of the original power law fit. We close by remarking that the high p-values of course only arise because the distribution has a perfect power law component.

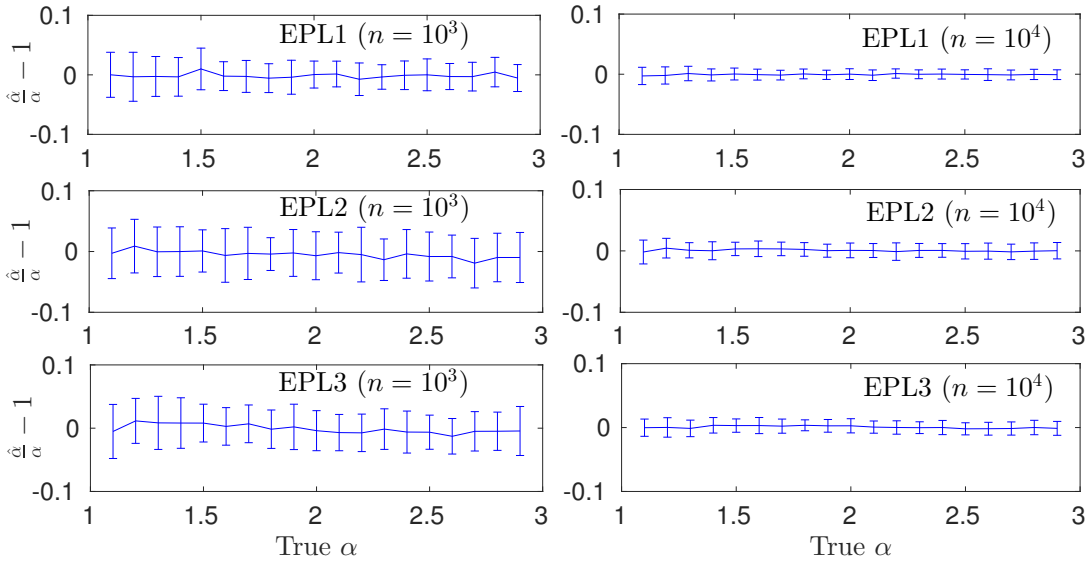

Figure A4. The apKS method is applied to 100 random samples which are simulated from probability distributions with exact power law intervals with exponents $1.1 \leq \alpha \leq 2.9$. The figure shows the statistics for the roughly 90% of the simulations that resulted in a validated bounded power law. Error bars are centered at the relative bias in $\hat{\alpha}$ (i.e. $\mathbb{E}(\hat{\alpha})/\alpha - 1$) with bars going as far as one relative standard deviation, i.e. $\mathrm{SD}(\hat{\alpha})/\alpha$. From top to bottom, we plot the results for the EPL1, EPL2, and EPL3 distributions, displayed from left to right in Figure 8. The results shown in the first column correspond to random samples of size $n = 10^3$ and the second column corresponds to random samples of size $n = 10^4$.

A.2. Supplementary Results for Power Law Fitting over Bounded Interval

Averages and relative errors of the estimator $\hat{\alpha}$ by the apKS method are shown in Figure A4. The apKS method identified validated bounded power laws roughly 90% of the time which is consistent with our choice of the $p-$value threshold of 0.1. The relative error in $\hat{\alpha}$ (out of these roughly 90 of 100 simulations) is about 5% for random samples of size $n = 10^3$ and it is about 2% for random samples of size $n = 10^4$. The results corresponding to the KS method (not shown) for power law fitting over bounded intervals are comparable to those of the apKS method.

Appendix B. Compatibility of Bounds Study

We generated 100 random samples of size $n = 10^4$ from the EPL3 model, with power law exponent $\alpha = 1.5$ and power law region $[x_*, x^*] = [1, 100]$. For each random sample $j = 1, \ldots, 100$, the KS method was used to estimate a power law exponent $\hat{\alpha}^{(j)}$ and power law interval $[\hat{x}_*^{(j)}, \hat{x}^{*(j)}]$. 88 out of these 100 random samples passed the usual semiparametric bootstrap validation test. Then, for each j, we loop over all other random samples $j' \neq j$, and check whether a power law fit using the parameters $\{\hat{\alpha}^{(j)}, \hat{x}_*^{(j)}, \hat{x}^{*(j)}\}$ would pass the semiparametric bootstrap validation step for random sample j'. We found that no more than 19 of the 100 random samples would validate a power law fit for any of the collection of fitting parameters $\{\hat{\alpha}^{(j)}, \hat{x}_*^{(j)}, \hat{x}^{*(j)}\}$, and conversely, no more than 21 of these 100 collections of fitting parameters would be validated on any given random sample.

Part of the reason for the low success rate for this compatibility of bounds study

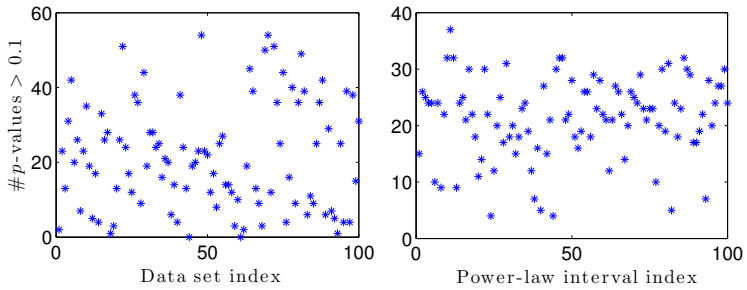

Figure B1. Compatibility of bounds study of bounded power law fits to 100 independent random samples of size $n = 10^4$ from the EPL3 distribution with true exponent $\alpha = 1.5$ and true power law region $[x_*, x^*] = [1, 100]$. The bounded power law parameters $\{\hat{\alpha}^{(j)}, \hat{x}_*^{(j)}, \hat{x}^{*(j)}\}$ fit to each random sample $j = 1, \ldots, 100$ using the KS method were adjusted, as described in the main text, to the other random samples $j' \neq j$, and the semiparametric bootstrap validation was used to calculate a p-value for the hypothesis that the power law fit to the jth random sample, after adjustment, was also a good fit to the random sample j'. The left panel indicates, for each random sample $j = 1, \ldots, 100$, the number of adjusted power law fits from other random samples that resulted in a p-value greater than 0.1, indicating agreement with the random sample in question. The right panel indicates, for each collection of power law fit parameters $\{\hat{\alpha}^{(j)}, \hat{x}_*^{(j)}, \hat{x}^{*(j)}\}$ obtained from random samples $j = 1, \ldots, 100$, how many other random samples would give a p-value greater than 0.1, indicating agreement with those fitting parameters, after adjustment.

might be that the quality of fit as measured by the KS metric might be quite sensitive to fluctuations in the estimators. Therefore, we next repeated this compatibility of bounds study, with some adaptation of the fitting parameters to each other random sample. We are really interested in the question of whether the power law interval $[\hat{x}_*^{(j)}, \hat{x}^{*(j)}]$ estimated from one random sample also serves as a good power law interval for other random samples. So in the second compatibility of bounds study, when we apply the fitting parameters $(\hat{\alpha}^{(j)}, \hat{x}_*^{(j)}, \hat{x}^{*(j)})$ from random sample j to another random sample j', we first adjust the bounds of the power law interval $[\hat{x}_*^{(j)}, \hat{x}^{*(j)}]$ by moving them to the closest values in random sample j', and then refit the power law exponent using the KS method over this adjusted interval. The adjustment of the power law interval tries to remove quality-of-fit discrepancies that could arise from the fact that the power law interval bounds fitted to random sample j also correspond to data points in that set, and the refitting of the power law exponent tries to remove the possible quality-of-fit deficiency because the estimated power law exponent $\hat{\alpha}^{(j)}$ from random sample j will have statistical fluctuations adapted to that random sample. Note, however, that even with these adjustments, we are still essentially checking whether the power law bounds $[\hat{x}_*^{(j)}, \hat{x}^{*(j)}]$ can serve as suitable power law bounds (with possibly different exponent) for a different random sample j', drawn from the same EPL3 distribution. The results of this cross-validation study, with adjusted fits, are presented in Figure B1.

We see the adjustment of the fitting parameters does somewhat improve the statistical self-consistency of the power law fits across the random samples drawn from the common distribution, but not to a high degree. From the left panel, we see that no sample will pass the semiparametric bootstrap validation step for more than 60% of the power law intervals $[\hat{x}_*^{(j)}, \hat{x}^{*(j)}]$ obtained from the other samples, and many samples are statistically compatible with fewer than 20% of the power law intervals obtained from other random samples. Similarly, the right panel shows that any given power law interval $[\hat{x}_*^{(j)}, \hat{x}^{*(j)}]$ obtained from a given random sample will yield a validated power law fit (with possibly different exponent!) for no more than 40% of the other

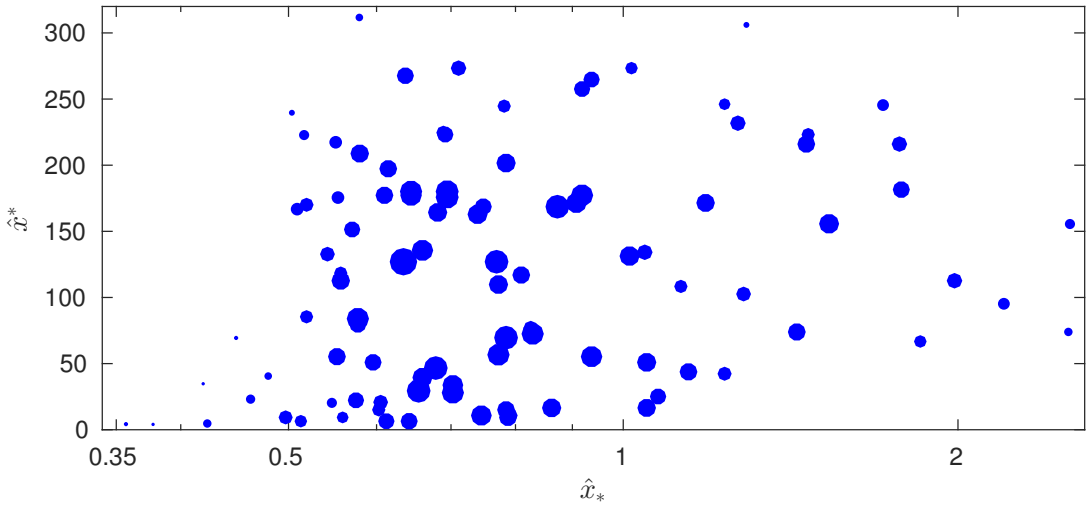

Figure B2. Each one of the 100 points in this plot represents the bounded power law interval selected by the KS method, validated or not, for a random sample of size $n = 10^4$ drawn from the EPL3 distribution with a bounded power law interval $[1, 100]$ and exponent $\alpha = 1.5$. The size of the solid circle about each point is in proportion to the fraction of the other 99 random samples for which, after the adjustment step described in the main text, a power law fit over that interval is validated by the semiparametric bootstrap.

random samples, with many intervals being statistically compatible with a power law over fewer than 25% of the other random samples. Experiments with larger samples of size $n = 10^6$ gave similar results.

To see how the actual values of the power law bounds $[\hat{x}_*^{(j)}, \hat{x}^{*(j)}]$ influence their compatibility with other random samples, we present in Figure B2 a scatter plot of the 100 combinations of inferred power law intervals $\{\hat{x}_*^{(j)}, \hat{x}^{*(j)}\}_{j=1}^{100}$, with the size of the dot indicating the fraction of other random samples to which that interval (after adjustment) gave rise to a validated power law fit. One might expect that more conservative bound values satisfying $\hat{x}_* > x_* = 1$ and $\hat{x}^* < x^* = 100$ should be statistically compatible for most other random samples, but this is not seen to be the case. Presumably the failure of the compatibility of power law fits over these conservative intervals suffer from a reduction in the number of data points falling in the putative power law interval $[\hat{x}_*, \hat{x}^*]$. The intervals with the highest success in compatibility appear to be those with $0.6 < \hat{x}_* < 0.8$ and $\hat{x}^* < 200$, which give moderate overestimates of the interval of power law validity. In particular, the compatibility of bounds seems to be most sensitive to the value of $\hat{x}_*$, presumably because it affects more severely the number of data points falling within the putative power law interval $[\hat{x}_*, \hat{x}^*]$.

Appendix C. Comparison with Another Variation of KS Method

In this section, we make some remarks concerning the relation of our proposed apKS method for identifying bounded power law regions with the method proposed in [1, 22], wherein the interval of the power law fit is selected as the interval with the largest number of data points or log-range such that the MLE power law fit passes a parametric bootstrap validation test based on KS distance (at some prescribed p-value). For shorthand, we will refer to the method of [1,22] as the *maximal validated interval* KS (mviKS) method. More specifically, selection of the validated power-law

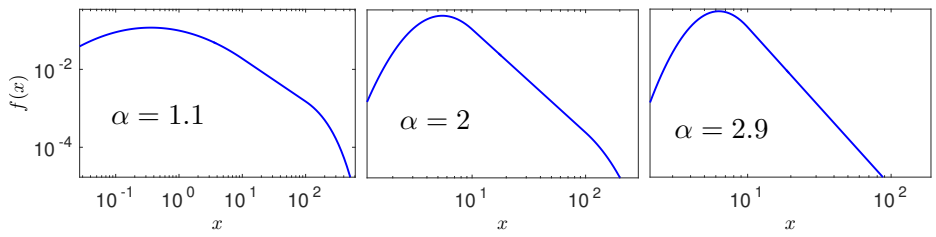

Figure C1. EPL3 distributions used to generate the random samples: $\alpha = 1.1, 2$ and 2.9 from left to right.

interval with the largest number of data points will be referred to as the mviKS-N method and selection of the validated power-law interval with the largest log-range will be referred to as the mviKS-R method. In [23], an idea is developed similar to the mviKS method for determining the bound of a power law tail, but with the p-value computed from a novel "ring test" rather than from a bootstrap. While our method has some common goals and philosophy with these approaches, we would suggest some relative advantages of the apKS method.

First of all, as discussed in Subsection 4.3, the apKS method balances the quality of fit (as measured by the KS metric) and size of the interval in the selection of the power law interval, whereas the mviKS approach makes no reference to the actual quality of fit in the selection of the power law fit, other than it passing some validation criterion. Secondly, both methods use a validation measure (the p-value of the semiparametric or parametric bootstrap) in the actual selection of the power law model, which is a bit statistically unorthodox. While this validation measure plays a primary role in the mviKS method (by defining the set of candidate intervals, to be optimized by range or number of data points), the validation measure is used somewhat more indirectly in the apKS method in the choice of the penalty coefficient (a tuning parameter) in the penalized KS metric.

The mviKS method is also relatively computationally expensive because it must validate all candidate intervals in order to choose the largest validated interval, and the validation step is the most time consuming part of the power law fitting procedure. To compare the computational cost with the apKS method, recall from Subsection 4.3 that we consider candidate power law intervals whose endpoints are selected from a certain number (m) of candidate data points in each decade of the data range and also by only testing intervals that are larger than a decade long. Suppose a sample $X_{1:n}$ is given whose elements span a range of d decades. This would give approximately $(md - m)(md - m + 1)/2$ power law fit and validation computations for the mviKS method. On the other hand, the apKS method requires at most 14 fit-and-validation steps (Subsection 3.3), and in fact usually requires no more than 5 such computations. So with $m = 10$ candidate data points in each decade and $d = 3$ decades of data, the mviKS method would require about 210 fit-and-validation steps, a forty-fold increase in cost relative to the apKS method. Of course these computational costs are no great consideration if only a single sample is considered, but could be relevant if applied to a large collection of experimental or computational data sets.

We turn next to the central concern, which is the relative accuracy of the methods. Numerically, we find the apKS method to be more accurate than the mviKS method in estimating the power-law exponent as well as the power-law interval. The apKS method searches for a power-law interval that is longer than a certain length threshold. The default choice is a decade long, that is intervals $[x_*, x^*]$ such that $x^*/x_* > 10$. Though the mviKS method as described in [1,22] does not make any remarks on any

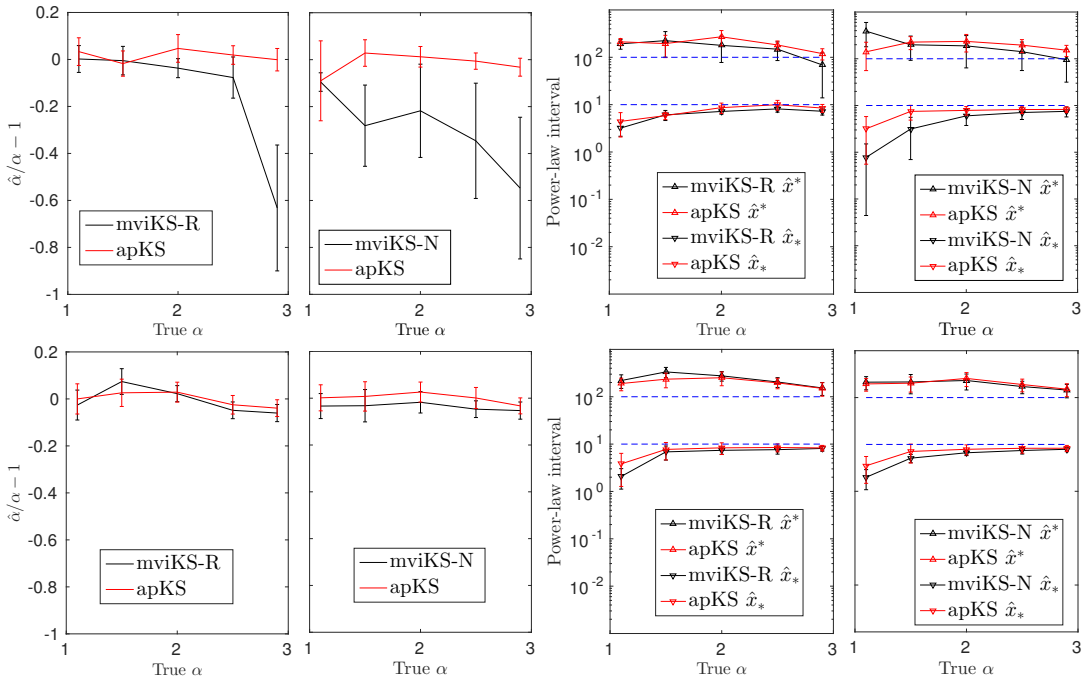

Figure C2. The apKS method and two versions of the mviKS method (-R and -N) are applied to 100 random samples of size $n = 10^3$ from the EPL3 distribution with power law exponents $\alpha = 1.1, 1.5, 2, 2.5$ and 2.9 over the interval $[x_*, x^*] = [10, 100]$. (Top) Comparisons of apKS and original mviKS methods. (Bottom) Comparisons of apKS and mviKS both restricted to consideration of power-law intervals at least a decade long. (Left) Comparisons of relative bias and standard error of power law exponent estimates $\hat{\alpha}$ by the mviKS and by the apKS methods. (Right) Comparisons of absolute bias and standard error of power law interval bound estimates $[\hat{x}_*, \hat{x}^*]$.

length threshold for candidate power-law interval, we will compare the apKS against the original mviKS method and also against the mviKS method that searches for power-law intervals among those which are at least a decade long.

We can see from the top panels in Figure C2 that, for the EPL3 test distribution (Appendix D.4) illustrated in Figure C1, that the mviKS method appears to be generally estimating the upper bound parameter x^* with more variability than the apKS method, though in some cases with less bias. The lower bound x_* to the power law region is estimated with comparable quality by the apKS and mviKS-R methods, but with significantly greater bias and error by the mviKS-N method. Most seriously, we see that the power law exponent α is estimated by the apKS method with substantially less bias and standard error than the mviKS-R method when $\alpha \geq 2.5$ and than the mviKS-N method when $\alpha \geq 1.5$.

An apparent contradiction in the results shown in the top half of Figure C2 is that the mviKS methods seem to be estimating the bounds of the power law interval more accurately as α increases, but the estimation of the power law exponent deteriorates for both mviKS-R and mviKS-N methods. The resolution is that the estimation of the power law exponent α becomes more sensitive to the estimated lower bound $\hat{x}_*$ as α increases, due to the sharper transition between the non-power-law core and power-law region (Figure C1), and the estimation of $\hat{x}_*$ by the mviKS methods is not improving sufficiently rapidly as α increases to permit a reliable estimator $\hat{\alpha}$ of the power law exponent.

When the two approaches are compared, with both restricting power law fits to intervals of at least one decade in length, the apKS method has just slightly less bias and comparable standard error relative to the mviKS methods, as we see in the bottom half of Figure C2. Even though the mviKS methods are trying to find the largest or most data-rich intervals with a good power law fit, on some random samples this will drive the selection of short putative power law intervals into the non-power-law core and produce correspondingly poor bound estimates. The preclusion of consideration of intervals less than a decade long then simply forces the mviKS methods for these random samples to report no suitable validated power law fit, removing the poor bound estimates reported for that random sample from consideration in Figure C2. Though this results in power law interval bound estimations of comparable quality to those of the apKS method, the computational time required by the apKS method is considerably less than that for the mviKS methods.

Appendix D. Probability density functions used for simulation studies

In this appendix, we define the probability density functions used to generate random samples in order to test the KS and apKS methods for power law fitting. We label the three distributions with a bounded power law region as EPLk, where EPL denotes "exact power law" and the index $k \in \{1, 2, 3\}$. All model probability distributions have an analytically invertible CDF, allowing easy simulation via the inverse transform method.

D.1. Power law tail model

This distribution is an exponential for $0 < x < x_*$ and a power law for $x \geq x_*$.

We define the probability density function (2) where the constants A, C, and β are chosen so that

(1) $\int_0^\infty f(x)dx = 1$ (α must be greater than 1), that is
   - $A\left(\frac{e^{-\beta x_*}}{-\beta} + \frac{1}{\beta}\right) + C\left(-\frac{x_*^{1-\alpha}}{1-\alpha}\right) = 1.$

(2) $f(x)$ is continuous and has a continuous slope at $x = x_*$. That is
   - $Ae^{-\beta x_*} - Cx_*^{-\alpha} = 0$ and $\alpha x_*^{-1} = \beta$.

The only free parameters are the power law exponent α and power law tail lower bound x_*; the constants A and C are then uniquely determined by the above two conditions.

D.2. EPL1 distribution

This distribution is power law for $x_* \le x \le x^*$ and has zero density elsewhere, and is more commonly referred to as an "upper-truncated Pareto" distribution [12,18].

We define the probability density function:

$$f(x) = Cx^{-\alpha},\ x_* \le x \le x^*. \tag{D1}$$

The power law exponent $\alpha > 0$ is arbitrary and the normalization constant is

$$C = \frac{1-\alpha}{x^{*1-\alpha} - x_*^{1-\alpha}}.$$

D.3. EPL2 distribution

This distribution is an exponential for $0 < x < x_*$; a power law for $x_* \le x \le x^*$; and an exponential for $x > x^*$.

We define the probability density function

$$f(x) = \begin{cases} Ae^{-\beta x} & : 0 < x < x_* \\ Cx^{-\alpha} & : x_* \le x \le x^* \\ Ae^{-\beta x} & : x^* < x, \end{cases} \tag{D2}$$

where we choose the constants β, A and C so that

(1) $\int_0^\infty f(x)dx = 1$, that is
   - $A \cdot \frac{e^{-\beta x_0} + e^{-\beta x^*} - e^{-\beta x_*}}{\beta} + C \cdot \frac{x^{*-\alpha+1} - x_*^{-\alpha+1}}{-\alpha+1} = 1.$

(2) $f(x)$ is continuous but does not necessarily have a continuous slope at $x = x_*$ or $x = x^*$. These two conditions are reduced to the following equations:
   - $A \cdot e^{-\beta x^*} - C \cdot x^{*-\alpha} = 0,$
   - $\beta = \alpha \cdot \dfrac{\log x^* - \log x_*}{x^* - x_*}.$

The only free parameters are the power law exponent α and power law bounds x_* and x^*; the constants A, β, and C are then uniquely determined by the above three conditions.

D.4. EPL3 distribution

This distribution is a log-normal for $0 < x < x_*$; a power law for $x_* \le x \le x^*$; and an exponential for $x > x^*$.

Let $f_{\mathrm{LN}}(x)$ be the probability density function and $F_{\mathrm{LN}}(x)$ the cumulative distribution function for the log-normal distribution with mean μ and standard deviation σ.

We then define the probability density function:

$$f(x) = \begin{cases} f_{\mathrm{LN}}(x) & : 0 < x < x_* \\ Cx^{-\alpha} & : x_* \leq x \leq x^* \\ Ae^{-\beta x} & : x^* < x \end{cases} \tag{D3}$$

The parameter μ is chosen as 1. This seems to guarantee that the part of the distribution preceding the power law interval is visibly a non-power law. Then the parameters L, A, C, σ, and β are numerically calculated so that the following are satisfied:

(1) $\int_0^\infty f(x)dx = 1$, that is
   - $L \cdot F_{\mathrm{LN}}(x_*) + C\left(\frac{x^{*1-\alpha} - x_*^{1-\alpha}}{1-\alpha}\right) + A\left(\frac{e^{-\beta x^*}}{\beta}\right) = 1.$

(2) $f(x)$ is continuous and has a continuous slope at $x = x_*$ and $x = x^*$. These two conditions reduce to:
   - $L \cdot f_{\mathrm{LN}} f(x_*) = Cx_*^{-\alpha}$ and $L \cdot f'_{\mathrm{LN}}(x_*) = C(-\alpha)x_*^{-\alpha-1}$.
   - $Cx^{*-\alpha} = Ae^{-\beta x^*}$ and $\alpha x^{*-1} = \beta$.

The only free parameters are the power law exponent α and power law bounds x_* and x^*.