

A Graphical Test for Local Self-Similarity in Univariate Data

Rakhee Dinubhai Patel Frederic Paik Schoenberg

Department of Statistics
University of California, Los Angeles
Los Angeles, CA 90095-1554

Rakhee Dinubhai Patel is Doctoral Candidate, UCLA Department of Statistics, 8125 Math Sciences Bldg., Box 951554, Los Angeles, CA 90095-1554 (e-mail: rakhee@stat.ucla.edu), and Frederic Paik Schoenberg is Professor, UCLA Department of Statistics, 8125 Math Sciences Bldg., Box 951554, Los Angeles, CA 90095-1554 (e-mail: frederic@stat.ucla.edu).

Abstract

The Pareto distribution, or power-law distribution, has long been used to model phenomena in many fields, including wildfire sizes, earthquake seismic moments and stock price changes. Recent observations have brought the fit of the Pareto into question, however, particularly in the upper tail where it often overestimates the frequency of the largest events. This paper proposes a graphical self-similarity test specifically designed to assess whether a Pareto distribution fits better than a tapered Pareto or another alternative. Unlike some model selection methods, this graphical test provides the advantage of highlighting where the model fits well and where it breaks down. Specifically, for data that seem to be better modeled by the tapered Pareto or other alternatives, the test assesses the degree of local self-similarity at each value where the test is computed. The basic properties of the graphical test and its implementation are discussed, and applications of the test to seismological, wildfire, and financial data are considered.

Keywords: Goodness-of-fit testing, Pareto distribution, power-law, self-similarity, tapered Pareto distribution.

1 Introduction

1.1 Characterizing distributions of variables exhibiting seemingly power-law behavior

Many phenomena observed in a wide variety of scientific disciplines tend to exhibit heavy tails and are commonly modeled using power-law distributions such as the Pareto (see e.g. ch.9 of Johnson et al. 2005 for a review). However, recent work in several fields has shown that alternative distributions may be more appropriate, as in many cases the Pareto tends to overestimate the density in the upper tail. In seismology, for instance, earthquake sizes are typically measured in terms of scalar seismic moments, which have been commonly modeled to follow the Pareto distribution; this corresponds to the exponential Gutenberg-Richter law in terms of earthquake magnitudes (Vere-Jones 1992, Kagan 1994, Utsu 1999). However, several studies have shown that for many local and global catalogs, the tapered Pareto distribution provides significantly improved fit to the seismic moment distribution (Kagan 1993, Jackson and Kagan 1999, Kagan and Schoenberg 2001, Vere-Jones et al. 2001). In addition, earthquake inter-event times and distances have been conventionally modeled as Pareto distributed (Ogata 1998, Utsu 1999), but perhaps may also be better described by the tapered Pareto distribution (Schoenberg et al. 2008). Similarly, wildfire sizes have been commonly modeled as following the Pareto distribution, but were shown in Cumming (2001) and Schoenberg et al. (2003) to follow a truncated or tapered Pareto distribution significantly more closely. In finance, normalized stock returns have long been posited to exhibit power-law behavior consistent with the Pareto distribution, but recently Malevergne et al. (2005, 2006) and Pisarenko and Sornette (2006) have shown that the upper tail of stock returns decay significantly more quickly than the

power-law would suggest, though more slowly than some other heavy-tailed distributions such as the stretched exponential. In this paper, we investigate whether the changes in stock prices can also be described using the tapered Pareto distribution. Because of the extensive range of applications to which this model selection problem arises, it is of interest to develop a graphical tool designed to examine the fits of these distributions and to aid in the discrimination between competing models for these types of heavy-tailed distributions.

In this paper, we propose a graphical test that can be used to characterize whether a distribution is well described by the Pareto distribution or if it more closely follows an alternative such as the tapered Pareto distribution. In particular, the test will focus in detail on a special property of the Pareto distribution and power-law distributions in general: self-similarity. By examining this property in a graphical fashion, one may visualize trends in the data in a convenient way, evaluate the level of self-similarity and fractality in the data, and determine for what *ranges* a particular heavy-tailed model fits well and where it breaks down.

1.2 Model Selection

1.2.1 Pareto Distribution

Vilfredo Pareto originally introduced the *power-law* or *Pareto* distribution to describe the differential allocation of wealth (Pareto 1897). Since that time, a host of other observable phenomena have been characterized as following a Pareto distribution (see e.g. Mandelbrot 1963, Vere-Jones 1992, Utsu 1999). The key feature of the Pareto distribution is its extremely heavy tail. Another defining property of power-law dis-

tributions is that they exhibit self-similarity consistent with certain fractal and diffusion processes. The probability density function of the Pareto distribution is given by $f(x) = \beta\alpha^\beta/x^{\beta+1}$, $\alpha \leq x < \infty$, where $\alpha > 0$ is a scale parameter and $\beta > 0$ serves as a shape parameter. The cumulative distribution function is given by $F(x) = 1 - (\alpha/x)^\beta$. Thus the logarithm of the survival function is a linear function of $\log(x)$, with slope β .

The tails of the Pareto distribution are very heavy, and the relatively high frequency of extremely large events implied by the fitted Pareto distribution is often contradicted by the data and in some cases by basic physical principles (Knopoff and Kagan 1977; Sornette and Sornette 1999). To provide a more sensible and better-fitting alternative, several other distributions with less heavy tails have been suggested, such as the lognormal, half-normal, exponential, Frechet, truncated Pareto and tapered Pareto, among others (see e.g. Kagan and Schoenberg 2001). Figure 1(a) shows that for the sizes of Los Angeles County wildfires, the Pareto distribution (represented by the dashed line) overestimates the frequency of the largest events. Schoenberg et al. (2003) have shown that the tapered Pareto distribution (represented by the dotted-dashed curve) provides a much better fit for these data. Similarly, superior fit appears to be offered by the tapered Pareto distribution to the inter-event times of Southern California earthquakes, as shown in Figure 1(b). The wildfire data in Figure 1(a) are based on fires burning at least 0.405 sq. kilometers (100 acres) in Los Angeles County from 1950-2000, recorded by the Los Angeles County Department of Public Works and the Los Angeles County Fire Department. For further details about this data set as well as images of the centroid locations of these wildfires, see Peng et al. (2005). The earthquake data in Figure 1(b) were compiled by the Southern California Earthquake Center (SCEC) and consist of the 6,796 earthquakes of magnitude at least 3.0 occurring between January 1, 1984 and June 17, 2004 in a rectangular area around

Los Angeles, California, between longitudes -122 and -114 and latitudes 32 and 37 (approximately 733km by 556km).

1.2.2 Tapered Pareto Distribution

The first person to notice the shortcomings of the upper tail of the Pareto distribution was apparently Vilfredo Pareto himself, who suggested a modified alternative now known as the tapered Pareto distribution (Pareto 1897). As Figure 1 indicates, the tapered Pareto behaves similarly to the Pareto in the lower quantiles, but then decays more quickly (in fact, exponentially) in the upper tail. The probability density function of the tapered Pareto distribution is given by

$$f(x) = \left(\frac{\beta}{x} + \frac{1}{\theta}\right) \left(\frac{\alpha}{x}\right)^\beta \exp\left(\frac{\alpha - x}{\theta}\right), \alpha \leq x < \infty$$

and the cumulative distribution function is given by

$$F(x) = 1 - \left(\frac{\alpha}{x}\right)^\beta \exp\left(\frac{\alpha - x}{\theta}\right).$$

The parameter θ governs the shape of the upper tail. As θ approaches infinity, the tapered Pareto approaches the Pareto distribution, and as β approaches 0, the tapered Pareto approaches the exponential distribution. For events following the tapered Pareto distribution, the survival function plotted on a log-log scale appears nearly linear for smaller events, but gradually decays in an exponential fashion, as seen in Figure 1.

1.2.3 Existing model selection methods

Though there are several methods with which one can distinguish between two distributions (see e.g. Burnham et al. 2002), this research aims to capture elements that the typical test might not. Since the Pareto distribution is nested within the tapered Pareto, the likelihood ratio statistic $LR = \sup\{L_0\}/\sup\{L_1\}$ provides an optimal and asymptotically efficient method to select the appropriate distribution for a given set of data (see e.g. Bickel and Doksum 2007). Other common goodness of fit methods include Akaike's information criterion or AIC (Akaike 1974) and the Bayesian information criterion or BIC (Schwarz 1978) which measure the tradeoff between bias and variance when fitting a model by imposing penalties for additional parameters. These are given by $AIC = -2\ln(L) + 2k$ and $BIC = -2\ln(L) + k\ln(n)$, where each statistic aims to prevent overfitting by penalizing the model for fitting more free parameters. Residual analyses such as comparing mean squared errors or sums of squared residuals and cross-validation techniques (see e.g. Stone 1977, Geisser 1993) are other common model selection methods that may be used to distinguish between the Pareto and tapered Pareto, but like likelihood statistics, their utility in practice may be limited by the fact that they offer only a single numerical summary of the overall fit.

A goal of this paper is to explore graphical tests that allow one to compare features of the empirical distribution of the data to those of a model such as the Pareto distribution, and also offer a means to assess for what parts of the distribution the Pareto model might fit well and where it might break down. In particular, the proposed graphical representation examines the property of self-similarity, which is a characteristic of the Pareto distribution, in detail.

1.3 Self-similarity of Pareto distribution

Mandelbrot (1967) very simply described self-similarity as an instance where “each portion can be considered a reduced-scale image of the whole.” The self-similarity or fractal property of the Pareto distribution lies in the fact that the survival function, as exhibited in Figure 1, is linear when plotted on log-log scale. As a result, the shape of the density function (on log-log scale) looks essentially the same at all scales. In other words, the density at any value x , relative to the density at any other value y , only depends on the ratio of x to y and not the values themselves. The slope of the log survival function is the shape parameter β , which is often called the fractal dimension of the data. The test statistic constructed in this paper will focus extensively on this self-similarity property of the Pareto distribution.

2 Graphical self-similarity test

2.1 Test statistic for local self-similarity

2.1.1 Ratio of densities under the Pareto null hypothesis

The self-similarity of the Pareto distribution stems from the fact that the ratio of the Pareto density for any two given values of x is given by

$$\frac{f(x)}{f(kx)} = k^{\beta+1}, \forall x, k > 0,$$

which is independent of the value of x . One may thus standardize this ratio, for any x and any k , yielding

$$g_k(x) := \frac{1}{k} \left(\frac{f(x)}{f(kx)} \right)^{\frac{1}{\beta+1}} = 1, \forall x, k > 0.$$

This suggests estimating $g_k(x)$ given data and comparing the result with unity as a means of obtaining a representation of the self-similarity of the data for each value of x and k . That is, for a given data set of independent observations $\{x_1, x_2, \dots, x_n\}$, we estimate $g_k(x_i)$ for each $x_i, k > 0$ ($i = 1, \dots, n$) and interpret the result, in relation to unity, as a measure of the local self-similarity of the data.

2.1.2 Ratio of densities under the tapered Pareto null hypothesis

If instead one seeks to test data against a tapered Pareto null hypothesis, it is clear that the corresponding ratio of tapered Pareto densities will be dependent on x because of the lack of self-similarity of this distribution:

$$\frac{f(x)}{f(kx)} = k^{\beta+1} \left(\frac{\beta\theta + x}{\beta\theta + kx} \right) \exp \left(\frac{x(k-1)}{\theta} \right), \forall x, k > 0.$$

Therefore, the corresponding value of $g_k(x)$, for tapered Pareto distributed random variables, is given by

$$g_k(x) = \frac{1}{k} \left(\frac{f(x)}{f(kx)} \right)^{\frac{1}{\beta+1}} = \left(\frac{\beta\theta + x}{\beta\theta + kx} \right)^{\frac{1}{\beta+1}} \exp \left(\frac{x(k-1)}{\theta(\beta+1)} \right), \forall x, k > 0,$$

and because of the exponential component of the tapered Pareto distribution, $g_k(x)$ depends on both x and k . This suggests estimating the local self-similarity $g_k(x)$ using independent observations $\{x_1, x_2, \dots, x_n\}$, and assessing its agreement with the

local self-similarity for the tapered Pareto distribution.

2.2 Estimating local self-similarity

A natural estimate of $g_k(x)$ is given by

$$\hat{g}_k(x) = \frac{1}{k} \left(\frac{\hat{f}(x)}{\hat{f}(kx)} \right)^{\frac{1}{\hat{\beta}+1}}, \quad \forall x, k > 0,$$

where \hat{f} is a non-parametric density estimate, and $\hat{\beta}$ is an estimate of the parameter β . For instance, one may obtain non-parametric density estimate $\hat{f}(x)$ using kernel density estimation and estimate $\hat{\beta}$ by maximum likelihood, as described briefly below, in order to obtain a local self-similarity test statistic $\hat{g}_k(x)$ for any positive values of x and k .

2.2.1 Density estimation

The list of useful non-parametric density estimators is extensive (see e.g. Silverman, 1986). The standard kernel density estimator takes the form

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K \left(\frac{x - x_i}{h} \right),$$

where n is the sample size of the data, h is the bandwidth, K is a kernel function and $\{x_1, \dots, x_n\}$ are the observations. A particularly important determination in kernel density estimation is the choice of bandwidth. Because a fixed (or global) bandwidth may result in undersmoothing areas where data are sparse and oversmoothing areas where data are concentrated, variable bandwidth methods may be especially appropriate for the purpose of estimating a heavy tailed distribution, to account for the

high degree of irregularity in the data in terms of the distance between sequential ordered observations (see e.g. Silverman, 1986).

In particular, one could apply the commonly used two-stage adaptive kernel method (see e.g. Abramson 1982, Silverman 1986, Fox 1990, Salgado-Ugarte et al. 1993), where the bandwidth varies with each observation in the data. The first stage in this method is to estimate a pilot density $\hat{f}(x_i)$ for each observation in the data using, for instance, a kernel estimator with fixed bandwidth. To select an appropriate pilot bandwidth, one could use Silverman's rule of thumb (Silverman 1986) or Scott's variation of Silverman's rule (Scott 1992). A host of other alternatives such as the Sheather-Jones direct plug-in approach (Sheather and Jones, 1991) are also appropriate. Using these pilot density estimates $\hat{f}(x_i)$, bandwidth adjustment factors λ_i can be calculated for each observation in the data as follows:

$$\lambda_i = \left(\frac{G}{\hat{f}(x_i)} \right)^{0.5},$$

where G is the geometric mean of the initial fixed bandwidth density estimates $\hat{f}(x_i)$, $i = 1, \dots, n$. In the second stage of the estimation, the variable bandwidths and resulting kernel density estimates can be calculated as follows:

$$\hat{f}_A(x) = \frac{1}{nh} \sum_{i=1}^n \frac{1}{\lambda_i} K \left(\frac{x - x_i}{\lambda_i h} \right),$$

where h is the fixed bandwidth used in the first stage of the density estimation. The variance of this particular density estimate is given by

$$V \left(\hat{f}_A(x) \right) = \frac{f(x)}{nh\lambda(x)} \int_{-\infty}^{\infty} [K(s)]^2,$$

where $K(s)$ represents the kernel function (see e.g. Scott 1992).

Note that the choice of kernel function (Gaussian, Epanechnikov, etc.) may also be influential in determining the quality and variability of the density estimates. Further, because data modeled by the Pareto distribution generally achieve a maximum density at some minimum boundary of the data, one may consider using reflection across this lower boundary (see e.g. Silverman 1986). For heavy-tailed distributions with sparse data in the upper tail, the variance of the density estimates generally increases with the value of x (as the concentration of data decreases). Thus, obtaining a density estimate with low variance is particularly important for heavy-tailed distributions.

2.2.2 Maximum likelihood estimation

The fractal dimension parameter, β , used in the local self-similarity test statistic may be estimated using maximum likelihood. Under a Pareto null hypothesis, the maximum likelihood estimate of the shape parameter or fractal exponent β has a closed form solution given by

$$\hat{\beta} = \frac{n}{\sum_{i=1}^n \log(x_i/\hat{\alpha})},$$

where n is the sample size of the data and $\hat{\alpha} = \min_i x_i$ is the maximum likelihood estimate of the scale parameter α . The variance of $\hat{\beta}$ also has a closed form solution (see e.g. Johnson et al. 1994) :

$$V(\hat{\beta}) = \frac{n^2\beta^2}{(n-2)^2(n-3)}.$$

For a tapered Pareto null hypothesis, for example, the estimate of β has no closed form solution and must be solved by optimizing the following log-likelihood function with respect to both $\hat{\beta}$ and $\hat{\theta}$:

$$-\sum_{i=1}^n \left(\log \left(\frac{\hat{\beta}}{x_i} + \frac{1}{\hat{\theta}} \right) \right) - \hat{\beta} n \log(\hat{\alpha}) + \hat{\beta} \sum_{i=1}^n \log(x_i) - \frac{\hat{\alpha} n}{\hat{\theta}} + \frac{1}{\hat{\theta}} \sum_{i=1}^n x_i,$$

where n is the sample size and $\hat{\alpha} = \min_i x_i$ is the maximum likelihood estimate of the scale parameter α . As usual, the variance of tapered Pareto maximum likelihood estimates can be obtained from the diagonal elements of the inverse Hessian matrix from the above likelihood function. For other null distributions, parameter estimates can be calculated in a similar fashion.

2.3 Simulation studies

The performance of the self-similarity test can be assessed graphically using simulations.

Figure 2 illustrates empirical histograms and survival functions along with theoretical densities and survival functions fit by maximum likelihood for Pareto and tapered Pareto simulations. The Pareto and tapered Pareto random variables are generated using inverse transform sampling, using sample sizes of $n = 1000$ and the parameters $\alpha = 1$ and $\beta = 2$, as well as $\theta = 3$ for the tapered Pareto simulation. Note that the difference between the two distributions is extremely difficult to distinguish from the histograms alone. However, the log survival plots in Figure 2 show some differences between the two distributions in the upper tail. As expected, the empirical survival of the Pareto simulation appears rather linear, and thus self-similar, throughout the

entire range of its survival plot, while the empirical survival of the tapered Pareto simulation shows some noticeable curvature in the upper tail.

2.3.1 Graphical representation of self-similarity

One may inspect the statistic $\hat{g}_k(x)$ for various choices of k and x in order to assess the degree of self-similarity for a given simulated or real data set. As an illustration, consider the two Pareto and tapered Pareto simulations shown in Figure 2, and imagine testing them against a Pareto null hypothesis. For each simulated data set, one may choose equally spaced grid values over the range of x and estimate the density of each simulation by implementing the two-stage adaptive kernel density estimation with reflection using a Gaussian kernel and Scott's rule to estimate the pilot bandwidth as described in Section 2.2.

The resulting test statistic $\hat{g}_k(x)$ for the Pareto simulation is shown in Figure 3(a). Most of the values are close to unity, as expected. On the other hand, the test statistic for the tapered Pareto simulation, represented by Figure 3(b), seems fairly consistent with the null Pareto distribution up to approximately $x = 6$, for most values of k . However, the self-similarity apparently breaks down in the upper tail, for $x > 6$. This is to be expected since the values in the tapered Pareto simulation are independently drawn from a tapered Pareto distribution with $\theta = 3$, and in general the tapering of the tail in the tapered Pareto distribution becomes especially pronounced for values larger than θ .

2.3.2 Inference

In order to evaluate the distribution and power of the test statistic $g_k(x)$ under a given null hypothesis, one may construct m simulated data sets each of n iid random variables drawn from the desired null distribution. Note that the null distribution will generally be the Pareto distribution, but as mentioned in Section 2.1, one may also investigate the same test statistic under the assumption of another null distribution, e.g. the tapered Pareto, Frechet, etc. For each set of simulated data, $\hat{g}_k(x)$ is calculated over the same grid used for the data. For each pixel in the grid of values for which the test statistic is calculated (i.e. for each x and k), the statistical significance of that pixel for the data is determined using the quantile method, simply by comparing the value of $\hat{g}_k(x)$ for the data with those for the simulations.

As an illustration, Figures 4(a) and 4(b) depict the one-sided p-values corresponding to the local non-self-similarity test \hat{g} for the Pareto and tapered Pareto simulations used in Figures 3(a) and 3(b), respectively. A p-value of 0.03, for instance, means that, for the particular values of x and k in question, $g_k(x) - \hat{g}_k(x)$ for the simulated data set is greater than the 97th percentile of $g_k(x) - \hat{g}_k(x)$ for the m (in this case, 100) simulations generated from the null Pareto distribution.

As seen in Figure 4(a), the simulation drawn from the Pareto distribution shows few significant departures from self-similarity, as expected. The tapered Pareto simulation in Figure 4(b) consists of draws from a tapered Pareto distribution with $\theta = 3$ and shows significant deviation from the Pareto generally beyond $x = 3$ for most values of k as one might expect. Also of note in Figure 4(b) are the nonsignificant values of $\hat{g}_k(x)$ for the largest values of x and for large k . These may be due to the relatively low

power of the graphical test in the upper tail where very few observations are recorded. Thus the sample size and choice of method used for density estimation, particularly the bandwidth, has great influence on the power of the test statistic $\hat{g}_k(x)$.

2.4 Comparison to empirical survival function with confidence bounds

An alternative, commonly used method for assessing goodness-of-fit is to compare the empirical survival function with the proposed (e.g. Pareto) distribution. Confidence bounds, as depicted in Figures 1 and 2, can be constructed using the binomial distribution, since the empirical cumulative distribution function $\hat{F}(x)$ can be represented by A/n , where A is the number of observations less than or equal to x , n is the sample size, and $A \sim \text{Binomial}(n, p = F(x))$.

The survival function has an advantage over $\hat{g}_k(x)$ in that accurate density estimation is not necessary for the estimation of the survival function. However, because the survival function is an aggregate measure, it has comparatively low power at discerning local departures from self similarity. On the other hand, the test statistic $\hat{g}_k(x)$ may be more powerful at detecting lack of fit locally, i.e. for values of x in a particular range. An illustration is given below.

As an illustration, Figure 5 shows the empirical survival function and the graphical test $g_k(x)$ applied to a simulated data set consisting of draws from the Pareto distribution, but with an increased density for $4 \leq x \leq 4.5$ and a density of zero for $5 \leq x \leq 5.5$. Note that the survival function, along with 95% confidence bounds,

does not show any significant lack of fit to the Pareto distribution. The proposed local self-similarity test, $g_k(x)$, rejects the Pareto distribution in the appropriate range, particularly for large values of k . The test may thus be useful in detecting local lack of self-similarity for a particular data set.

Although values of the test statistic $\hat{g}_k(x)$ are correlated, the correlations for various values of x are actually much smaller than the corresponding correlations between survival function estimates. Table 1 shows the correlations between the test statistic $\hat{g}_k(x_i)$ and $\hat{g}_k(x_j)$ for 100 Pareto simulations generated using the parameters $\alpha = 1$, $\beta = 2$ and sample size $n = 1000$, where $k = 0.8$ and the x values are ten equally-spaced points between 3 and 12. Meanwhile, Table 2 represents the correlations of the estimated survival function at the same ten equally spaced values of x for the same 100 Pareto simulations. Similar results are obtained using other values of fixed k for the test statistic.

Note that, for the local self-similarity test $g_k(x)$, different choices of k result in a test of differential power depending on the alternative hypothesis in question. If, for instance, the alternative is that the distribution of a given data set is Pareto except very locally near a given x , as in the simulated example in Figure 5, then k close to 1 will have optimal power, whereas smaller values of k may yield a more powerful test in the case where the data in question deviate from self-similarity over a slightly broader range.

Because one may choose to examine different k for a given value of x , it is important to bear in mind that for a given x , the test statistic $\hat{g}_k(x)$ may be highly correlated

for various values of k , especially for values of k close to one another. For the 100 Pareto simulations used in Tables 1 and 2, the correlations between $\hat{g}_{k_i}(x)$ and $\hat{g}_{k_j}(x)$, where $x = 5$ and each i, j is in the sequence of ten values of k between 0.50 and 0.95, are given in Table 3. Results are similar for other values of x .

3 Application of graphical self-similarity test

3.1 Application to wildfire sizes and earthquake inter-event times

Section 1 introduced two data sets, the sizes of Los Angeles County wildfires burning at least 100 acres (0.405 sq. kilometers) from 1950-2000, and inter-event times between earthquakes of magnitude at least 3.0 occurring between January 1, 1984 and June 17, 2004 in Southern California, which have traditionally been thought to be self-similar in nature, but which have more recently been shown to follow a tapered Pareto distribution more closely (see Schoenberg et al. 2003 and Schoenberg et al. 2008). We can apply the graphical self-similarity test against a Pareto null hypothesis (using the two-stage adaptive kernel method with reflection, a Gaussian kernel and Scott's rule to estimate the pilot bandwidth for the density estimates and maximum likelihood estimates as the parameters for the null distribution) in order to determine which ranges of these data sets are indeed self-similar and which ranges are not. Figure 6 shows the p-values for local non-self-similarity for both data sets.

The survival plot in Figure 1(a) shows that the wildfire size data appears to conform rather closely to the Pareto distribution for burn areas under approximately 75 sq.

kilometers. However, Figure 6(a) shows that the non-self-similarity is significant for values of x between approximately 30 and 50 sq. kilometers at a significance level of 0.05. This is also not depicted by the 95% confidence bounds around the empirical survival and indicates a lack of local self-similarity in this range. Additionally, while Figure 1(a) shows that the self-similarity of the data appears to subside for wildfires burning greater than 75 sq. kilometers or so, the significance region using the test in Figure 6(a) does not reappear until after 150 sq. kilometers. The lack of statistical significance at large wildfire sizes appears to be attributable to the relatively small sample size ($n = 594$) and the scarcity of data in the extreme upper tail, as there were only 15 wildfire records larger than 75 sq. kilometers in the data set, and the power of the test is relatively low in such situations. A physical explanation for the deviation from the Pareto for large events may be the effect of human efforts to suppress wildfires that have spread to larger sizes.

The survival plot in Figure 1(b) shows that the earthquake inter-event time data do not appear to be self-similar for essentially the entire range of the data. The p-values in Figure 5(b) show that the local non-self-similarity appears to be statistically significant (using a significance level of 0.05) for inter-event times greater than five days. On the other hand, though the survival plot does not seem consistent with the Pareto distribution for inter-event times smaller than five days, the graphical test \hat{g}_k does not show significant departures from self-similarity for this range of the data. The tapered Pareto maximum likelihood estimate of β for this data is only 0.0393, so that the fitted tapered Pareto distribution is similar to an exponential distribution, but the apparent lack of non-self-similarity in the lower tail of the data seems consistent with the general properties of a tapered Pareto distribution. However, when dealing with estimates of β this small, Pareto simulations used for inference can generate random

variables spread across extremely large ranges, which can make density estimation quite difficult. Thus, for small values of x , the non-significant p-values may indicate consistency with self-similarity or may be a result of unstable density estimates in this range.

3.2 Application to normalized stock returns

Modeling the distribution of stock price returns is a key piece in any risk analysis for a publicly traded company. Stock price returns have often been modeled as following a power-law distribution (see e.g. Mandelbrot 1963, Cont 2001, Barari and Mitra 2008, Plerou and Stanley 2008, Gu and Zhou 2009), though several authors have noted recently that the upper tail of the distribution is typically not well described by a power-law (Malevergne et al. 2005, Malevergne et al. 2006, Pisarenko and Sornette 2006). As with the natural processes observed earlier, we propose the tapered Pareto may also provide a suitable fit to these stock price returns and will examine the self-similarity of the data (or lack thereof) using the test statistic $\hat{g}_k(x)$ and corresponding graphics outlined in the previous Section.

3.2.1 Stock price data

Consider an examination of the historical split and dividend-adjusted daily closing prices of Google, Inc., from the date of its initial public offering on August 19, 2004 through April 30, 2010. Following Kaizoji (2004), we will consider modeling the absolute values of the logarithmic returns (1,433 observations), in order to capture the full volatility in the market.

3.2.2 Applying the self-similarity test under the Pareto null hypothesis

Figure 7 shows the results of the graphical self-similarity test to determine whether a Pareto distribution, fit by maximum likelihood to the data, provides a reasonable approximation to the data. Figure 7(a) displays the test statistic $\hat{g}_k(x)$ for the Google return data and Figure 7(b) illustrates the p-values of the pixels compared to 100 simulations of the Pareto null hypothesis fit to the data by maximum likelihood. The density estimates used to calculate the test statistic are calculated, once again, using the two-stage adaptive method using a Gaussian kernel and Scott's rule (Scott, 1992) to determine the pilot bandwidth along with reflection. The test clearly indicates significant departures between the Pareto distribution and the Google returns. Specifically, the data set displays significant non-self-similarity for returns greater than 5% in absolute value, using a significance level of 0.05. The tapered Pareto maximum likelihood estimate of θ is $\hat{\theta}_{tap} = 1.61\%$ and we can see that mainly returns beyond this level tend to deviate significantly from self-similarity. As with the earthquake inter-event times, the tapered Pareto maximum likelihood estimate for β is close to zero ($\hat{\beta}_{tap} = 8.74 * 10^{-9}$), indicating a nearly exponential distribution, but the test fails to reject self-similarity for much of the lower tail of the data, suggesting that a tapered Pareto may be more appropriate. Like in the previous example, an alternative explanation could be unstable density estimates due to the small β . It should be noted that the results of the local self-similarity test using other methods of obtaining density estimates were slightly different. The general departure from self-similarity is still apparent for the same range of x seen in Figure 7 using the other methods, but for some values of k and x the significance levels differ.

3.2.3 Applying the self-similarity test under the tapered Pareto null hypothesis

We can also use the local self-similarity test statistic $\hat{g}_k(x)$ to test data against the tapered Pareto distribution. Figure 8(a) represents the expected value of $g_k(x)$ for the tapered Pareto distribution, and is seen to approximate the test statistic $\hat{g}_k(x)$ for the Google data, shown in Figure 7(a), better than the Pareto. Figure 8(b) illustrates the p-value for $\hat{g}_k(x)$ for each pixel versus 100 simulations of the tapered Pareto distribution fitted to the Google data by maximum likelihood. In this case, there are only a few significant departures from the tapered Pareto around absolute returns of 4% to 5% using a significance level of 0.05. It is plausible that these could be due to chance. Note that since many tests are performed simultaneously, by chance one might expect some Type 1 errors. In practice one may use a Bonferroni correction to account for this, if such errors are to be avoided. Thus, the tapered Pareto seems to provide a reasonable fit to the majority of the data.

3.2.4 Discussion

The application of our test statistic $\hat{g}_k(x)$ on Google's normalized stock returns points out significant departures from self-similarity and suggests that these returns are not well described by the Pareto distribution for a large range of the data. However, it should be noted that these results do not preclude the possibility that small subintervals of the data, when inspected separately, may individually exhibit self-similarity. Note also that since Google is a relatively young stock, the recent market volatility should be expected to contribute more to the upper tail than one would anticipate for an older stock with thousands of observations. Our investigations using alternative data sets, including other stocks that have been traded for many decades, as well as

mutual funds and indexes, seem to suggest that Google is not at all unusual; returns of most other public securities appear to be significantly non-self-similar for the bulk of the range of the data, and tend to be far better described by the tapered Pareto instead of the Pareto distribution.

4 Summary

The graphical test presented in this paper is a convenient way to examine local self-similarity, particularly in data thought to exhibit power-law behavior. Not only can the test confirm departures of data from a theoretical Pareto distribution as seen in a survival plot, but it can also detect pockets of non-self-similarity where obvious departures in the survival plot are not evident. On the other hand, the test may also identify areas in the data where non-self-similarity cannot be confirmed, even if the survival plot does not show a close fit of the data to the theoretical Pareto. In the lower tail, this latter result could be due to either chance or the fact that even a tapered Pareto distribution is roughly Pareto in the lower tail where the bulk of the data resides. Even if the data appears to fit a tapered Pareto distribution with small values of β and θ and the survival plot suggests a nearly exponential fit throughout the range of the data rather than a Pareto fit in the lower quantiles, the test seems to be quite powerful in the lower quantiles of the data provided that the density estimates are stable. In the upper tail, however, where data are far more sparse, conventional density estimates are typically highly variable.

The main shortcoming of the proposed test seems to be that the choice of density estimation technique greatly influences the test statistic $\hat{g}_k(x)$. This choice is often

rather arbitrary, and while objective and adaptive methods are available, especially for estimating bandwidths in smoothing procedures, it is important to bear in mind that the particular choice of density estimation method may have a large impact on the test statistic. This particularly comes into play in the upper quantiles where data can be quite sparse or in cases where density needs to be estimated over an extremely large range of data. Thus the test statistic may lack power in the upper tail, especially for relatively small data sets.

The results and methods used in this paper can extend to many other applications. For instance, the self-similarity of several other phenomena, such as earthquake seismic moments and distances between main shocks and aftershocks, among other natural occurrences can be further examined using the graphical test. In finance, in addition to exploring the distributions of various individual stocks, which may vary depending on the age of the stock and corresponding market conditions, one can apply the test to overall market indicators such as exchange-traded funds and stock indexes.

An important direction for future work is the investigation of the power of the test against various specific alternatives such as the exponential, Frechet and lognormal distributions, for example, or mixtures or sums of Pareto distributions with different parameters, particularly in the upper tail where the power is relatively low. This also involves looking for the most powerful choice of k for a particular alternative hypothesis. In addition, the investigation of means of mitigating the covariance between two values of the test statistic $\hat{g}_k(x)$, for different values of k and x , especially in the context of constructing appropriate individual and simultaneous confidence bounds, are important subjects for future research.

5 References

- Abramson, I. S. (1982), "On bandwidth variation in kernel estimates a square root law," *Annals of Statistics*, 10(4), 1217-1223.
- Akaike, H. (1974), "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, 19 (6), 716-723.
- Barari, M. and Mitra, S. (2008), "Power law versus exponential law in characterizing stock market returns," *Atlantic Economic Journal*, 36(3), 377-379.
- Ben-Zion, Y. (2008), "Collective Behavior of Earthquakes and Faults: Continuum-Discrete Transitions, Progressive Evolutionary Changes and Different Dynamic Regimes," *Rev. Geophysics*, 46, RG4006.
- Bickel, P. and Doksum, K. (2007), *Mathematical Statistics* (Vol. I, 2nd ed.), New Jersey: Pearson Prentice Hall.
- Burnham, K.P., and Anderson, D.R. (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.), New York: Springer-Verlag.
- Cont, R. (2001), "Empirical properties of asset returns: stylized facts and statistical issues," *Quantitative Finance*, 1, 223-236.
- Cumming, S.G. (2001), "A parametric model of the fire-size distribution." *Canadian Journal of Forest Research*, 31, 1297-1303.
- Fox, J. (1990), "Describing univariate distributions," *Modern Methods of Data Analysis*, eds. J. Fox and J. S. Long, 581-25. Newbury Park, CA: Sage Publications.
- Geisser, S. (1993), *Predictive Inference*, New York: Chapman and Hall.
- Gu, G.F. and Zhou, W.X. (2009), "On the probability distribution of stock returns in the Mike-Farmer model," *European Physical Journal B*, 67(4), 585-592.
- Jackson, D.D., and Kagan, Y.Y. (1999), "Testable earthquake forecasts for 1999," *Seism. Res. Lett.*, 70(4), 393-403.
- Johnson, N.L., Kemp, A.W. and Kotz, S. (2005), *Univariate Discrete Distributions* (3rd ed.), Hoboken, NJ: John Wiley & Sons.
- Johnson, N.L., Kotz, S. and Balakrishnan, N. (1994), *Continuous Univariate Distributions* (Vol. 1, 2nd ed.), New York: John Wiley & Sons.
- Kagan, Y. Y. (1993), "Statistics of characteristic earthquakes," *Bull. Seismol. Soc. Amer.*, 83(1), 7-24.

- Kagan, Y.Y. (1994), "Observational evidence for earthquakes as a nonlinear dynamic process," *Physica D*, 77, 160-192.
- Kagan, Y., and Schoenberg, F. (2001), "Estimation of the upper cutoff parameter for the tapered Pareto distribution," *J. Appl. Prob.* 38A, Supplement: Festschrift for David Vere-Jones, D. Daley, editor, 158-175.
- Kaizoji, T. (2004), "Inflations and deflations in financial markets," *Physica A*, 343, 662-668.
- Knopoff, L., and Kagan, Y. (1977), "Analysis of the theory of extremes as applied to earthquake problems," *J. Geophys. Res.*, 82(36), 5647-5657.
- Malevergne, Y., Pisarenko, V.F., and Sornette, D. (2005), "Empirical Distributions of Log>Returns: between the Stretched Exponential and the Power Law?" *Quantitative Finance*, 5(4), 379-401.
- Malevergne, Y., Pisarenko, V.F., and Sornette, D. (2006), "Empirical Distributions of Log>Returns: between the Stretched Exponential and the Power Law?" *Applied Financial Economics*, 16, 271-289.
- Mandelbrot, B. (1963), "The Variation of Certain Speculative Prices," *Journal of Business*, 36, 394.
- Mandelbrot, B. (1967), "How Long Is the Coast of Britain? Statistical Self-Similarity and Fractional Dimension," *Science*, Vol. 156, No. 3775, 636-638.
- Ogata, Y. (1998), "Space-time point-process models for earthquake occurrences," *Annals of the Institute of Statistical Mathematics*, 50(2), 379-402.
- Pareto, V. (1897), *Cours d'Économie Politique*, Tome Second, Lausanne, F. Rouge, quoted by Pareto, V., 1964, *Œuvres Complètes*, Publ. by de Giovanni Busino, Genève, Droz, v. II.
- Peng, R.D., Schoenberg, F.P., Woods, J.A. (2005), "A space-time conditional intensity model for evaluating a wildfire hazard index," *Journal of the American Statistical Association*, 100(469), 26-35.
- Pisarenko, V.F. and Sornette, D. (2006), "New statistic for financial return distributions: power-law or exponential?," *Physica A*, 366, 387-400.
- Plerou, V., and Stanley, H.E. (2008), "Stock return distributions: Tests of scaling and universality from three distinct stock markets," *Physical Review*, E77(3), 037101-037104.
- Salgado-Ugarte, I. H. and Pérez-Hernández, M. A. (2003), "Exploring the use of variable bandwidth kernel density estimators," *Stata Journal*, 3(2).
- Schoenberg, F.P., Barr, C., and Seo, J. (2008), "The distribution of Voronoi cells

- generated by Southern California earthquake epicenters,” *Environmetrics*, 20(2), 159-171.
- Schoenberg, F.P., Peng, R., and Woods, J. (2003), “On the distribution of wildfire sizes,” *Environmetrics*, 14(6), 583-592.
- Schwarz, G. (1978), “Estimating the dimension of a model,” *Annals of Statistics*, 6(2):0, 461-464.
- Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York: John Wiley & Sons.
- Sheather, S. J. and Jones, M. C. (1991), “A reliable data-based bandwidth selection method for kernel density estimation,” *Journal of Royal Statistical Society, Series B* 53, 683-690.
- Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, New York: Chapman and Hall.
- Sornette, D. and Sornette, A. (1999), “General theory of the modified Gutenberg-Richter law for large seismic moments,” *Bull. Seism. Soc. Am.*, 89, N4:1121-1130.
- Stone, M. (1977), “Asymptotics for and against cross-validation,” *Biometrika*, 64(1), 29-35.
- Utsu, T. (1999), “Representation and analysis of the earthquake size distribution: a historical review and some new approaches,” *Pure Appl. Geophys.*, 155, 509-535.
- Vere-Jones, D. (1992), “Statistical methods for the description and display earthquake catalogues,” *Statistics in the Environmental and Earth Sciences*, A. T. Walden and P. Guttorp, eds., E. Arnold, London, 220-244.
- Vere-Jones, D., Robinson, R., and Yang, W.Z. (2001), “Remarks on the accelerated moment release model: problems of model formulation, simulation and estimation,” *Geophys. J. Int.*, 144, 517-531.

List of Figures

1	Empirical log survival (solid line) with 95% confidence bounds (dotted lines) of wildfire sizes in Los Angeles (left) and of earthquake inter-event times in Southern California (right), along with fitted Pareto (dashed line) and tapered Pareto (dotted-dashed line) distributions.	28
2	Top panel: simulated Pareto histogram (left) and tapered Pareto histogram (right). Bottom panel: empirical log survival (solid line) with 95% confidence bounds (dotted lines) of Pareto simulation (left) and of tapered Pareto simulation (right), along with fitted Pareto (dashed line) and tapered Pareto (dotted-dashed line) distributions. The simulations are generated using the parameters $\alpha = 1$, $\beta = 2$, $\theta = 3$, and sample size $n = 1000$	29
3	Left: $\hat{g}_k(x)$ for Pareto simulation. Right: $\hat{g}_k(x)$ for tapered Pareto simulation. The simulations are generated using the parameters $\alpha = 1$, $\beta = 2$, $\theta = 3$, and sample size $n = 1000$. For each simulation, the parameter $\hat{\beta}$ used to calculate $\hat{g}_k(x)$ is the maximum likelihood Pareto estimate of β . For Pareto simulation, $\hat{\beta} = 2.03$. For tapered Pareto simulation, $\hat{\beta} = 2.49$	30
4	Left: one-sided p-values for local non-self-similarity for the Pareto simulation vs. H_0 . Right: one-sided p-values for local non-self-similarity for the tapered Pareto simulation vs. H_0 . Assumes H_0 is the Pareto distribution fitted to each simulation by maximum likelihood. For Pareto simulation, H_0 : Pareto($\alpha = 1.00, \beta = 2.03$). For tapered Pareto simulation, H_0 : Pareto($\alpha = 1.00, \beta = 2.49$).	31
5	Left: empirical survival of a modified Pareto simulation vs. H_0 . Right: one-sided p-values for local non-self-similarity for a modified Pareto simulation vs. H_0 . In the modified Pareto simulation, the simulated Pareto variables used in Section 2.3 between $x = 5$ and $x = 5.5$ have been reduced by 1 so that they now fall between $x = 4$ and $x = 4.5$. H_0 is the Pareto distribution fitted to the modified data by maximum likelihood: Pareto($\alpha = 1.00, \beta = 2.03$).	32
6	Left: one-sided p-values for local non-self-similarity for wildfire sizes in Los Angeles vs. H_0 . Assumes H_0 : Pareto($\alpha = 0.405, \beta = 0.623$). Right: one-sided p-values for local non-self-similarity for earthquake inter-event times in Southern California vs. H_0 . Assumes H_0 : Pareto($\alpha = 9.99 \times 10^{-7}, \beta = 0.087$).	33
7	Left: $\hat{g}_k(x)$ for Google stock returns. Right: one-sided p-values of non-self-similarity for Google stock returns vs. H_0 . Assumes H_0 : Pareto($\alpha = 2.10 \times 10^{-5}, \beta = 0.1638$). The grid values of stock returns over which the test statistic is estimated range from 0.5% to 18% increasing by increments of 0.5%.	34

8 Left: theoretical $g_k(x)$ under H_0 . Right: one-sided p-values of non-self-similarity of Google stock returns vs. H_0 . Assumes H_0 : tapered Pareto($\alpha = 2.10 \times 10^{-5}, \beta = 8.74 * 10^{-9}, \theta = 0.0161$). 35

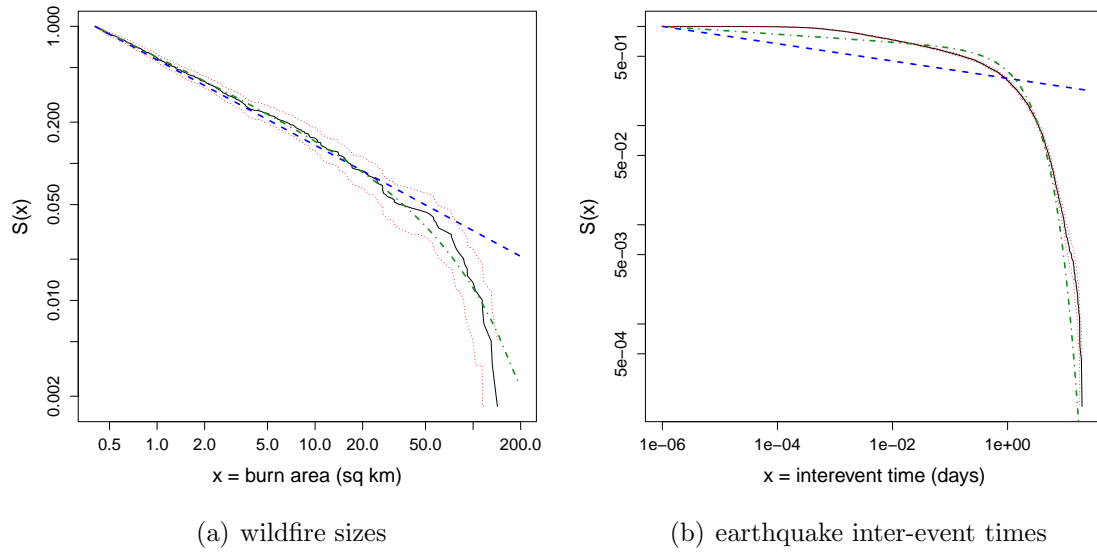
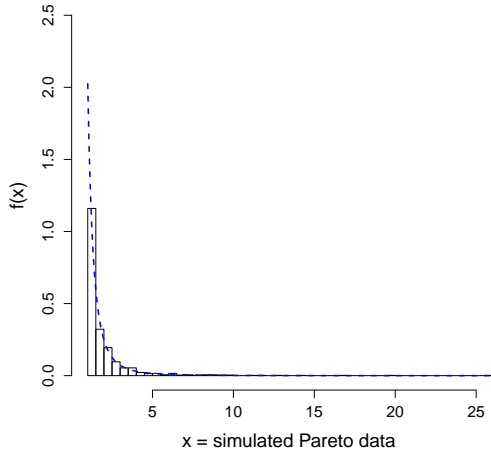
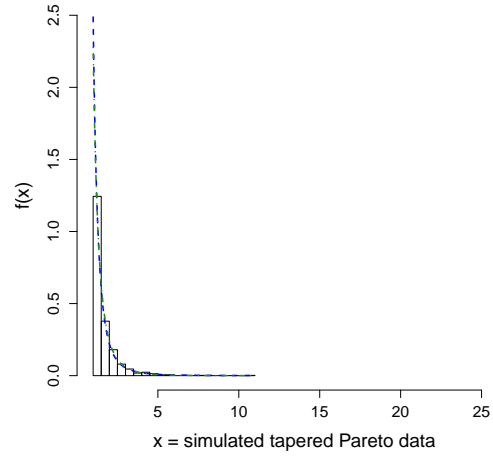


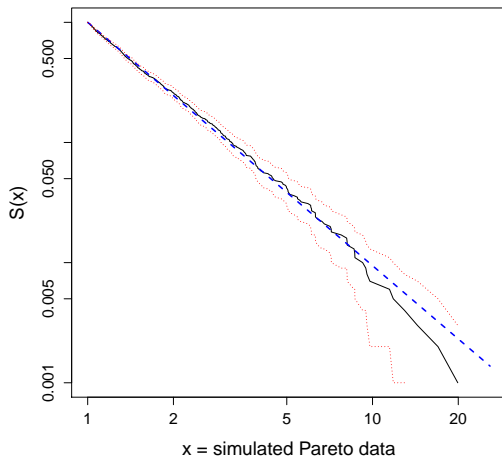
Figure 1: Empirical log survival (solid line) with 95% confidence bounds (dotted lines) of wildfire sizes in Los Angeles (left) and of earthquake inter-event times in Southern California (right), along with fitted Pareto (dashed line) and tapered Pareto (dotted-dashed line) distributions.



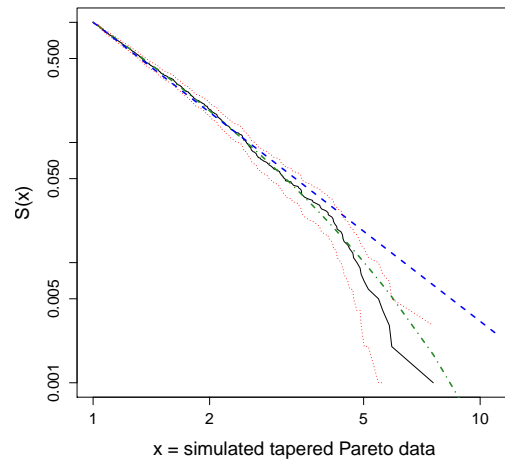
(a) Pareto simulation histogram



(b) tapered Pareto simulation histogram



(c) Pareto simulation survival



(d) tapered Pareto simulation survival

Figure 2: Top panel: simulated Pareto histogram (left) and tapered Pareto histogram (right). Bottom panel: empirical log survival (solid line) with 95% confidence bounds (dotted lines) of Pareto simulation (left) and of tapered Pareto simulation (right), along with fitted Pareto (dashed line) and tapered Pareto (dotted-dashed line) distributions. The simulations are generated using the parameters $\alpha = 1$, $\beta = 2$, $\theta = 3$, and sample size $n = 1000$.

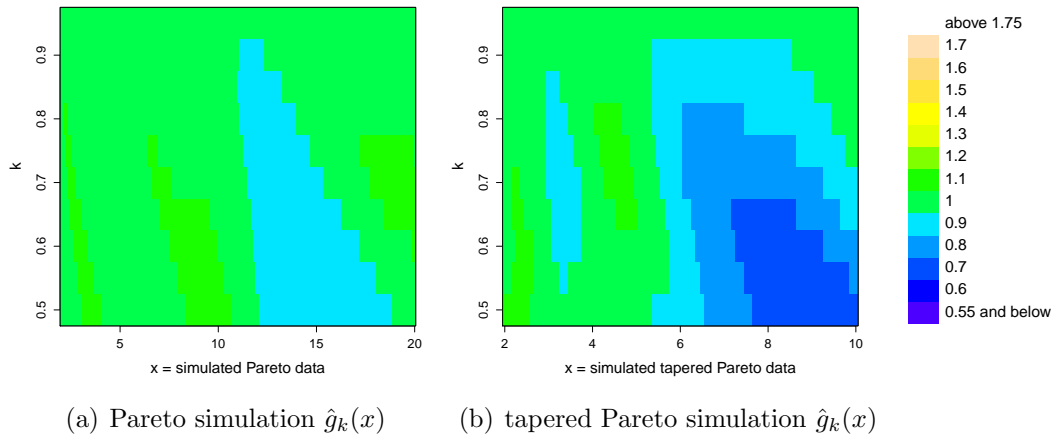
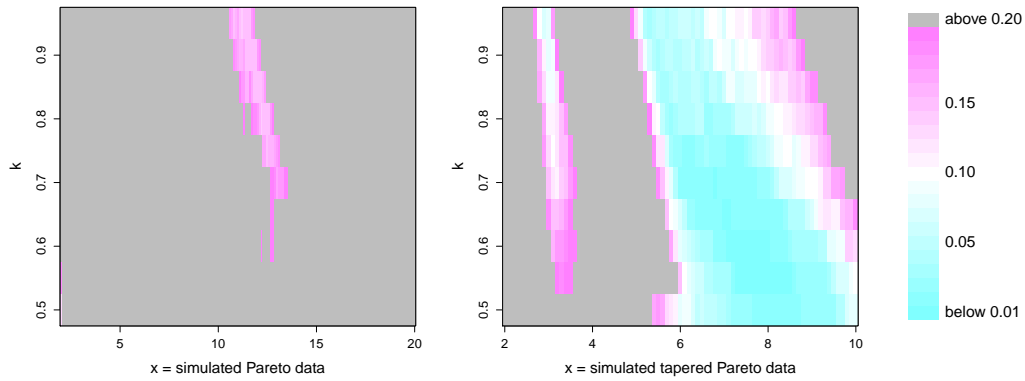
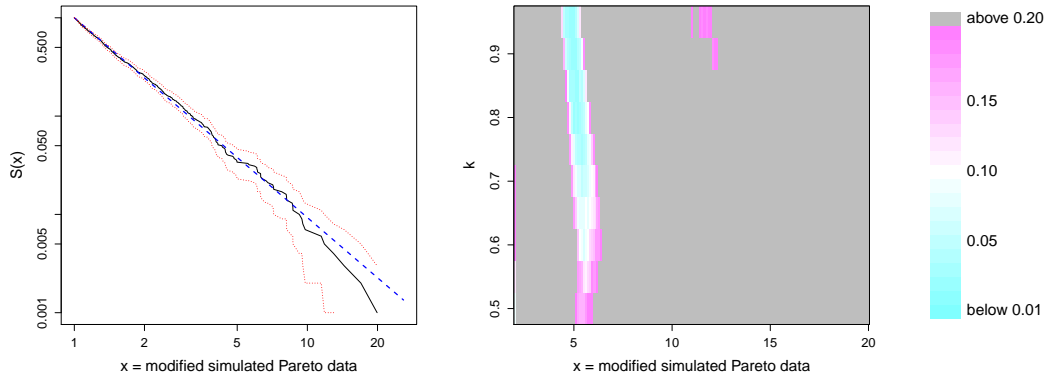


Figure 3: Left: $\hat{g}_k(x)$ for Pareto simulation. Right: $\hat{g}_k(x)$ for tapered Pareto simulation. The simulations are generated using the parameters $\alpha = 1$, $\beta = 2$, $\theta = 3$, and sample size $n = 1000$. For each simulation, the parameter $\hat{\beta}$ used to calculate $\hat{g}_k(x)$ is the maximum likelihood Pareto estimate of β . For Pareto simulation, $\hat{\beta} = 2.03$. For tapered Pareto simulation, $\hat{\beta} = 2.49$.



(a) Pareto simulation p-values (b) tapered Pareto simulation p-values

Figure 4: Left: one-sided p-values for local non-self-similarity for the Pareto simulation vs. H_0 . Right: one-sided p-values for local non-self-similarity for the tapered Pareto simulation vs. H_0 . Assumes H_0 is the Pareto distribution fitted to each simulation by maximum likelihood. For Pareto simulation, H_0 : Pareto($\alpha = 1.00, \beta = 2.03$). For tapered Pareto simulation, H_0 : Pareto($\alpha = 1.00, \beta = 2.49$).



(a) modified Pareto simulation survival (b) modified Pareto simulation p-values

Figure 5: Left: empirical survival of a modified Pareto simulation vs. H_0 . Right: one-sided p-values for local non-self-similarity for a modified Pareto simulation vs. H_0 . In the modified Pareto simulation, the simulated Pareto variables used in Section 2.3 between $x = 5$ and $x = 5.5$ have been reduced by 1 so that they now fall between $x = 4$ and $x = 4.5$. H_0 is the Pareto distribution fitted to the modified data by maximum likelihood: $\text{Pareto}(\alpha = 1.00, \beta = 2.03)$.

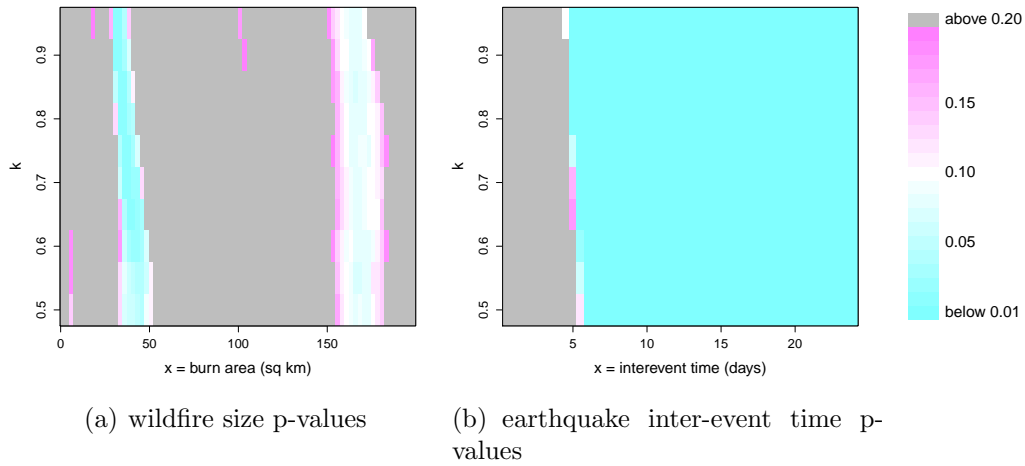


Figure 6: Left: one-sided p-values for local non-self-similarity for wildfire sizes in Los Angeles vs. H_0 . Assumes H_0 : $\text{Pareto}(\alpha = 0.405, \beta = 0.623)$. Right: one-sided p-values for local non-self-similarity for earthquake inter-event times in Southern California vs. H_0 . Assumes H_0 : $\text{Pareto}(\alpha = 9.99 \times 10^{-7}, \beta = 0.087)$.

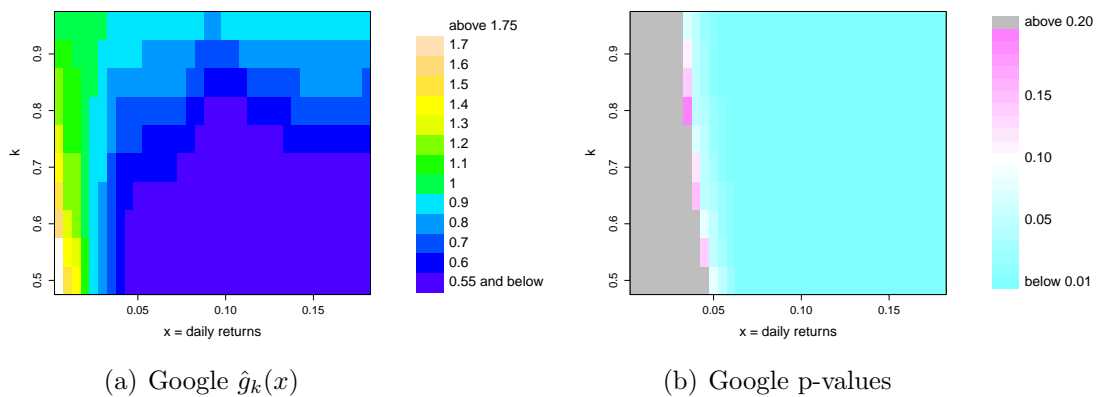


Figure 7: Left: $\hat{g}_k(x)$ for Google stock returns. Right: one-sided p-values of non-self-similarity for Google stock returns vs. H_0 . Assumes H_0 : Pareto($\alpha = 2.10 \times 10^{-5}$, $\beta = 0.1638$). The grid values of stock returns over which the test statistic is estimated range from 0.5% to 18% increasing by increments of 0.5%.

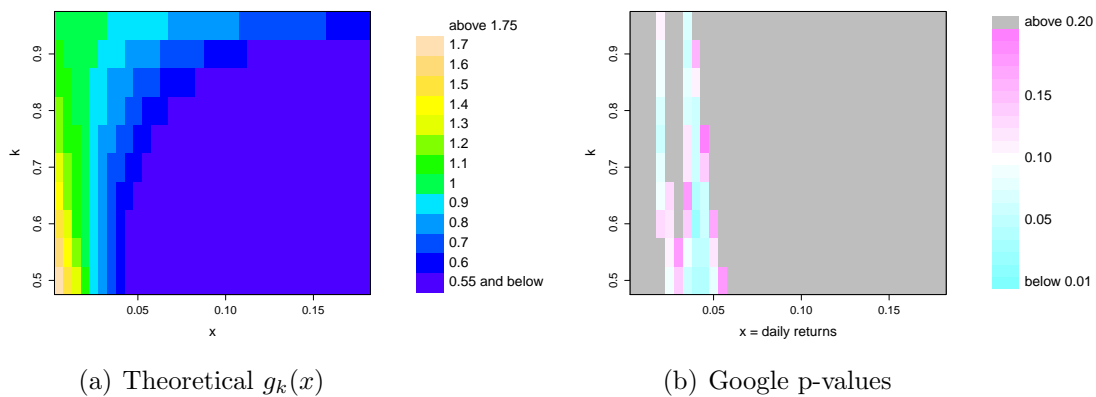


Figure 8: Left: theoretical $g_k(x)$ under H_0 . Right: one-sided p-values of non-self-similarity of Google stock returns vs. H_0 . Assumes H_0 : tapered Pareto($\alpha = 2.10 \times 10^{-5}$, $\beta = 8.74 * 10^{-9}$, $\theta = 0.0161$).

List of Tables

1	Correlations between test statistic $\hat{g}_{0.8}(x_i)$ and $\hat{g}_{0.8}(x_j)$ for 100 Pareto simulations generated using $\alpha = 1$, $\beta = 2$ and sample size $n = 1000$, where each i, j is in the sequence of ten equally spaced values of x between 3 and 12.	37
2	Correlations between estimated survival $\hat{S}(x_i)$ and $\hat{S}(x_j)$ for 100 Pareto simulations generated using $\alpha = 1$, $\beta = 2$ and sample size $n = 1000$, where each i, j is in the sequence of ten equally spaced values of x between 3 and 12.	38
3	Correlations between $g_{k_i}(5)$ and $g_{k_j}(5)$ for 100 Pareto simulations generated using $\alpha = 1$, $\beta = 2$ and sample size $n = 1000$, where each i, j is in the sequence of ten values of k between 0.50 and 0.95.	39

x	3	4	5	6	7	8	9	10	11	12
3	1.000	-0.561	-0.109	0.116	0.013	-0.094	-0.025	0.092	0.116	0.072
4	-0.561	1.000	-0.156	-0.414	-0.125	0.009	-0.019	-0.014	0.026	0.042
5	-0.109	-0.156	1.000	0.147	-0.398	-0.244	-0.191	-0.200	-0.081	0.084
6	0.116	-0.414	0.147	1.000	0.247	-0.313	-0.290	-0.170	-0.141	-0.147
7	0.013	-0.125	-0.398	0.247	1.000	0.530	-0.041	-0.258	-0.323	-0.333
8	-0.094	0.009	-0.244	-0.313	0.530	1.000	0.615	-0.042	-0.357	-0.368
9	-0.025	-0.019	-0.191	-0.290	-0.041	0.615	1.000	0.649	0.135	-0.252
10	0.092	-0.014	-0.200	-0.170	-0.258	-0.042	0.649	1.000	0.770	0.178
11	0.116	0.026	-0.081	-0.141	-0.323	-0.357	0.135	0.770	1.000	0.700
12	0.072	0.042	0.084	-0.147	-0.333	-0.368	-0.252	0.178	0.700	1.000

Table 1: Correlations between test statistic $\hat{g}_{0.8}(x_i)$ and $\hat{g}_{0.8}(x_j)$ for 100 Pareto simulations generated using $\alpha = 1$, $\beta = 2$ and sample size $n = 1000$, where each i, j is in the sequence of ten equally spaced values of x between 3 and 12.

x	3	4	5	6	7	8	9	10	11	12
3	1.000	0.676	0.580	0.417	0.366	0.322	0.293	0.278	0.246	0.249
4	0.676	1.000	0.838	0.687	0.638	0.502	0.440	0.445	0.386	0.373
5	0.580	0.838	1.000	0.836	0.774	0.631	0.543	0.537	0.458	0.412
6	0.417	0.687	0.836	1.000	0.879	0.784	0.672	0.650	0.507	0.480
7	0.366	0.638	0.774	0.879	1.000	0.871	0.759	0.740	0.620	0.585
8	0.322	0.502	0.631	0.784	0.871	1.000	0.880	0.821	0.695	0.644
9	0.293	0.440	0.543	0.672	0.759	0.880	1.000	0.926	0.825	0.784
10	0.278	0.445	0.537	0.650	0.740	0.821	0.926	1.000	0.917	0.857
11	0.246	0.386	0.458	0.507	0.620	0.695	0.825	0.917	1.000	0.918
12	0.249	0.373	0.412	0.480	0.585	0.644	0.784	0.857	0.918	1.000

Table 2: Correlations between estimated survival $\hat{S}(x_i)$ and $\hat{S}(x_j)$ for 100 Pareto simulations generated using $\alpha = 1$, $\beta = 2$ and sample size $n = 1000$, where each i, j is in the sequence of ten equally spaced values of x between 3 and 12.

k	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.50	1.000	0.946	0.822	0.715	0.656	0.617	0.566	0.488	0.389	0.283
0.55	0.946	1.000	0.933	0.803	0.694	0.617	0.544	0.458	0.360	0.259
0.60	0.822	0.933	1.000	0.947	0.840	0.726	0.603	0.469	0.336	0.214
0.65	0.715	0.803	0.947	1.000	0.958	0.857	0.709	0.533	0.358	0.205
0.70	0.656	0.694	0.840	0.958	1.000	0.960	0.840	0.662	0.468	0.288
0.75	0.617	0.617	0.726	0.857	0.960	1.000	0.953	0.821	0.642	0.456
0.80	0.566	0.544	0.603	0.709	0.840	0.953	1.000	0.952	0.827	0.664
0.85	0.488	0.458	0.469	0.533	0.662	0.821	0.952	1.000	0.957	0.848
0.90	0.389	0.360	0.336	0.358	0.468	0.642	0.827	0.957	1.000	0.963
0.95	0.283	0.259	0.214	0.205	0.288	0.456	0.664	0.848	0.963	1.000

Table 3: Correlations between $g_{k_i}(5)$ and $g_{k_j}(5)$ for 100 Pareto simulations generated using $\alpha = 1$, $\beta = 2$ and sample size $n = 1000$, where each i, j is in the sequence of ten values of k between 0.50 and 0.95.