Advanced Methods in Applied Statistics 2018

Power-law distributions in empirical data

Article summary

Christian Anker Rosiek*

2018-03-07

What follows is a summary of the article "Power-law distributions in empirical data" by Aaron Clauset, Cosma Rohilla Shalizi and M.E.J. Newman (ref. [1]). The article presents tools to analyze datasets w.r.t. power-laws. Power-laws occur in diverse scientific fields and are made difficult to characterize due to large tail-fluctuations. It is therefore of scientific interest to develop methods for analyzing data hypothesized to follow such a distribution.

After introducing discrete and continuous power-laws, the article describes the maximum likelihood estimators for relevant parameters for both continuous and discrete distributions, and subsequently goes through goodness-of-fit test. Finally, the methods introduced are applied to 24 real-world datasets. This summary follows a similar structure emphasizing the methodology.

Power-law distributions. A quantity x obeys a power-law if it is drawn from a distribution proportional to $x^{-\alpha}$. The parameter α is known as the *scaling parameter*. A given quantity commonly obeys the power law only in some subinterval of $(0, \infty)$. Lower and upper bounds $x_{\min}$ and $x_{\max}$ for the power law may be introduced. The article [1] only considers distributions unbounded from above. For the doubly bounded discrete case, see e.g. ref. [2]. The normalized continuous distribution has probability density function (PDF)

$$p(x) = \frac{\alpha - 1}{x_{\min}} \left(\frac{x}{x_{\min}} \right)^{-\alpha}, \tag{1}$$

whereas normalized discrete distribution has probability mass function (PMF)[1]

$$p(x) = \frac{x^{-\alpha}}{\zeta(\alpha, x_{\min})}. \tag{2}$$

$\zeta(\alpha, x_{\min})$ is the *generalized* or *Hurwitz zeta-function* $\zeta(\alpha, x_{\min}) = \sum_{n=0}^{\infty} (n + x_{\min})^{-\alpha}$.

Note also the definitions of the complementary cumulative distribution function (CDF) for the continuous case, $P(x) = \int_x^\infty p(x')dx' = (x/x_{\min})^{-\alpha+1}$, and for the discrete case, $P(x) = \sum_{y=x}^{\infty} p(y) = \zeta(\alpha, x)/\zeta(\alpha, x_{\min})$.

Parameter estimation. The article [1] discusses the estimation of distribution parameters α and $x_{\min}$. Following article notation, estimators are denoted by "hatted" symbols, e.g. $\hat{\alpha}$ is the maximum likelihood estimator (MLE) for α.

Since a power-law becomes linear in a log-log plot, a common approach is to perform a linear least squares-fit to binned data on a log-log plot. This method is demonstrated to be inaccurate regardless of binning convention, as is also displayed in Figure 1.

Continuous distribution α MLE. For observations $\{x_i\}_{i \in \mathbb{N}}$, the MLE for the scaling parameter is given by

$$\hat{\alpha} = 1 + n \left(\sum_{i=1}^{n} \ln \frac{x_i}{x_{\min}} \right)^{-1} \tag{3}$$

with corresponding standard error $\sigma = (\hat{\alpha} - 1)/\sqrt{n} + \mathcal{O}(1/n)$. n is the number of observations.

Discrete distribution α MLE. Generalized to arbitrary integer $x_{\min}$, the MLE $\hat{\alpha}$ for the discrete case is found by maximizing the likelihood

$$\mathcal{L}(\alpha) = -n \ln \zeta(\alpha, x_{\min}) - \alpha \sum_{i=1}^{n} \ln x_i \tag{4a}$$

as a function of α, or equivalently solving the equation

$$\frac{\zeta'(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})} = -\frac{1}{n} \sum_{i=1}^{n} \ln x_i \,. \tag{4b}$$

The standard error on this $\hat{\alpha}$ may be estimated as

$$\sigma = \frac{1}{\sqrt{n}} \left(\frac{\zeta''(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})} - \left(\frac{\zeta'(\hat{\alpha}, x_{\min})}{\zeta(\hat{\alpha}, x_{\min})} \right)^2 \right)^{-1/2}. \tag{4c}$$

Estimating the lower bound $x_{\min}$. The article also discusses the estimation of the lower bound. The estimate of α is highly dependent on accurate estimation of $x_{\min}$, as is displayed in Figure 2—underestimating $x_{\min}$ will include non-power-law data whereas overestimating it will discard valid power-law data, increasing sensitivity to statistical fluctuation. Two estimators

*xvm706@alumni.ku.dk

[1] This summary tries to match the notation of the article. Thus depending on context, x represents either a discrete or a continuous variable, $p(x)$ represents a PDF or a PMF, etc.

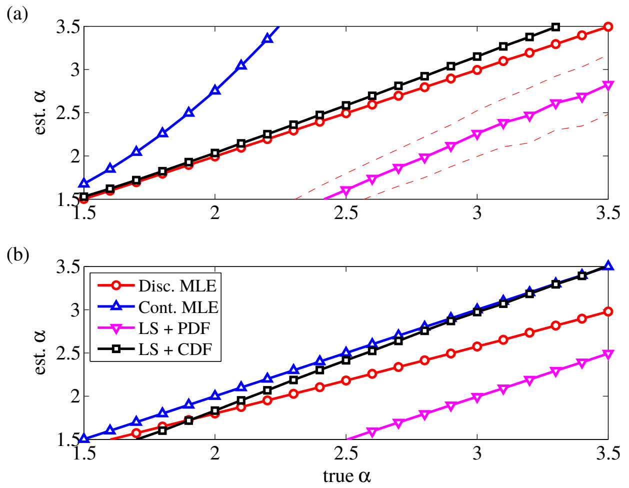

Figure 1: Figure 3.2 from article [1]. Comparison of different parameter estimation methods for α with data drawn from (a) discrete and (b) continuous distributions, namely: the discrete MLE (3), the continous MLE (4b), a linear least squares fit to constant-width bins of PDF, and a linear least squares fit to the CDF rank-frequency plot. Note how poorly all shown estimators, with the exception of the relevant MLE, perform.

are presented and tested on a particular sampled distribution. The second one is found to perform better, although both are described as reasonable.

The first estimator relies on an approximation known as a *Bayesian information criterion* (BIC) and mentioned to be valid for only discrete distributions. To estimate $x_{\min}$, one models the distribution as a set of independent probabilities for the discrete events below $x_{\min}$ in combination with the expected power-law above, and then maximizes the marginal likelihood for $x_{\min}$.

The alternate estimator (KS), valid for both discrete and continuous data, maximizes the similarity between the best-fit power-law and the empirical distribution. Similarity between CDFs of the data $S(x)$ and the fit $P(x)$ is here described by the Kolmogorov-Smirnov (KS) test statistic

$$D = \max_{x \geq x_{\min}} |S(x) - P(x)|, \tag{5}$$

but any test statistic can in principle be used. Minimizing D as a function of $x_{\min}$ yields an estimate for $x_{\min}$. Alternate test statistics are proposed, in particular a modified KS test statistic, re-weighted to distribute sensitivity uniformly across the entire data range.

Goodness-of-fit tests and model comparison. A Monte Carlo procedure for performing goodness-of-fit tests on fitted datasets is described, Using parameters obtained through the methods described in the previous sections, a number of synthetic datasets are sampled from a distribution with the same parameter values. The KS statistic is then calculated for the synthetic

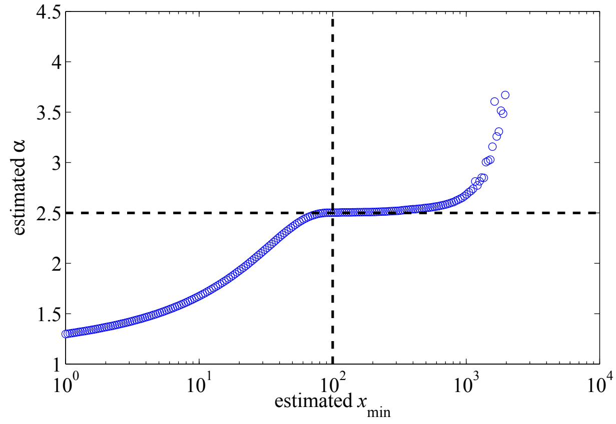

Figure 2: Figure 3.3 from article [1]. For a sampled true distribution ($\alpha = 2.5$ and $x_{\min} = 100$) with power-law behavior beyond $x_{\min}$, this shows how the estimate of α depends on the chosen cut-off (the lowest value of any fitted point) $x_{\min}$. The α-estimate appears relatively forgiving when overestimating $x_{\min}$, however also appears to quickly deteriorate with underestimation.

datasets and the p-value for the original dataset is then given as the fraction of synthetic datasets that perform worse than the original data in the KS test. Note that in this case, a larger p-value indicates a "better" fit to the data. Datasets with $p \leq 0.1$ are rejected.

For comparing plausibility different models for a given dataset, the article proposes comparison based on the likelihood ratio between the two models. This method may be used to reject a model in comparison with another and a p-value for the statistical significance of the rejection is computed based on ref. [3].

Application to real-world data. Finally, the methods just described are applied to 24 real-world datasets from a diverse series of fields, estimating scaling parameters and comparing different heavy-tailed distributions. One key observation is that the distinction between a log-normal and a power-law distribution is very difficult. For some datasets, scaling parameter estimates incompatible with previously published estimates are found, suggesting reevaluation of any resultant conclusions.

References.

[1] Aaron Clauset, Cosma Rohilla Shalizi, and M.E.J. Newman. *Power-law distributions in empirical data.* SIAM Review **51**, 661–703 (2009).

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

[3] Vuong, Q. H. *Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses Econometrica.* Econometric Society **57**, 307–333 (1989).