A practical recipe to fit discrete power-law distributions

Álvaro Corral[1], Anna Deluca[1,2], and Ramon Ferrer-i-Cancho[3]

[1]Centre de Recerca Matemàtica, Bellaterra, Barcelona, Spain
[2]Departament de Matemàtiques, UAB, Bellaterra, Barcelona, Spain
[3]Departament de Llenguatges i Sistemes Informàtics, UPC, Barcelona, Spain

Abstract

Power laws pervade statistical physics and complex systems [1, 2], but, traditionally, researchers in these fields have paid little attention to properly fit these distributions. Who has not seen (or even shown) a log-log plot of a completely curved line pretending to be a power law? Recently, Clauset et al. have proposed a method to decide if a set of values of a variable has a distribution whose tail is a power law [3]. The key of their procedure is the identification of the minimum value of the variable for which the fit holds, which is selected as the value for which the Kolmogorov-Smirnov distance between the empirical distribution and its maximum-likelihood fit is minimum. However, it has been shown that this method can reject the power-law hypothesis even in the case of power-law simulated data [4]. Here we propose a simpler selection criterion, which is illustrated with the more involving case of discrete power-law distributions.

1 Procedure

This method is similar in spirit to the one by Clauset et al. [3,4], but with important differences [5]. Here we just present the recipe, the justification is available in Ref. [6].

Consider a **discrete power-law** distribution, defined for $n = a, a+1, a+2, \dots \infty$ (with a natural),

$$f(n) = \text{Prob}[\text{variable} = n] = \frac{1}{\zeta(\beta+1, a) n^{\beta+1}}$$

$$S(n) = \text{Prob}[\text{variable} \geq n] = \frac{\zeta(\beta+1, n)}{\zeta(\beta+1, a)}$$

with $\beta > 0$ and ζ the Hurwitz zeta function [3] (Riemann function for $a = 1$),

$$\zeta(\gamma, a) = \sum_{k=0}^{\infty} \frac{1}{(a+k)^\gamma}.$$

Note then that $f(n)$ is a power law but $S(n)$ is not (only asymptotically).

For a fixed, the data values verifying $n \geq a$ are numbered from $i = 1$ to N_a, and the remainder is removed.

Then, the method consists of the following steps:

1. **Maximum likelihood estimation** of the exponent β.

   Calculate the log-likelihood function,

   $$\ell(\beta) = \frac{1}{N_a}\sum_{i=1}^{N_a} \ln f(n_i) = -\ln \zeta(\beta+1, a) - (\beta+1)\ln G_a,$$

   with G_a the geometric mean of the data in the range, $\ln G_a = N_a^{-1}\sum \ln n_i$.

   Calculate the maximum of $\ell(\beta)$ (for instance through the downhill simplex method [7]),

   $$\beta_{emp} = \max_{\forall \beta} \ell(\beta),$$

   which has an error (standard deviation [3])

   $$\sigma = \frac{\beta_{emp}}{\sqrt{N_a}}.$$

   The computation of the zeta function uses the Euler-Maclaurin formula [8, 9],

   $$\sum_{k=0}^{\infty} \tilde{f}(k) \simeq \sum_{k=0}^{M-1} \tilde{f}(k) + \int_M^{\infty} \tilde{f}(k)dk + \frac{\tilde{f}(M)}{2} - \sum_{k=1}^{P} \frac{B_{2k}}{(2k)!}\tilde{f}^{(2k-1)}(M),$$

   where B_{2k} are the Bernoulli numbers ($B_2 = 1/6, B_4 = -1/30, B_6 = 1/42, B_8 = -1/30, \dots$) [8]. So,

   $$\zeta(\gamma, a) \simeq \sum_{k=0}^{M-1} \frac{1}{(a+k)^{\gamma}} + \frac{(a+M)^{1-\gamma}}{\gamma - 1} + \frac{1}{2(a+M)^{\gamma}} + \sum_{k=1}^{P} B_{2k}C_{2k-1}(M),$$

   with

   $$C_{2k-1}(M) = \frac{(\gamma+2k-2)(\gamma+2k-3)}{2k(2k-1)(a+M)^2}C_{2k-3}(M) \text{ and } C_1(M) = \frac{\gamma}{2(a+M)^{\gamma+1}}.$$

   The second sum in the formula runs from $k = 1$ to a fixed P, taken $P = 18$, except if a minimum value term ($B_{2k}C_{2k-1}(M)$) is reached, case in which the sum is stopped; this ensures a better convergence [9]. We also take $M = 14$.

   Once we obtain β_{emp}, how do we know if the fit is good or bad?

2. **Calculation of the Kolmogorov-Smirnov statistic** [7],

   $$d_{emp} = \max_{\forall n \geq a} \left| \frac{N_n}{N_a} - S(n; \beta_{emp}) \right|,$$

   with N_n the number of data taking values larger or equal to n. The maximization is performed for all values of $n \geq a$, integer and not integer.

   Large and small values of d_{emp} denote respectively bad and good fits. But what is large and small? This is determined in Step 3.

3. **Simulation of the discrete power-law distribution**, with exponent β_{emp} and $n \geq a$.

   We use a generalization of the rejection method of Ref. [10]:

   (a) Generate a uniform random number u between 0 and u_{max}, with $a = 1/u_{max}^{1/\beta_{emp}}$.

   (b) Obtain a new random number
   $$y = \text{int}(1/u^{1/\beta_{emp}}),$$
   where int(x) means the integer part of x. Notice that its probability function is
   $$q(y) = (a/y)^{\beta_{emp}} - (a/(y+1))^{\beta_{emp}}.$$

   (c) Accept y as the simulated value if a new uniform random number v (between 0 and 1) fulfills
   $$v \leq \frac{f(y)q(a)}{f(a)q(y)}$$
   and reject y otherwise. If accepted, take $n = y$.
   Notice that the computation of the ζ function is not required.
   Defining $\tau = (1 + y^{-1})^{\beta_{emp}}$ and $b = (a+1)^{\beta_{emp}}$ the acceptation condition becomes simpler,
   $$vy\frac{\tau - 1}{b - a^{\beta_{emp}}} \leq \frac{a\tau}{b},$$

   (d) Repeat the process until N_a values of $n = y$ are obtained.

4. Apply step 1 (maximum likelihood estimation) to the simulated data.

   Call the obtained exponent β_{sim}.

5. Apply step 2 (calculation of the Kolmogorov-Smirnov statistic) to the simulated data, using the fit obtained in step 4, as
   $$d_{sim} = \max_{\forall n \geq a} \left| \frac{N_{sim}(n)}{N_a} - S(n; \beta_{sim}) \right|,$$
   with $N_{sim}(n)$ the number of simulated data taking values larger or equal to n.

6. Comparison of the 2 statistics d_{emp} and d_{sim} is not enough, so:

   Repeat steps 3, 4, and 5 a large enough number of times (e.g., 100 or more, as allowed by computational resources), in order to get an ensemble of values of d_{sim}.

7. Compute $p-$value as
   $$p = \frac{\text{number of simulations with } d_{sim} > d_{emp}}{\text{number of simulations}}.$$
   The error of the $p-$value comes from that of a binomial distribution,
   $$\sigma_p = \sqrt{\frac{p(1-p)}{\text{number of simulations}}}.$$
   Low values of p, like $p \leq 0.05$ are considered bad fits.
   For higher values, $p > 0.05$, the power-law fit with β_{emp} cannot be rejected.

Repeating the whole procedure for "all" values of a we obtain a set of acceptable pairs of a and β_{emp}. Select the one that gives the smallest value of a provided that p is above 0.20 (for instance). In a formula,

$$a^* = \min\{a \text{ such that } p > 0.20\},$$

which has associated the resulting exponent β^*_{emp}.

Note that the final $p-$value of the procedure is not the one obtained for fixed a, but this is not relevant in order to provide a good fit (as long as the latter is larger than, say, 0.20).

The figures illustrate the results for n = word frequencies in the Finnish novel *Seitsemän veljestä* by Aleksis Kivi, for which $a^* = 1$ and $\beta^*_{emp} = 1.13 \pm 0.01$, with $N_{a^*} = 22035$ and 8.1×10^4 word tokens. Notice that $f(n)$ is a power law but $S(n)$ is not, but both are representations of a power-law distribution.

We thank R. D. Malmgren (for discussions), L. Devroye (for his book), G. Boleda (for many things!), and the assistants to the 2012 meeting of the network `complexitat.cat` (for their interest).

References

1. Bak P (1996) How Nature Works: The Science of Self-Organized Criticality. Copernicus, New York.

2. Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. Cont Phys 46: 323 –351.

3. Clauset A, Shalizi CR, Newman MEJ (2009) Power-law distributions in empirical data. SIAM Rev 51: 661–703.

4. Corral A, Font F, Camacho J (2011) Non-characteristic half-lives in radioactive decay. Phys Rev E 83: 066103.

5. Peters O, Deluca A, Corral A, Neelin JD, Holloway CE (2010) Universality of rain event size distributions. J Stat Mech P11030.

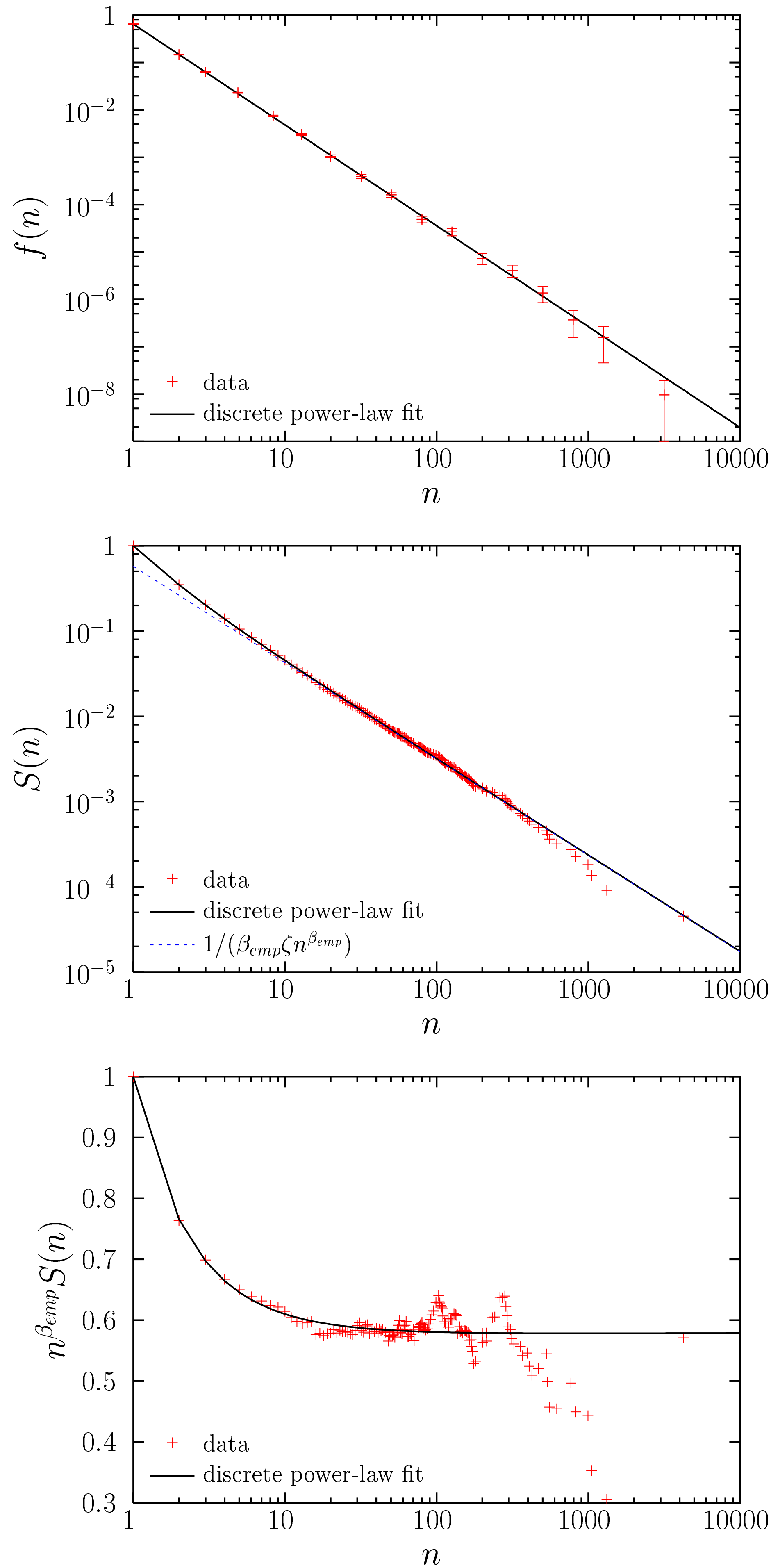
6. Corral A, Boleda G, Ferrer-i-Cancho R (2012) in preparation .

7. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical Recipes in FORTRAN. Cambridge University Press, Cambridge, 2nd edition.

8. Abramowitz M, Stegun IA, editors (1965) Handbook of Mathematical Functions. Dover, New York.

9. Vepstas L (2007) An efficient algorithm for accelerating the convergence of oscillatory series, useful for computing the polylogarithm and Hurwitz zeta functions. ArXiv : math/0702243.

10. Devroye L (1986) Non-Uniform Random Variate Generation. Springer-Verlag, New York.



$f(n)$
n
data
discrete power-law fit
$S(n)$
n
data
discrete power-law fit
$1/(\beta_{emp}\zeta n^{\beta_{emp}})$
$n^{\beta_{emp}}S(n)$
n
data
discrete power-law fit