A Generalized Power Law Model of Citations

Ruheyan Nuermaimaiti[1], Leonid V. Bogachev[2] and Jochen Voss[3]

[1] *mmrn@leeds.ac.uk* [2] *L.V.Bogachevz@leeds.ac.uk* [3] *J.Voss@leeds.ac.uk*
School of Mathematics, University of Leeds, Leeds LS2 9JT (United Kingdom)

Abstract
The classical power law model is widely used in informetrics to describe citations of scientific papers, although it is not addressing variability across individual authors. We report our preliminary results for a novel model based on a certain parametric form of the expected individual citation profile which generalizes the power law frequency formula. The new model interpolates between large citation numbers, where the power law tail is reproduced, and low citation numbers, which are usually truncated when fitting the power law model to the data. In addition, we derive a deterministic limit shape of the citation profile, which can be used to make predictions about various citation function such as the h-index.

Introduction

The classical power-law model was introduced by Lotka (1926) as an empirical match with observed frequencies of citations in scientific publications. In a later development, Price (1965) discovered an important connection with networks, whereby citations were interpreted as nodes' degrees. Examples of fitting the power law to the citation data can be found in Coile (1977), Redner (1998), and Clauset, Shalizi & Newman (2009). In particular, it was found that the power law frequencies do not necessarily fit well in the entire citation spectrum, so that a suitable truncation of lower citation values may be needed.

Importantly, no assumptions are made in the power law model about the frequency distribution of citations for an individual author randomly chosen from the population of authors. This makes it difficult to project the model fitted to a pooled corpus of publications onto individual authors, for example, for the purposes of evaluating their productivity.

The power-law model can be fitted to real-life data using standard statistical methods such as the maximum likelihood or ordinary least squares estimation. As has been documented across many use cases (Clauset, Shalizi & Newman, 2009), the power law usually fits quite well but only in the tail region of the frequency range, which motivates the use of truncated power-law models by excluding the lower values. This may decrease the utility of the model in estimation of various functions of citations, such as the popular h-index, introduced by Hirsch (2005) and defined as the maximum number h of an author's papers, each one cited at least h times.

In an attempt to overcome this shortcoming, we propose a novel model by modifying the power law setting. The new model interpolates between slow (almost flat) decay of the citation frequencies at the bottom of the citation spectrum and then reproducing the power-law behavior at the tail of the frequency distribution. As we will demonstrate below using a small real data set, the model provides a very good fit across the entire citation spectrum. In addition, and in contrast to the scale-free power law, our model possesses a deterministic limit shape of the citation profile, which can be used, for example, to make meaningful estimation of the h-index. In particular, the estimation of the h-index based on the modified model appears to be significantly more accurate as compared to that in the standard power law model.

Power law frequencies

In its classical setting, the power law model states that the relative frequency f_j of exactly j citations accumulated by a randomly sampled paper is proportional to the a-th power of j, with some exponent $a > 1$ (typically lying in the range $2 < a < 3$), that is (to include the case $j = 0$ and to normalize the sum of frequencies to unity),

$$f_j = \frac{c}{(j+1)^a} \qquad (j = 0,1,2,\dots), \tag{1}$$

where the normalizing constant c is given by the reciprocal Riemann zeta function,

$$c^{-1} = \sum_{j\geq 0} \frac{1}{(j+1)^a} = \zeta(a). \tag{2}$$

Here, "randomly sampled" means that a paper is sampled from a *pooled corpus* of papers written by a (large) population of authors. For simplicity, interaction effects due to joint authorship are not taken into account; such effects are complicated but have minor impact.

In terms of analysis of citation data, if there are M_j papers with j citations, out of the total number of papers M, then the power law model predicts that the relative frequencies M_j/M are approximately given by formula (1),

$$f_j \approx \frac{M_j}{M} \qquad (j = 0,1,2,\dots). \tag{3}$$

Note that the total number of papers and the total number of citations in this corpus are given, respectively, by

$$M = \sum_{j\geq 0} M_j\,, \qquad N = \sum_{j\geq 0} j\,M_j\,. \tag{4}$$

Of course, in any real-life data set the numbers M_j will reduce to zero for j big enough (so that the series in (4) are in fact finite sums), but this is reconciled with the prediction (1) simply by the fact that the theoretical frequencies f_j tend to zero as $j \to \infty$.

As mentioned in the Introduction, the power law model does not address the frequency distribution of citations for an *individual author* randomly chosen from the observed population of authors (say, of size K) and featured by a collection of *citation counts* ν_j, that is, the numbers of papers by this author that have j citations ($j = 0,1,2,$ ...). Moreover, the number of observed authors, K, is often omitted in popular citation data sets (cf. Redner, 1998).

Having in mind statistically homogeneous populations of authors who produce their research outputs according to the same probability distribution, it is natural to assume that these authors are independent from one another due to lack of interaction. Equally reasonable is the assumption of mutual independence of the counts $\nu_j(k)$ for each individual author $k = 1, \dots, K$. In this notation, the pooled numbers M_j are given by

$$M_j = \sum_{k=1}^{K} \nu_j(k) \qquad (j = 0,1,2,\dots). \tag{5}$$

Then, according to the law of large numbers, for $K \gg 1$ we have

$$\frac{M_j}{K} = \frac{\nu_j(1) + \cdots + \nu_j(K)}{K} \approx E(\nu_j), \tag{6}$$

where E stands for expectation (statistical mean) and a random variable ν_j represents the counts $\nu_j(k)$. Recalling (4), the mean number of papers per author is approximated as

$$\frac{M}{K} = \sum_{j\geq 0} \frac{M_j}{K} \approx \sum_{j\geq 0} E(\nu_j) \qquad (j = 0,1,2,\dots). \tag{7}$$

provided that the series on the right is convergent. Thus, combining formulas (3), (6) and (7), we obtain the link between the power-law frequencies f_j and the expected counts $E(\nu_j)$,

$$f_j \approx \frac{M_j/K}{M/K} \approx \frac{E(\nu_j)}{\sum_{j\geq 0} E(\nu_j)} \qquad (j = 0,1,2,\dots). \tag{8}$$

In practice, power law is often fitted in the tail of the frequency distribution, that is, for $j \geq j_*$, with a suitably chosen truncation point j_*. This leads to readjustment of the normalizing constant c_a in the frequency formula (1). Fitting such a model to the data requires optimization over two parameters, a and j_*. Specifically, a natural heuristic tool to fit a truncated power-law model is by looking at the frequency plots (e.g., histograms) with logarithmic scales on both axes, whereby one seeks a straight-line fit, with the slope corresponding to $(-a)$ (cf. Nicholls, 1987). An alternative approach (Clauset, Shalizi & Newman, 2009), which provides the helpful smoothing of the discrete data, is via the complementary cumulative frequencies

$$F_j = \sum_{\ell \geq j} f_\ell \qquad (j \geq j_*). \tag{9}$$

Using again the log-log plots, a good fit corresponds to a straight line, with slope $(1 - a)$.

Generalized power-law model

Setting of the model

The generalized power-law (GPL) model introduced in this section is set out using two hyper-parameters n and m, interpreted as the mean numbers per author of citations and papers, respectively. These parameters can be estimated from the observed pooled corpus by

$$n \approx \frac{N}{K}\,, \qquad m \approx \frac{M}{K}. \tag{10}$$

There are also two shape parameters, $a > 2$ (akin to the power-law exponent) and $b > 0$. Namely, the citation frequencies are now assumed to be of the form (cf. (1))

$$f_j = \frac{C m^{ab-1}}{(j + m^b)^a} \qquad (j = 0,1,2, \dots). \tag{11}$$

For small j ($j \ll m^b$) we have $f_j \approx C/m$, while for larger j we get a power-law dependence,

$$f_j \approx \frac{C m^{ab-1}}{j^a} \qquad (j \gg m^b). \tag{12}$$

This may be viewed as an effective sewing of the formerly truncated lower values with the power-law tail. Recalling the link (8) between the pooled frequencies f_j and the individual expected counts $E(\nu_j)$, we have

$$E(\nu_j) \approx m\, f_j = \frac{C}{(jm^{-b} + 1)^a}\,. \tag{13}$$

Hence, we can calibrate the model using (10) to make it consistent with the hyper-parameters,

$$m \approx \sum_{j \geq 0} E(\nu_j) = C \sum_{j \geq 0} \frac{1}{(jm^{-b} + 1)^a} \approx C m^b \int_0^\infty \frac{dx}{(x+1)^a} = \frac{C m^b}{a - 1}\,, \tag{14}$$

and similarly

$$n \approx \sum_{j \geq 0} j\, E(\nu_j) = C \sum_{j \geq 0} \frac{j}{(jm^{-b} + 1)^a} \approx C m^{2b} \int_0^\infty \frac{x\, dx}{(x+1)^a} = \frac{C m^{2b}}{(a-1)(a-2)}\,. \tag{15}$$

Considering the ratio n/m (i.e., the mean number of citations per paper), we get

$$\frac{n}{m} \approx \frac{m^b}{a - 2}\,, \qquad C \approx m^{1-b}(a - 1). \tag{16}$$

Hence, b is expressed in terms of the hyper-parameters and the parameter a,

$$b \approx \frac{\log n + \log(a-2)}{\log m} - 1\,. \tag{17}$$

The shape parameter a can be fitted to the data using either ordinary least squares or a suitable version of the maximum likelihood estimation (cf. Nicholls, 1987).

Limit shape

It is useful to represent the citation profile of an individual author by ranking their papers according to the citation scores (i.e., accumulated numbers of citations) $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$; for example, $\lambda_1 = \max\{\lambda_i\}$ is the score of the most cited paper. The citation profile is succinctly visualized by the *Young diagram* formed by (left- and bottom-aligned) row blocks with $\lambda_1, \lambda_2, \ldots$ unit square cells; its upper boundary is the graph of the step function

$$Y(x) = \sum_{j \geq x} \nu_j \qquad (x \geq 0), \tag{18}$$

where ν_j are the author's citation counts. In particular, $Y(0)$ is their total number of papers.

A useful insight into the structure of the citations data may be available by looking at the shape of suitably rescaled Young diagrams when both n and m are large (Vershik, 1996). Specifically, set

$$A = m^b, \qquad B = n/m^b, \tag{19}$$

and consider the rescaled (expected) shape

$$E\big(\tilde{Y}(x)\big) = \frac{1}{B} E\big(Y(Ax)\big) = \frac{1}{B} \sum_{j \geq Ax} E(\nu_j). \tag{20}$$

Substituting expressions (13) and approximating the sum by an integral like in (14) and (15), it is easy to show that there is a limit shape given by

$$E\big(\tilde{Y}(x)\big) \approx \varphi(x; a) = \frac{a-2}{(x+1)^{a-1}} \qquad (x \geq 0). \tag{21}$$

Data analysis

Goodness-of-fit

In this section, we fit the two models discussed above to real citation data, collected in January 2020 for a small population of 113 authors identified as those who had published at least one paper in the *Electronic Journal of Probability* in the first 10 issues (January–October) of volume 24 in 2019 (https://projecteuclid.org/euclid.ejp/1546571125) and who are also featured on Google Scholar (https://scholar.google.com). The total counts for this data set are $K = 113$ (authors), $N = 245{,}567$ (citations) and $M = 15{,}400$ (papers), including $M_0 = 6{,}472$ papers with zero citations. Noting that the observed frequency $f_0 = M_0/M = 0.42$ is quite high, it was decided to omit the value at $j = 0$ in the GPL model fitting (as seen in Figure 2, this value does appear to be an outlier). According to (10), the hyper-parameters of the GPL are given by $m = (M - M_0)/K = 79.01$ and $n = 2{,}173.16$. In turn, the corresponding estimates for the shape parameters are $a = 2.50$ and $b = 0.60$, found numerically using the *optim* command in R.

The power-law model was fitted using a suitable truncation as explained after equation (8); the fitted values, obtained using the *poweRlaw* package in R (https://cran.r-project.org/web/packages/poweRlaw), are $a = 2.32$ and $j_* = 48$. As we will see, the goodness-of-fit of the power-law model is excellent, but a high value of j_* is disappointing. Note that if we opted to ignore truncation and tried to fit a power-law model in the entire range, the fitted value of the exponent would change to $a = 2.42$.

First, let us report the results for the GPL model regarding the match of the theoretical limit shape $y = \varphi(x; a)$ specified in equation (21). In Figure 1, this limit shape is plotted in scaled-back coordinates for a better comparison with the data (represented by an empirical Young diagram as explained above), that is, $y = B\ \varphi(x/A; a)$, with the scaling coefficients given by formula (19) and estimated from the data as $A = 13.75$ and $B = 158.08$.

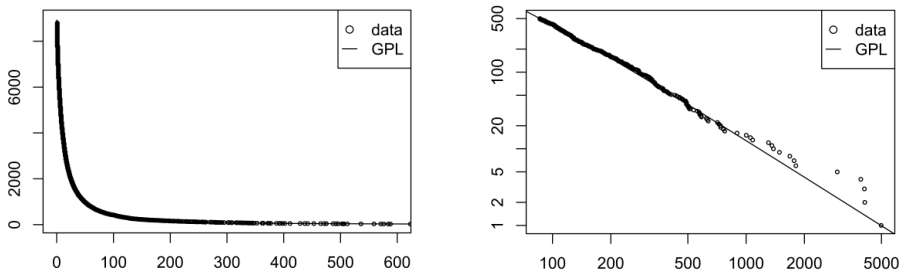

Figure 1. Observed data in the Young diagram representation and the fitted limit shape, with estimated parameters $a = 2.50$ and $b = 0.60$. The left panel illustrates a match for lower values of citations, while the right panel shows the tail comparison in the log-log coordinates.

We see from Figure 1 that the fit of the GPL is remarkably accurate, especially over a large initial part of the citation spectrum. To inspect details of the tail behavior, we use the log-log scale, revealing some minor discrepancies due to few extreme points.

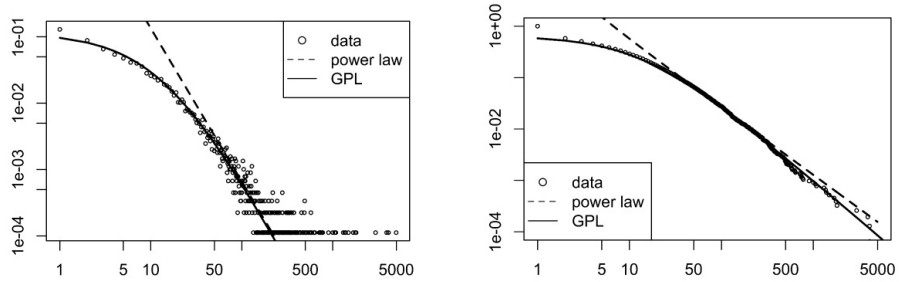

Figure 2. Log-log plots comparing data and the fitted models. The left panel shows frequencies f_j, while the right panel features complementary cumulative frequencies F_j. The left part of the dashed lines indicates an extrapolation of the power law below the truncation point $j_* = 48$. The observed frequency at $j = 0$ (zero citations) is an outlier.

Next, we compare the fit of both models to the data using standard frequency plots (Figure 2). Not surprisingly, the power law works extremely well at the tail but it is useless for smaller values of j. In contrast, the GPL strongly outperforms the power law over the initial range (despite a visible outlier at $j = 0$) but also works equally well at the tail.

Estimation of the h-index

Let us now look at what our models can tell about the h-index. According to Egghe & Rousseau (2006), in the non-truncated power-law model the h-index is estimated by the formula

$$h \approx m^{1/a}, \tag{22}$$

where $m = M/K$ is the mean number of papers per author. In our case, $m = 79.01$; using the non-truncated estimate $a = 2.42$ we get $h = 6.08$, while if we (formally) use the truncated fit $a = 2.31$ then the estimated value of h would slightly change to 6.63.
In the GPL model, the h-index geometrically corresponds to inscribing a biggest square inside the empirical Young diagram of citations. Using the limit shape (21) and scaling back using the coefficients A and B, we find that h is the approximate solution of the equation

$$h\left(\frac{h}{A}+1\right)^{a-1} = B(a-2). \tag{23}$$

Solving this equation numerically yields $h = 20.29$. Comparing with the empirical (mean) value $h = 17.52$, we see that the GPL estimation is superior to that of the power law, which cannot capture the true h-index lying deep below the truncation point $j_* = 48$.

Conclusion

In this paper, we have attempted to connect the pooled frequencies modeled via power law with individual citations per author. We have also introduced the generalized power-law (GPL) model which has been demonstrated to fit to the real data well in the entire range of citations, unlike the power law which frequently needs to get truncated at the beginning. In a real data set that we have studied, the GPL model reveals an inflated frequency of zero citations, which draws attention to a possible lack of impact in scientific output. An additional novel feature of the GPL model is that it possesses a deterministic limit shape which can be useful in estimating informative function of the citation data such as the h-index. Finally, it would be interesting to compare our GPL model with alternative fitting approaches (Sichel, 1985).

Acknowledgments

Work by Ruheyan Nuermaimaiti is part of her PhD research at the School of Mathematics, University of Leeds. The authors are grateful to Anna Senkevich for helpful insights concerning data retrieval from the web and to an anonymous referee for suggested bibliographic references.

References

Clauset, A., Shalizi, C.R. & Newman, M.E.J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51, 661–703.

Coile, R.C. (1977). Lotka's frequency distribution of scientific productivity. *Journal of the American Society for Information Science*, 28, 366–370.

Egghe, L. & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics*, 69, 121–129.

Hirsch, J.E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 16569–16572.

Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16, 317–323.

Nicholls, P.T. (1987). Estimation of Zipf parameters. *Journal of the American Society for Information Science*, 38, 443–445.

Price, D.J.D.S. (1965). Networks of scientific papers. *Science*, 149, 510–515.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *The European Physical Journal B – Condensed Matter and Complex Systems*, 4, 131–134.

Sichel, H.C. (1985). A bibliometric distribution which really works. *Journal of the American Society for Information Science*, 36, 314–321.

Vershik, A.M. (1996). Statistical mechanics of combinatorial partitions, and their limit shapes. *Functional Analysis and Its Applications*, 30, 90–105.