Research Paper

Modeling citation concentration through a mixture of Leimkuhler curves

Emilio Gómez-Déniz, Pablo Dorta-González *,1

Department of Quantitative Methods in Economics and TIDES Institute, University of Las Palmas de Gran Canaria, Spain

ARTICLE INFO

ABSTRACT

When a graphical representation of the cumulative percentage of total citations to articles, ordered from most cited to least cited, is plotted against the cumulative percentage of articles, we obtain a Leimkuhler curve. In this study, we noticed that standard Leimkuhler functions may not be sufficient to provide accurate fits to various empirical informetrics data. Therefore, we introduce a new approach to Leimkuhler curves by fitting a known probability density function to the initial Leimkuhler curve, taking into account the presence of a heterogeneity factor. As a significant contribution to the existing literature, we introduce a pair of mixture distributions (called PG and PIG) to bibliometrics. In addition, we present closed-form expressions for Leimkuhler curves. Some measures of citation concentration are examined empirically for the basic models (based on the Power and Pareto distributions) and the mixed models derived from these. An application to two sources of informetric data was conducted to see how the mixing models outperform the standard basic models. The different models were fitted using non-linear least squares estimation.

1. Introduction

Power law relationships or similar long-tailed distributions are pervasive in various domains of informetrics, such as authors and their publications (Lotka, 1926), topics and journals related to them (Bradford, 1934), web sites and their links (Faloutsos et al., 1999), or research literature and its obsolescence (Dorta-González & Gómez-Déniz, 2022). Furthermore, these types of distributions are widely observed in diverse fields, including demography (cities and their inhabitants), linguistics (words and their uses), economics (incomes in a market economy), ecology (forest fragmentation), astronomy (initial mass functions), and more (Newman, 2005; Pareto, 1895; Salpeter, 1955; Saravia et al., 2018; Zipf, 1941). The ubiquity of power laws is evident across different domains, showcasing their widespread presence in various disciplines.

Power laws are mathematical relationships that describe a phenomenon in which a small number of elements have a disproportionately high impact or frequency compared to the majority. In the context of citation networks, power laws suggest that a few highly cited papers have a much greater impact than the majority of less cited papers. This phenomenon is known as preferential attachment or cumulative advantage (Brzezinski, 2015). Some other papers on the application of power laws to citations are those of Thelwall and Wilson (2014), Thelwall (2016a), and Thelwall (2016b). The distribution of citations among scientific papers often

---

* Corresponding author.
  E-mail address: pablo.dorta@ulpgc.es (E. Gómez-Déniz).
1 Department of Quantitative Methods in Economics, University of Las Palmas de Gran Canaria, 35017 Las Palmas de Gran Canaria, Spain.

follows a power law distribution, which is characterized by a straight line on a log-log plot. This means that while most papers receive few citations, a small fraction receive an exceptionally large number of citations and contribute significantly to the overall impact of the field. This pattern is similar to the well-known "80/20" rule, where roughly 80% of citations are contributed by 20% of papers.

Lorenz and Leimkuhler curves have proven to be useful tools in bibliometrics for analyzing the distribution and growth of scientific output. These curves can provide insights into the concentration of research output, the distribution of citations, and the factors influencing the growth rate of scientific output. For example, a study by Abramo et al. (2011) used Lorenz curves to analyze the concentration of research output among Italian universities. Egghe (2006) examined the distribution of citations among scientific papers. He found that the distribution was highly skewed, with a few papers receiving a disproportionate number of citations. More recently, Nair and Vineshkumar (2022) derived novel properties of Leimkuhler curves by employing quantile functions and applied these findings to the informetrics domain.

These distributions share the common characteristic of having a few sources with a high concentration of items, leading to the study of inequality or concentration as a major focus of the field (Rousseau et al., 2018). The social implications of this are significant, prompting research into techniques for quantifying and modeling inequality. One such technique is Leimkuhler ordering, which provides a partial ordering of non-negative random variables based on their concentration properties. This allows the comparison of distributions and the identification of more concentrated ones.

An area of interest within the field of informetrics, particularly in bibliometrics, revolves around the challenge of measuring journal productivity through citation analysis to identify a core of influential journals that are at the forefront of research in a particular field. This issue concerns the distribution pattern of citations among different journals over a given period of time and how to rank journals based on the decreasing number of citations received. Research on this topic has received considerable attention in recent decades. For a detailed study of this problem, see for example Hubert (1977). If a graphical representation of the cumulative percentage of the total sample of cited papers, ordered from largest to smallest, is plotted against the cumulative percentage of journals, we get a Leimkuhler curve, or double cumulative curve (the curve starts at the largest value) of a productivity distribution, as it is known in the informetric literature. It is well known that this tool is related to the graph of the Lorenz curve used in economics to model income distribution (see Burrell, 2005; Sarabia, 2008a, 2008b; and Sarabia et al., 2010, among others).

In this research, our findings suggest that conventional Leimkuhler functions may not adequately capture the complexities of diverse empirical informetrics data. To address this limitation, we propose a novel approach to Leimkuhler curves that involves fitting a known probability density function to the original Leimkuhler curve, while accounting for the presence of a heterogeneity factor. We then analyze different measures of citation concentration using the traditional basic model in the literature, based on Power and Pareto distributions (Egghe, 2005), and the mixed models derived from it. To assess the effectiveness of our approach, we apply it to two sources of informetric data and compare the performance of the mixed models with the standard basic models.

The structure of this paper is as follows. Section 2 deals with the standard definition of a Leimkhuler curve and provides two standard Leimkhuler curves, the one based on the power distribution and the one based on the Pareto distribution. Two numerical examples are provided to show how these two basic models work. The mixture model of Leimkhuler curves is shown in section 3. Here we also give the expression for the Gini and Pietra indices, two measures of citation concentration. Numerical applications are given in section 4 and conclusions in the last section 5.

2. The standard Leimkuhler curve

The Leimkuhler curve is a crucial instrument in information processing, especially in informetrics (see for instance, Burrell, 2005). This curve is closely associated with the Lorenz curve, differing only in that the contributions are arranged in a descending order along the Leimkuhler curve. This curve plots the cumulative proportion of total citations against the cumulative proportion of sources. Specifically, given a set of n ordered numbers $x_1 \geq x_2 \geq \cdots \geq x_n \geq 0$, the Leimkuhler curve K obtained from them is defined as the points i/n, $i = 0, 1, \ldots, n$ by $K(0) = 0$ and $K(i/n) = s_i/s_n$, where $s_i = \sum_{j=1}^{i} x_j$. Therefore, $K(i/n)$ represents the fraction of the total s_n contributed by the ith largest numbers. Thus, by the fraction i/n of the largest numbers, the Leimkuhler curve is defined for all $u \in [0, 1]$ by linear interpolation between adjacent point $(i/n, K(i/n))$, thus producing a polygonal or broken-line graph in rectangular coordinates from the origin to the end $(1, 1)$.

The following formal definition of a Leimkuhler curve is due to Sarabia (2008a).

Definition 1. Let $X \in \mathcal{X}$ a non-negative random variable with cumulative distribution function $F_X(x) = \Pr(X \leq x)$ and inverse distribution function $F_X^{-1}(z) = \inf\{x : F_X(x) \geq z\}$. The Leimkuhler curve $K_X(u)$, $0 \leq u \leq 1$, corresponding to X is defined by

$$K_X(u) = \frac{1}{\mu_X} \int_{1-u}^{1} F_X^{-1}(y)\, dy, \quad 0 \leq u \leq 1, \tag{1}$$

where $\mu_X = \int_{\mathcal{X}} x\, dF_X(x) < \infty$, the mathematical expectation of X, is assumed to be finite.

The definition above allows for the continuous and discrete case. Suppose that $K_X(u)$ given in (1) has first and second derivatives and that F_X has a continuous nonzero first derivative. Then $K_X(u)$ satisfies the following properties (see for instance Sarabia, 2008a):

(i) $K_X(0) = 0$, $K_X(1) = 1$.

(ii) $K_X(u)$ is a non-decreasing function.
(iii) $K_X(u)$ is a concave function.

Other alternatives definitions of a Leimkuhler curve can be seen in Sarabia (2008a). Furthermore, taking into account that Leimkuhler and Lorenz curves (say $L_X(u)$) are related via

$$K_X(u) = 1 - L_X(1-u) \tag{2}$$

the proof of the Theorem above can be carried out by taking into account the characterization of a Lorenz curve given, among others, in Theorem 1 in Sarabia et al. (2010) and Theorem 2 in Sordo et al. (2014).

The counterpart of the Gini index (see Burrell, 1991, 1992) defined for Lorenz curves to the Leimkuhler curve is given by

$$G(\theta) = 2 \int_0^1 K_X(u;\theta)\,du - 1. \tag{3}$$

This index takes values in the unit interval and measures the subject category's egalitarianism. The lower the index, the less egalitarian the subject category; the higher the index, the more egalitarian the subject category.

The authors Kakwani (1980) and Yitzhaki (1983) proposed a generalization of the Gini index that applies to Lorenz curves. Their proposed measure, which has been adapted for the Leimkuhler curve, is known as the generalized Gini index. It is defined by the equation

$$G_r(\theta) = r(r+1) \int_0^1 (1-u)^{r-1} K_X(u;\theta)\,du - 1, \quad r > 0. \tag{4}$$

The parameter r plays a crucial role in determining the sensitivity of the generalized Gini index to changes in the income distribution. Specifically, when $r \to 0$, changes at the upper end of the distribution have the only impact on the value of the index and this index does not distinguish among Leimkuhler curves. Conversely, as r increases towards infinity, the index becomes increasingly sensitive to changes at the lower end of the distribution. In the limit, the value of the index is determined exclusively by the lowest income, reflecting the notion that the overall social welfare is contingent solely on the well-being of the poorest member of society. It is worth noting that the standard Gini index, as represented by equation (3), is retrieved when r is set to 1.

On the other hand, the Pietra index (Eliazar & Sokolov, 2010), defined as

$$P(\theta) = \max_{0 \le u \le 1} |K_X(u;\theta) - u|, \tag{5}$$

i.e. the distance between the Leimkuhler curve and the line of perfect equality.

Although, in practice, the citation distributions are integer-valued, despite the variable of interest in the paper being discrete, the study is limited to considering distributional modeling with continuous approximations.

This work does not intend to keep the distribution of the number of citations. In this sense, classical discrete distributions, such as the Poisson and negative binomial distributions, are the appropriate modeling for this. We will follow the spirit of Leimkuhler's work (see Leimkuhler, 1967) in which the curve named after him will be used to adjust the fraction of documents, x, that are most productive, i.e., $0 < x < 1$, while $F_X(x)$ denotes the proportion of total productivity contained in that fraction. It will be shown that the Leimkuhler curve constructed from (1), regardless of the distribution function from which it is based and its domain or support is a valuable tool for the intended purpose. Later, this task can be improved by enriching the Leimkuhler curve on which it is based.

Example 1. Let Z be a power distribution with cumulative distribution function given by $F_Z(z) = z^{1/\theta}$, $z \in (0,1)$, $\theta > 0$ and mean $\mu_Z = (1+\theta)^{-1}$. Now, from (1) we get the corresponding Leimkuhler curve given by (this model has been considered by Sarabia, 2008b)

$$K_Z(u;\theta) = 1 - \bar{u}^{1+\theta}, \tag{6}$$

where $\bar{u} = 1 - u$, $u \in [0,1]$. A simple generalization of this Leimkuhler curve can be obtained from the Lorenz curve given in Sarabia et al. (1999) by using (2) (see expression (15) in that paper) and is given by

$$K_Z(u;\theta,k) = 1 - (1 - u^k)\bar{u}^{\theta}, \tag{7}$$

where $0 < k \le 1$. Note that when $k = 1$ expression (7) reduces to (6) as a special case. The corresponding Gini indices to these Leimkuhler curves are given by

$$G(\theta) = \frac{\theta}{2+\theta}, \tag{8}$$

$$G(\theta,k) = \frac{2\Gamma(k+1)\Gamma(\theta+1)}{\Gamma(\theta+k+2)} + \frac{\theta-1}{\theta+1}, \tag{9}$$

for (6) and (7), respectively.

Table 1
Descriptive statistics of the variable number of citations.

	Stat. & Prob.	OR & MS
	cites	cites
min	0	0
max	908	84
Mean	13.81	14.08
Variance	273.90	50.41
ID	19.83	3.58
total citations	94267	83649
n	6826	5941

Now, from (4) we get

$$G_r(\theta) = \frac{r\theta}{1 + r + \theta},$$

for the Leimkuhler curve given in (6). An expression of this index, which is not reproduced here, can be obtained for (7) in terms of the incomplete beta function.

On the other hand, the Pietra index given in (5), for (6), is reached at $u = 1 - (1 + \theta)^{-1/\theta}$ and is given by

$$P(\theta) = \theta(1 + \theta)^{-1/\theta - 1}. \tag{10}$$

Numerical computation is needed to get the Pietra index for the Leimkuhler curve given in (7). These two Leimkuhler curves will be denoted as the power Leimkuhler curve and the generalized power Leimkuher curves, respectively. □

Example 2. Let Z be a classical Pareto distribution with cumulative distribution function given by $F_Z(z) = 1 - (z/\sigma)^{-1/\theta}$, $z \geq \sigma > 0$, $0 < \theta < 1$ and mean $\mu_Z = \sigma(1 - \theta)^{-1}$. Now, from (1) we get the corresponding Leimkuhler curve given by (this model has been considered by Sarabia, 2008a)

$$K_Z(u; \theta) = u^{1-\theta}. \tag{11}$$

The corresponding Gini index to this Leimkuhler curve is given by

$$G(\theta) = \frac{\theta}{2 - \theta}. \tag{12}$$

Now, from (4) we get

$$G_r(\theta) = \frac{\Gamma(2 + r)\Gamma(2 - \theta)}{\Gamma(2 + r - \theta)} - 1.$$

On the other hand, the Pietra index given in (5) is reached at $u = (1 + \theta)^{1/\theta}$ and is given by

$$P(\theta) = \theta(1 + \theta)^{1/\theta - 1}. \tag{13}$$

In advance this Leimkuhler curve will be denoted as the Pareto Leimkuhler curve. □

We have considered two data sources to see how these two models work. They correspond to two JCR subject categories in the WOS database: Operations Research & Management Science (OR & MS), and Statistic & Probability (Stat. & Prob.), and some descriptive statistics of them are shown in Table 1. The total number of "research articles" in 2016, referred to these subject categories (OR & MS and Stat. & Prob.) was 12,767. In addition, we identified all citations received within a 5-year citation window (2016-2020).

To estimate the parameters in equations (6) and (11), we employed a non-linear least squares fit. To derive our results, we adopted a methodology that entailed minimizing the sum of squared discrepancies between the empirical data points of the Leimkuhler curve and the corresponding values predicted by the theoretical models. More specifically, if we denote a Leimkuhler curve as $K_X(u; \theta)$, the objective of the analysis can be expressed mathematically as a minimization problem:

$$\min_{\theta} \sum_{i=1}^{n} \left(\frac{s_i}{s_n} - K_X\left(\frac{i}{n}; \theta\right) \right)^2,$$

that is, the sum of the squared discrepancies between the empirical data points of the Leimkuhler curve, $K(u_i) = s_i/s_n$, $u_i = i/n$ and the corresponding values predicted by the theoretical model, $K_X(u_i; \theta)$. Here θ represents the parameter vector to be estimated. The non-linear least squares method allows for the estimation of non-linear parameters, such as those in equations (6), (7) and (11) by iteratively minimizing the sum of the squared differences between the theoretical and empirical curves until convergence is reached.

Table 2
Estimated parameters (standard error between parenthesis), using non-linear least square, for the different Leimkuhler curves considered and some statistics.

	Model	Estimated parameters				MSE	MAX	MAE	CAIC	Gini	Pietra
		$\widehat{\theta}$	$\widehat{k}$	$\widehat{\alpha}$	$\widehat{\beta}$						
Stat. & Prob.	Power	3.832 (0.0131)				0.0027	0.0835	0.0471	-20778.40	0.6571	0.5276
	PG			0.701 (0.0275)	0.102 (0.0058)	0.0003	0.0318	0.0149	-36227.40	0.5910	0.4574
	PIG			9.305 (0.0275)	2.227 (0.0058)	$7.7 \cdot 10^{-5}$	0.0169	0.0076	-45277.80	0.6011	0.4536
	GP	1.859 (0.0042)	0.514 (0.0005)			$1.3 \cdot 10^{-4}$	0.0319	0.0089	-41525.70	0.6204	0.4636
	GPG		0.554 (0.0007)	1.514 (0.0184)	0.596 (0.0096)	$6.7 \cdot 10^{-5}$	0.0236	0.0060	-46149.90	0.6071	0.4608
	GPIG		0.571 (0.0010)	2.919 (0.0182)	2.416 (0.0230)	$5.9 \cdot 10^{-5}$	0.0209	0.0061	-46983.00	0.6070	0.4591
	Pa	0.645 (0.0011)				0.0067	0.1244	0.0730	-14744.30	0.4756	0.3373
	PaGB		-25.483 (0.0360)	11.042 (0.1543)	14.080 (0.1381)	0.0078	0.4231	0.0797	-15703.10	0.4984	0.3840
OR & MS	Power	2.767 (0.0140)				0.0044	0.0968	0.0601	-15361.20	0.5804	0.4548
	PG			0.392 (0.0013)	0.055 (0.0003)	$1.5 \cdot 10^{-5}$	0.0261	0.0108	-35415.90	0.5021	0.3838
	PIG			14.035 (0.033)	1.029 (0.0008)	$9.2 \cdot 10^{-6}$	0.0096	0.0024	-52011.20	0.5188	0.3812
	GP	0.928 (0.004)	0.492 (0.0005)			0.0002	0.0324	0.0110	-34430.30	0.5282	0.3803
	GPG		0.558 (0.0008)	0.445 (0.0028)	0.212 (0.0026)	$4.1 \cdot 10^{-5}$	0.0170	0.0052	-43094.90	0.5115	0.3849
	GPIG		0.799 (0.0039)	10.765 (0.1181)	0.742 (0.0056)	$7.5 \cdot 10^{-6}$	0.0088	0.0019	-53166.10	0.5165	0.3813
	Pa	0.606 (0.0009)				0.0028	0.0783	0.0478	-18027.20	0.4344	0.3305
	PaGB		-28.083 (0.096)	45.010 (1.967)	41.080 (1.369)	0.0029	0.473	0.0489	-18506.80	0.4307	0.3276

PG: Mixture power-gamma Leimkuhler curve; PIG: Mixture power-inverse Gaussian Leimkuhler curve.
GPG: Mixture generalized power-gamma Leimkuhler curve; GPIG: Mixture generalized power-inverse Gaussian Leimkuhler curve.
PaGB: Mixture Pareto-generalized beta Leimkuhler curve.

Table 2 shows the estimation results for the three models considered given in equations (6), (7) and (11), respectively. To estimate the standard errors of the coefficients, we used a consistent estimate of the covariance matrix, as described in Amemiya (1985). We also provide the consistent Akaike's information criteria (CAIC), proposed by Bozdogan (1987). The CAIC is expressed as $k(1 + \log n) - 2 \log \ell$. This last measure of model selection was chosen to overcome the tendency of the AIC to overestimate the complexity of the underlying model since it lacks certain properties of asymptotic consistency as it does not directly depend on the sample size. Then, to calculate the CAIC, a correction factor based on the sample size is used to compensate for the overestimating nature of AIC. Note that a model with a lower CAIC value is preferred to one with a higher value. To assess the goodness of fit among the models, various measures were employed, including the mean squared error (MSE), the maximum absolute error (MAX), and the mean absolute error (MAE). These measures are formulated as follows:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{s_i}{s_n} - K_X(u_i; \widehat{\theta}) \right]^2,$$

$$\text{MAX} = \max_{i=1,\ldots,n} \left| \frac{s_i}{s_n} - K_X(u_i; \widehat{\theta}) \right|,$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{s_i}{s_n} - K_X(u_i; \widehat{\theta}) \right|.$$

We also present the Gini and the Pietra indices obtained by using expressions provided in (8), (9), (12), (10) and (13). We used a code implemented in Wolfram Mathematica (v.13.0) to compute the estimated parameters, as well as the MSE, MAX, MAE, Gini, and Pietra indices. The code is straightforward and easy to use. It is worth noting that a plot of the CDF of the residuals against the CDF of a normal distribution has been represented in all the cases considered, not resulting in a straight line. Therefore, the residuals appear to violate the normality hypothesis. Hence, maximum likelihood estimation seems more appropriate.
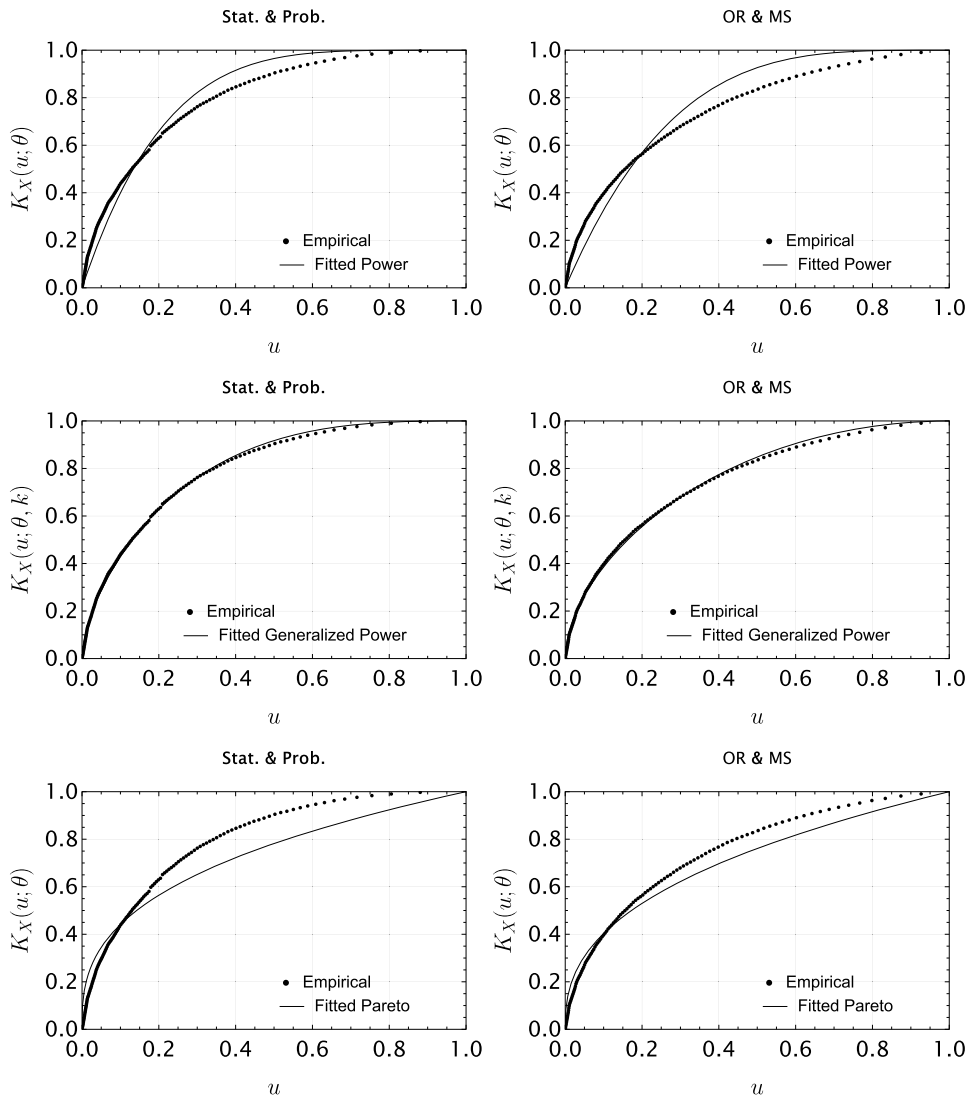
Fig. 1. Empirical and fitted, by non-linear least square, Leimkuhler curves based on the data considered. From top to down they correspond to equations given in (6), (7) and (11), respectively.

In Fig. 1 the cumulative number of citations, $K_X(u)$, is plotted against the cumulative proportion, u, of articles and compared with the estimated Leimkuhler curves given in (6), (7) and (11) for the sources of data, respectively. The graphs show that standard models accurately represent the empirical Leimkuhler curves.

3. Beyond the standard model

Empirical evidence suggests that a simple model of this nature may be too inflexible and insufficient in explaining the phenomenon under study. It is more appropriate to assume that the phenomenon is heterogeneous and that a mixture model may better capture the underlying complexity. Furthermore, the contagious and other missing characteristics of the data suggest that the pattern of citation behavior varies intrinsically across different journals in the subject category.

In practice, over-dispersion, i.e. the variance is greater than the mean, arises due to the heterogeneity of the population, meaning that individuals have constant but unequal probabilities of experiencing an event. In this case, each member of the population has a value of the parameter $\theta \in \Theta$, which follows a probability density function $g(\theta)$. Therefore, θ can be considered an inhomogeneity parameter that varies from journal to journal, following a probability density function $g(\theta; \Upsilon)$, and the actual Leimkuhler curve can be constructed in this way. Here, Υ represents the hyperparameters of the mixing distribution.

The use of mixture models in informetric and bibliometrics is familiar and has been used to get long-tail citation distributions. See for example Burrell and Cane (1982) and Burrell (2005), among others. It is widely recognized that informetric data poses some specific challenges because they tend to exhibit a heavily skewed distribution.

In short, although the basic model may be exact within a group of journals, the parameter θ frequently will vary randomly across groups of journals because of random variation in citations and articles considered, for example. Both of these data sources exhibit empirical over-dispersion, meaning that their variance is greater than their mean. Table 1 displays the index of dispersion $(ID = var(X)/\mathbb{E}(X))$, as well as the mean and variance of the empirical data for both the citations and articles variables across the two subject categories of journals. Furthermore, there is evidence to suggest that the over-dispersion may be linked to the heterogeneity of the population across subject categories. In such cases, the parameter θ can be treated as a random variable that assumes different values across various journals within the subject category, thus capturing the inherent uncertainty associated with this parameter, and varying from individual to individual based on a probability density function.

Then, we propose the following Leimkuhler curve instead of the one provided in (1).

Proposition 1. If $K_X(u;\theta)$ is a Leimkuhler curve obtained from (1), then

$$K_X(u;\Upsilon) = \mathbb{E}_{\Theta}[K_X(u;\theta)] = \int_{\Theta} K_X(u;\theta)g(\theta;\Upsilon)\,d\theta \tag{14}$$

is also a proper Leimkuhler curve.

Proof. The result is direct. ☐

The Gini index in this case can be computed also by mixture as we show in the next result.

Proposition 2. If $K_X(u;\Upsilon)$ is a proper Leimkuhler curve obtained from (14), then the Gini index is given by

$$G(\Upsilon) = \mathbb{E}_{\Theta}[G(\theta)] = \int_{\Theta} G(\theta)g(\theta;\Upsilon)\,d\theta. \tag{15}$$

Proof. It is simple. ☐

3.1. Specific examples

Note that the Power Leimkuhler curve provided in (7) can be rewritten as

$$K_X(u;\theta) = 1 - (1 - u^k)\exp(\theta\log\bar{u}), \quad 0 \le u < 1. \tag{16}$$

Thus, attending to (16) any probability density function with a closed form for its moment generating function, such as the gamma, the inverse Gaussian, the generalized inverse Gaussian, and Rayleigh distributions should be good candidates to introduce here.

3.1.1. The gamma case

Assume first that the parameter $\theta \in \Theta$ follows a gamma distribution with shaper parameter $\alpha > 0$ and rate parameter $\beta > 0$, i.e. $g(\theta;\Upsilon) \propto \theta^{\alpha-1}\exp(-\beta\theta)$ where $\theta > 0$ and $\Upsilon = (\alpha, \beta)$. It is well-known that its moment generating function is given by

$$M_{\Theta}(t;\Upsilon) = \mathbb{E}[\exp(t\Theta)] = \left(\frac{\beta}{\beta - t}\right)^{\alpha}. \tag{17}$$

Now, using (6), (14) and (17) it is straightforward to get the following mixture Leimkuhler curve

$$K_X(u;\Upsilon) = 1 - \bar{u}\psi_G(u;\Upsilon), \quad 0 \le u < 1. \tag{18}$$

$$K_X(u;\Upsilon,k) = 1 - (1 - u^k)\psi_G(u;\Upsilon), \quad 0 \le u < 1, \tag{19}$$

where

$$\psi_G(u;\Upsilon) = \left(\frac{\beta}{\beta - \log\bar{u}}\right)^{\alpha}.$$

The special case for which $\alpha = 1$ derives in the mixture Leimkuhler curve obtained from the exponential mixing distribution.

From (15) we get that the mixture Gini index results

$$G(\Upsilon) = \alpha(2\beta)^{\alpha}\Gamma(-\alpha, 2\beta)\exp(2\beta),$$

where $\Gamma(a, b) = \int_b^{\infty} t^{a-1}\exp(-t)\,dt$ is the incomplete gamma function.

$$G_r(\Upsilon) = \alpha r((1 + r)\beta)^{\alpha}\Gamma(-\alpha, (1 + r)\beta)\exp[(1 + r)\beta].$$

In advance (18) will be referred as the PG, denoting that is the mixture of a power Leimkuhler curve with the gamma distribution and (19) as GPG.

3.1.2. The inverse Gaussian case

Consider now that Θ follows an inverse Gaussian distribution with probability density function (see Folks & Chhikara, 1978; and Seshadri, 1983) given by

$$g(\theta; \Upsilon) = \sqrt{\frac{\beta}{2\pi\theta^3}} \exp\left[-\frac{\beta}{2\alpha^2\theta}(\theta - \alpha)^2\right], \quad \theta > 0,$$

where $\alpha > 0$ is the mean of the distribution and $\beta > 0$. Its variance is given by α^3/β and the moment generating function is given by

$$M_\Theta(t; \Upsilon) = \mathbb{E}[\exp(t\Theta)] = \exp\left[\frac{\beta}{\alpha}\left(1 - \sqrt{1 - 2\alpha^2 t/\beta}\right)\right]. \tag{20}$$

Using now (6), (14) and (20) we have the mixture Leimkuhler curve given by

$$K_X(u; \Upsilon) = 1 - \bar{u}\exp[\psi_{IG}(u; \Upsilon)], \quad 0 \le u < 1, \tag{21}$$

$$K_X(u; \Upsilon, k) = 1 - (1 - u^k)\exp[\psi_{IG}(u; \Upsilon)], \quad 0 \le u < 1, \tag{22}$$

where

$$\psi_{IG}(u; \Upsilon) = \frac{\beta}{\alpha}\left(1 - \sqrt{1 - 2\alpha^2 \log \bar{u}/\beta}\right).$$

We will use the term PIG to denote the Leimkuhler curve given in (21) because it is the power Leimkuhler curve with the inverse Gaussian distribution and GPIG for the Leimkuhler curve given in (22). In this case, numerical computation is required to get the Gini index since non-closed-form expressions exist for (15).

3.2. The Pareto-confluent hypergeometric distribution case

The confluent hypergeometric distribution due to Gordy (1998) has a probability density function given by

$$g(\theta; \Upsilon) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}\exp(-k\theta)}{B(\alpha, \beta)\,_1F_1(\alpha; \alpha + \beta; -k)}, \quad 0 < \theta < 1, \tag{23}$$

where $\Upsilon = (\alpha, \beta, k)$ with $\alpha > 0$, $\beta > 0$, $-\infty < k < \infty$ and $_1F_1(\cdot; \cdot; \cdot)$ is the confluent hypergeometric function (see Abramowitz & Stegun, 1968, p. 505), given by

$$_1F_1(a; b; z) = \frac{\Gamma(b)}{\Gamma(b-a)\Gamma(a)}\int_0^1 t^{a-1}(1-t)^{b-a-1}\exp(zt)\,dt.$$

If $k = 0$, then this density reduces to the standard beta density. Now, by using (11), (14) and (23) we get the mixture Leimkuhler curve given by

$$\begin{aligned}
K_X(u; \Upsilon) &= \int_0^1 u\exp(-\theta\log u)\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}\exp(-k\theta)}{B(\alpha, \beta)\,_1F_1(\alpha; \alpha + \beta; -k)}\,d\theta \\
&= \frac{u\,_1F_1(\alpha; \alpha + \beta; -k - \log u)}{_1F_1(\alpha; \alpha + \beta; -k)} = \frac{_1F_1(\beta; \alpha + \beta; k + \log u)}{_1F_1(\beta; \alpha + \beta; k)}, \quad 0 \le u \le 1,
\end{aligned} \tag{24}$$

where we have used the Kummer transformation (see Abramowitz & Stegun, 1968, p. 505) $_1F_1(a; b; -c) = \exp(-c)\,_1F_1(b - a; b; c)$.

In advance (24) will be referred as the PaGB, denoting that is the mixture of a Pareto Leimkuhler curve with the confluent hypergeometric distribution (generalization of the standard beta distribution).

When $k = 0$ the Leimkuhler curve given in (24) reduces to

$$K_X(u; \Upsilon) = \,_1F_1(\beta; \alpha + \beta; \log u), \quad 0 \le u \le 1,$$

which corresponds to the mixture of the Leimkuhler curve given in (11) with the standard beta density.

3.3. Stochastic ordering

We are examining the Leimkuhler ordering in this study. The Leimkuhler curve can establish an unambiguous ordering when two distribution functions have non-intersecting Leimkuhler curves (see Atkinson, 1970, Dasgupta et al., 1973, and Shorrocks, 1983, among others). The class $\mathcal{L}$ is defined as all non-negative random variables with a positive finite expectation, and the Leimkuhler partial order $\le_L$ is defined on the class $\mathcal{L}$ by

$$X \le_L Y \iff F_X(x) \le F_Y(y) \iff K_X(u) \le K_Y(u), \forall u \in [0, 1].$$

If $X \leq_L Y$, then X exhibits less inequality than Y in the Leimkuhler sense. The next result shows that the family of Leimkuhler curves given in (18) and (21) are ordered with respect to the Υ vector of parameters.

Proposition 3. *Let $\Upsilon_1 = (\alpha_1, \beta_1, k)$, $\Upsilon_2 = (\alpha_2, \beta_2, k)$, then the Leimkuhler curves given in (18) and (19) are ordered with respect to α and β, i.e.*

(i) if $\alpha_1 < \alpha_2$ we have that $K_X(u; \Upsilon_1) \geq K_X(u; \Upsilon_2)$, $\forall \beta > 0$, $\forall k \in (0, 1]$.
(ii) if $\beta_1 < \beta_2$ we have that $K_X(u; \Upsilon_1) \leq K_X(u; \Upsilon_2)$, $\forall \alpha > 0$, $\forall k \in (0, 1]$.

Proof. By computing the derivative of (18) or (19) with respect to α and β we get after some simple computations that

$$\frac{\partial K_X(u; \Upsilon)}{\partial \alpha} = -\left[1 - K_X(u; \Upsilon)\right] \log\left(\frac{\beta}{\beta - \log \bar{u}}\right) > 0,$$

$$\frac{\partial K_X(u; \Upsilon)}{\partial \beta} = \frac{\left[1 - K_X(u; \Upsilon)\right] \alpha \log \bar{u}}{\beta(\beta - \log \bar{u})} < 0,$$

respectively. Hence the result. $\square$

Proposition 4. *Let $\Upsilon_1 = (\alpha_1, \beta_1, k)$, $\Upsilon_2 = (\alpha_2, \beta_2, k)$, then the Leimkuhler curves given in (21) and (22) are ordered with respect to α and β, i.e.*

(i) if $\alpha_1 < \alpha_2$ we have that $K_X(u; \Upsilon_1) \geq K_X(u; \Upsilon_2)$, $\forall \beta > 0$, $\forall k \in (0, 1]$.
(ii) if $\beta_1 < \beta_2$ we have that $K_X(u; \Upsilon_1) \geq K_X(u; \Upsilon_2)$, $\forall \alpha > 0$, $\forall k \in (0, 1]$.

Proof. Again, some computations provide the following derivative of (21) or (21) with respect to α and β,

$$\frac{\partial K_X(u; \Upsilon)}{\partial \alpha} = -\left[1 - K_X(u; \Upsilon)\right] \frac{\beta}{\alpha^2} \left[-\left(1 - \sqrt{1 - 2\alpha^2 \log \bar{u}/\beta}\right) + \frac{2\alpha^2 \log \bar{u}/\beta}{\sqrt{1 - 2\alpha^2 \log \bar{u}/\beta}}\right] > 0,$$

$$\frac{\partial K_X(u; \Upsilon)}{\partial \beta} = -\frac{\left[1 - K_X(u; \Upsilon)\right]}{\alpha} \left(1 - \sqrt{1 - 2\alpha^2 \log \bar{u}/\beta} - \frac{\alpha^2 \log \bar{u}/\beta}{\sqrt{1 - 2\alpha^2 \log \bar{u}/\beta}}\right) > 0,$$

respectively. Thus, the result follows. $\square$

Proposition 5. *Let $\Upsilon_1 = (\alpha, \beta, k_1)$, $\Upsilon_2 = (\alpha, \beta, k_2)$, then the Leimkuhler curves given in (19) and (22) are ordered with respect to k, i.e.*

(i) if $k_1 < k_2$ we have that $K_X(u; \Upsilon_1) \leq K_X(u; \Upsilon_2)$, $\forall \alpha, \beta > 0$.

Proof. Again, some computations provide the following derivative of (19) or (22) with respect to k,

$$\frac{\partial K_X(u; \Upsilon)}{\partial k} = \frac{\left[1 - K_X(u; \Upsilon_1)\right] k u^{k-1}}{1 - u^k} > 0.$$

Hence, the result follows. $\square$

4. Calculations

Table 2 displays the estimation results for the models considered here. The computational time needed for obtaining parameter estimates and model comparison measures is minimal compared to standard models.

As evidenced by the tables, the novel models exhibit superior performance with in general smaller values of MSE, MAX and MAE. Furthermore, Gini and Pietra indices are nearer to the empirical values in comparison with the initial Leimkuhler curves (0.6112 and 0.4625 for Stat. & Prob. data, and 0.5166 and 0.3803 for OR & MS data, respectively). Notably, the new proposed functional forms yield the best fit in this comparison, with the exception of the Pareto and Pareto mixture model which work worst than the rest of the models.

Furthermore, we provided a visual representation of the performance of each model in Fig. 2. As expected, the mixture models fit the empirical Leimkuhler curve reasonably well and outperform the Power Leimkuhler models shown in Fig. 1. In particular, the mixture based on the inverse Gaussian distribution is better.

In our study, we used a least squares approach to estimate the parameters of our model. It is worth noting that under the assumption of the normal distribution of the residuals, a basic assumption, many statistical packages allow the calculation of AIC and from this the CAIC. Therefore, although we did not directly use maximum likelihood estimates, we used these statistics for comparative purposes to quantify the improvement obtained in our model.
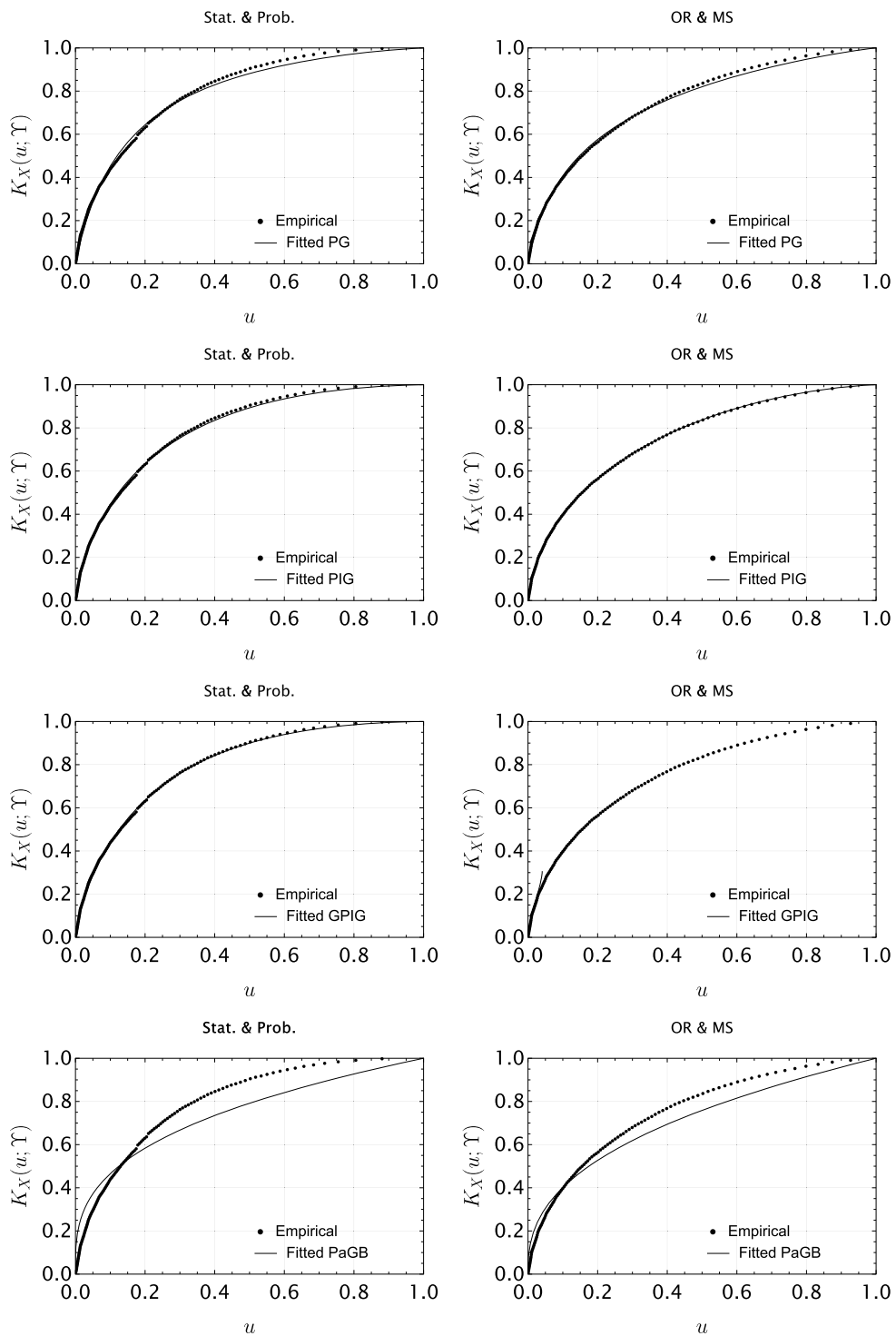
Fig. 2. Empirical and fitted, by non-linear least square, Leimkuhler curves based on the data considered and mixture models. The left and right sides correspond to the mixture power-gamma (PG) and mixture power-inverse Gaussian (PIG) Leimkuhler curves, given in equations (18) and (21), respectively.

In addition, we recognize the challenge posed by the difficulty of visually identifying subtle differences in the curves from direct observation of the graphs, especially when they appear to be closely clustered. Therefore, the use of CAIC statistic provides a quantitative measure to complement visual inspection and aid in the comparison of model fits.

We found higher citation concentration indices, as measured by the Gini and Pietra indices, in the case of Stat. & Prob. Several factors may explain these differences, as discussed below.

Citation frequency and volume: Research in OR & MS often involves the application of mathematical models, optimization techniques, and decision frameworks to practical problems. This often leads to more frequent citations as practitioners and researchers refer to specific methods and solution approaches. In contrast, citations in the field of Stat. & Prob. may be less frequent, especially for well-established theories and fundamental concepts. However, when new methodologies, algorithms, or data analysis techniques are introduced, citations may become more substantial (Adler et al., 2009; Ji et al., 2022).

Citation purposes: Citations in OR & MS are often used to demonstrate the effectiveness and feasibility of particular optimization techniques, decision models, or algorithms. These citations often support the implementation of solutions in real-world scenarios. On the other hand, citations in the area of Stat. & Prob. often serve to provide the theoretical foundation for data analysis methods, probability models, and statistical inference. They form the basis for rigorous methodologies and may be cited to rationalize statistical choices made in research (Adler et al., 2009).

Interdisciplinarity: While both fields have interdisciplinary connections, OR & MS, with its focus on problem solving and decision making across sectors, may attract citations from diverse fields such as engineering, business, health care, and logistics. Similarly, citations within Stat. & Prob. may have a stronger presence in disciplines related to data science, biology, economics, social sciences, and physical sciences, where statistical methods play a central role in analysis (Ji et al., 2022).

Data sources: OR & MS often relies on real-world data from operational and management contexts, such as supply chain or production data. In contrast, Stat. & Prob. employs a wide range of data sources, including experimental data, surveys, and observations (Devore, 2015).

It's important to recognize that these distinctions are generalizations, and there may be variation among individual researchers and specific research topics within each subject category.

5. Conclusions

The Leimkuhler curve has become a widely used tool in informetrics and other fields for analyzing the concentration and inequality of resources and productivity. This paper proposed a new method for parameterizing the citation distribution using a mixture of Leimkuhler curves, which provides a more flexible and accurate framework for modeling citation distributions. We also considered measures of citation concentration across different fields, such as the Gini and Pietra indexes, to capture the differences in citation patterns across various research fields. Our results demonstrate the effectiveness of our proposed method and highlight the importance of considering measures of concentration and inequality in bibliometric analysis and other fields.

The distribution of citations can be characterized by a combination of Leimkuhler curves, with citation concentration measurements providing insights into how citations are concentrated among a select group of highly cited journals and papers. By employing this proposed method, a more comprehensive understanding of citation patterns can be achieved, leading to the effective identification of a core group of leading journals or papers. Empirical results demonstrate the method's efficacy in capturing differences in citation patterns across various fields, as evidenced by goodness of fit and predictive accuracy analyses. These findings hold significant implications for comprehending the dynamics of scientific knowledge productivity and dissemination, thereby making a valuable contribution to the field of bibliometrics.

As a limitation, we used two field examples to test the method. However, we consider that the main contribution of the paper lies in the theoretical and mathematical results. Specifically, in relation to the introduction of mixture distributions (called PG and PIG) to bibliometrics, and the obtaining of closed-form expressions for Leimkuhler curves.

We conducted an analysis of citations within the subject categories Operations Research & Management Science (OR & MS) and Statistics & Probability (Stat. & Prob.). Our calculations shed light on these categories in the following ways. We found higher citation concentration indices, as measured by the Gini and Pietra indices, in the case of Stat. & Prob. Factors that may explain these differences include citation frequency and volume, citation purpose, interdisciplinarity, and data sources.

The results obtained from the use of Leimkuhler curves have significant applications in the field of library and information science. These curves provide valuable insight into trends and patterns in the development of collections and research areas, and allow for the visualization of the distribution of information sources over time (Gupta et al., 2022). When combined with bibliometric techniques, these curves can help identify influential authors or publications within a particular subject area, facilitate the building of comprehensive collections, and identify potential research collaborators or partners.

Leimkuhler curves serve as valuable tools for examining the evolution of a particular field or subject area over time. By comparing curves from different time periods, researchers can identify trends and changes in the distribution of information sources, providing crucial insights into the dynamic nature of research fields (Borgohain et al., 2021). Moreover, these curves are useful for evaluating the adequacy and balance of a library's collection over time. By analyzing the curves, librarians can pinpoint areas where adjustments in source acquisition may be necessary to ensure that the collection remains current and relevant. Furthermore, these curves hold relevance in the context of digital libraries, as they enable the study of usage patterns of electronic resources, which can inform collection development decisions and resource allocation strategies.

Finally, the use of gamma and inverse Gaussian distributions provides a lot of flexibility to the new models obtained, but beyond the underlying mathematical convenience, there is not theoretical justification for their use. This constitutes a weakness or limitation that must be underlined.

CRediT authorship contribution statement

Emilio Gómez-Déniz: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Pablo Dorta-González:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2011). Research productivity: Are higher academic ranks more productive than lower ones? *Scientometrics*, *88*, 915–928. https://doi.org/10.1007/s11192-011-0426-6.

Abramowitz, M., & Stegun, I. A. (Eds.). (1968). *In applied mathematics series: Vol. 55*. *Handbook of mathematical functions*. National Bureau of Standards.

Adler, R., Ewing, J., & Taylor, P. (2009). Citation statistics: A report from the international mathematical union (IMU) in cooperation with the international council of industrial and applied mathematics (ICIAM) and the institute of mathematical statistics (IMS). *Statistical Science*, *24*(1), 1–14. https://www.jstor.org/stable/20697661.

Amemiya, T. (1985). *Advanced econometrics*. Oxford: Basic Blackwell.

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*, *2*, 244–263.

Borgohain, D. J., Verma, M. K., Nazim, M., & Sarkar, M. (2021). Application of Bradford's law of scattering and Leimkuhler model to information science literature. *COLLNET Journal of Scientometrics and Information Management*, *15*(1), 197–212. https://doi.org/10.1080/09737766.2021.1943041.

Bozdogan, H. (1987). The general theory and its analytical extension. *Psychometrika*, *52*, 345–370.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering*, *137*, 85–86; reprinted in *Journal of Information Science*, *10*(4), 173–175.

Brzezinski, M. (2015). Power laws in citation distributions: Evidence from Scopus. *Scientometrics*, *103*, 213–228. https://doi.org/10.1007/s11192-014-1524-z.

Burrell, Q. L. (1991). The Bradford distribution and the Gini index. *Scientometrics*, *21*, 181–194.

Burrell, Q. L. (1992). The Gini index and the Leimkuhler curve for bibliometric processes. *Information Processing & Management*, *28*, 19–33.

Burrell, Q. L. (2005). Symmetry and other transformation features of Lorez/Leimkuhler representations of informetric data. *Information Processing & Management*, *41*, 1317–1329.

Burrell, Q. L., & Cane, V. R. (1982). The analysis of library data. *Journal of the Royal Statistical Society. Series A. General*, *145*(4), 439–471.

Dasgupta, P., Sen, A., & Starrett, D. (1973). Notes on the measurement of inequality. *Journal of Economic Theory*, *6*, 180–187.

Devore, J. L. (2015). *Probability and statistics for engineering and the sciences*. Boston: Cengage Learning.

Dorta-González, P., & Gómez-Déniz, E. (2022). Modeling the obsolescence of research literature in disciplinary journals through the age of their cited references. *Scientometrics*, *127*, 2901–2931.

Egghe, L. (Ed.). (2005). *Power laws in the information production process: Lotkaian informetrics*. Emerald Group Publishing Limited.

Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, *69*(1), 131–152.

Eliazar, I. I., & Sokolov, I. M. (2010). Measuring statistical heterogeneity: The Pietra index. *Physica A*, *389*, 117–125.

Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999). On power-law relationships of the Internet topology. In *Applications, technologies, architectures, and protocols for computer communication: Proceedings of the conference on applications, technologies, architectures, and protocols for computer communication* (pp. 251–262). New York: ACM.

Folks, J. L., & Chhikara, R. S. (1978). The inverse Gaussian distribution and its statistical application-a review. *Journal of the Royal Statistical Society, Series B, Methodological*, *40*(3), 263–289.

Gordy, M. B. (1998). Computationally convenient distributional assumptions for common-value auctions. *Computational Economics*, *12*, 61–78.

Gupta, S., Kanaujia, A., Lathabai, H. H., & Singh, V. K. (2022). Patterns in the growth and thematic evolution of Artificial Intelligence research: A study using Bradford distribution of productivity and path analysis. In *Research square preprint*.

Hubert, J. J. (1977). Bibliometric models for journal productivity. *Social Indicators Research*, *4*, 441–473.

Ji, P., Jin, J., Ke, Z. T., & Li, W. (2022). Co-citation and co-authorship networks of statisticians. *Journal of Business & Economic Statistics*, *40*(2), 469–485. https://doi.org/10.1080/07350015.2021.1978469.

Kakwani, N. (1980). On a class of poverty measures. *Econometrica*, *48*, 437–446.

Leimkuhler, F. F. (1967). The Bradford distribution. *Journal of Documentation*, *23*, 197–207.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16*(12), 317–323.

Nair, N. U., & Vineshkumar, B. (2022). Modelling informetric data using quantile functions. *Journal of Informetrics*, *16*(2), Article 101266. https://doi.org/10.1016/j.joi.2022.101266.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351.

Pareto, V. (1895). La legge della domanda. *Giornale degli Economisti 2nd Series*, *10*, 59–68.

Rousseau, R., Egghe, L., & Guns, R. (2018). *Becoming metric-wise. A bibliometric guide for researchers*. Chandos-Elsevier.

Salpeter, E. (1955). The luminosity function and stellar evolution. *The Astrophysical Journal*, *121*, 161–167.

Sarabia, J. (2008a). Explicit expressions for the Leimkuhler curve in parametric families. *Information Processing & Management*, *44*, 1808–1818.

Sarabia, J., Castillo, E., & Slottje, D. J. (1999). An ordered family of Lorenz curves. *Journal of Econometrics*, *91*, 43–60.

Sarabia, J. (2008b). A general definition of the Leimkuhler curve. *Journal of Informetrics*, *2*, 156–163.

Sarabia, J. M., Gómez-Déniz, E., Sarabia, M., & Prieto, F. (2010). A general method for generating parametric Lorenz and Leimkuhler curves. *Journal of Informetrics*, *4*(4), 524–539.

Saravia, L. A., Doyle, S. R., & Bond-Lamberty, B. (2018). Power laws and critical fragmentation in global forests. *Scientific Reports*, *8*, Article 17766. https://doi.org/ 10.1038/s41598-018-36120-w.

Seshadri, V. (1983). The inverse Gaussian distribution: Some properties and characterizations. *Canadian Journal of Statistics*, *11*(2), 131–136.

Shorrocks, A. F. (1983). Ranking income distributions. *Economica*, *50*, 2–17.

Sordo, M. A., Navarro, J., & Sarabia, J. M. (2014). Distorted Lorenz curves: Models and comparisons. *Social Choice and Welfare*, *42*(4), 761–780.

Thelwall, M. (2016a). The discretised lognormal and hooked power law distributions for complete citation data: Best options for modelling and regression. *Journal of Informetrics*, *10*(2), 336–346. https://doi.org/10.1016/j.joi.2015.12.007.

Thelwall, M. (2016b). Are the discretised lognormal and hooked power law distributions plausible for citation data? *Journal of Informetrics*, *10*(2), 454–470. https:// doi.org/10.1016/j.joi.2016.03.001.

Thelwall, M., & Wilson, P. (2014). Distributions for cited articles from individual subjects and years. *Journal of Informetrics*, *8*(4), 824–839. https://doi.org/10.1016/ j.joi.2014.08.001.

Yitzhaki, S. (1983). On an extension of the Gini inequality index. *International Economic Review*, *24*, 617–628.

Zipf, G. K. (1941). *National unity and disunity; the nation as a bio-social organism*. Bloomington: Principia Press.