Journal of Informetrics 8 (2014) 362–368

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Empirical modeling of the impact factor distribution

Michal Brzezinski*

Faculty of Economic Sciences, University of Warsaw, Poland

ARTICLE INFO

Article history:
Received 25 December 2013
Received in revised form 17 January 2014
Accepted 20 January 2014
Available online 11 February 2014

Keywords:
Impact factor
Two-exponent law
Dagum model
Singh-Maddala model
Maximum likelihood estimation
Model selection

ABSTRACT

The distribution of impact factors has been modeled in the recent informetric literature using two-exponent law proposed by Mansilla, Köppen, Cocho, and Miramontes (2007). This paper shows that two distributions widely-used in economics, namely the Dagum and Singh-Maddala models, possess several advantages over the two-exponent model. Compared to the latter, the former models give as good as or slightly better fit to data on impact factors in eight important scientific fields. In contrast to the two-exponent model, both proposed distributions have closed-from probability density functions and cumulative distribution functions, which facilitates fitting these distributions to data and deriving their statistical properties.

1. Introduction

The distribution of journal impact factors has been recently studied in the informetric literature from both theoretical and empirical perspectives. Mansilla et al. (2007) proposed the two-exponent law to model rank-frequency distributions of impact factors. This model was used by Campanario (2010) to study empirically changes in the distribution of impact factors over time. A theoretical derivation of the rank-frequency distribution of impact factors was derived by Egghe (2009); see also Egghe (2011) and Egghe and Waltman (2011).[1] Mishra (2010) has fitted several well-established statistical distributions to data on impact factors for journals from several scientific disciplines. The two-exponent law introduced by Mansilla et al. (2007) has been recently studied by Sarabia, Prieto, and Trueba (2012). The authors have obtained the probabilistic quantile function corresponding to the two-exponent law as well as derived several statistical measures and tools associated with this law like the moments, Lorenz and Leimkuhler curves and the Gini index of inequality. Moreover, they fitted the two-exponent law to data on impact factors for eight science categories and found that the fit of the model was satisfactory.

The present paper contributes to the literature on empirical modeling of the journal impact factor distribution by verifying in a statistically rigorous way if the two-exponent law is consistent with data on impact factors. In particular, the paper fits the two-exponent law to data using the maximum likelihood approach, which is more efficient than the least squares approach used in previous studies (Mansilla et al., 2007; Sarabia et al., 2012). The fit of the model to data is evaluated using an appropriate goodness-of-fit test. Finally, the two-exponent law is compared to alternative models using a likelihood ratio test. As alternatives we have chosen the Singh-Maddala and Dagum statistical models (see, e.g., Kleiber & Kotz, 2003), which

* Tel.: +48 22 5549155; fax: +48 22 8312846.
E-mail address: mbrzezinski@wne.uw.edu.pl

[1] See also Waltman and Van Eck (2009) for a criticism of Egghe (2009).

1751-1577/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2014.01.009

are widely-used in economics to model the distribution of income and other variables.[2] The analysis is performed for impact factors from eight science categories studied previously by Sarabia et al. (2012).

The reminder of the paper is structured as follows. Section 2 introduces definition and basic properties of the two-exponent law and the compared alternative models. Statistical tests used to assess goodness of fit to data and tests used for model selection are presented as well. Section 3 describes our data on impact factors, while Section 4 presents and discusses empirical findings. Finally, Section 5 concludes.

2. Methods

2.1. The two-exponent law versus Dagum and Singh-Maddala models

In order to model the distribution of impact factors, Mansilla et al. (2007) introduced the two-exponent law in terms of rank-frequency distribution taking the form:

$$f(r) = K\frac{(N+1-r)^b}{r^a}, \tag{1}$$

where $a > 0$, $b > 0$ are shape parameters, $K > 0$ is a scale parameter, N is the total number of sources (journals in the case of modeling impact factors), $r = 1, 2, \ldots, N$ is the ranking number and $f(r)$ it the impact factor. If $b = 0$, then (1) reduces to Zipf's law (Egghe, 2005); if $a = b$ it becomes the Lavalette law (Lavalette,, 1996), and when $a = 0$ it becomes a power law. Sarabia et al. (2012) derive the quantile function that corresponds to the two-exponent law (1), which takes the form:

$$F^{-1}(u) = K\frac{u^b}{(1-u)^a}, \quad \text{for} \quad 0 < u < 1. \tag{2}$$

The distribution defined by (2) was introduced in statistical literature by Hankin and Lee (2006), who called it the Davies distribution. Unfortunately, neither probability density function (pdf) nor cumulative distribution function (cdf) for (2) is available in closed form except for some special cases like the Zipf's law, the power law or the Lavalette law. However, they can be calculated numerically by inverting the quantile function (2).

The Singh-Maddala and Dagum models were introduced in economics in the context of modeling income distribution by, respectively, Singh and Maddala (1976) and Dagum (1977). In statistics, these distributions appeared first in the system of distributions of Burr (1942) and are known as Burr XII distribution (Singh-Maddala) and Burr III distribution (Dagum). For the Singh-Maddala and Dagum distributions, the pdfs, cdfs and quantile functions are available in closed forms. Specifically, for a sample of positive impact factors in a given scientific field, $x_1, \ldots, x_N$, the cdf for the Singh-Maddala distribution is given by:

$$F(x) = 1 - \left[1 + \left(\frac{x}{b}\right)^a\right]^{-q}, \tag{3}$$

where $a > 0$, $q > 0$ are shape parameters and $b > 0$ is a scale parameter. The cdf for the Dagum distribution is

$$F(x) = \left[1 + \left(\frac{x}{b}\right)^{-a}\right]^{-p}, \tag{4}$$

where $a > 0$, $p > 0$ are shape parameters and $b > 0$ is a scale parameter. The Singh-Maddala and Dagum distributions are closely related in the following way: $X \sim \text{Dagum}(a, b, p) \leftrightarrow (1/X)\text{SM}(a, (1/b), p))$, where $\sim$ means "is distributed as". The upper tail of the Singh-Maddala distribution is governed by two parameters (a and q), while the lower tail by a only (Kleiber, 1996). The opposite holds for the Dagum model (two parameters govern the behaviour of the lower tail and only one shapes the upper tail) and for this reason this model is more flexible in the lower tail. Therefore, the two models can be considered complementary as they have advantages in modeling different parts of the data. Theoretical properties of the Singh-Maddala and Dagum distributions are very well known; see Kleiber (1996, 2008) and Kleiber and Kotz (2003) for a detailed discussion.[3] In particular, while Sarabia et al. (2012) offer expressions for the Lorenz curve and the Gini index of inequality for the two-exponent distribution, in case of the Singh-Maddala and Dagum models expressions for a wide variety of inequality measures exist (Jenkins, 2009; Kleiber & Kotz, 2003). Moreover, the conditions that allow for testing Lorenz dominance (i.e. inference on inequality robust to the choice of an inequality measure) are available (Kleiber, 1996; Wilfling & Krämer, 1993).

When $a = b$ the two-exponent law becomes the Lavalette law (Lavalette,, 1996). As noticed by Sarabia et al. (2012), the Lavalette law is known in the economic literature as the Fisk (or log-logistic) distribution. The Fisk distribution is a special

[2] These distributions, among others, were previously used by Mishra (2010) to study the distribution of impact factors. However, they were not compared with the two-exponent model.

[3] The Singh-Maddala and Dagum distributions are nested within a four-parameter Generalized Beta of the Second Kind (GB2) model introduced by McDonald (1984). We have experimented with fitting this model to data on impact factors, but the gains from the additional complexity were small.