Journal of Informetrics 10 (2016) 719–731

Contents lists available at ScienceDirect

Journal of Informetrics

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

Measuring the robustness of the journal *h*-index with respect to publication and citation values: A Bayesian sensitivity analysis

C. Malesios*

Department of Rural Development, Democritus University of Thrace, 193 Pantazidou Str., Orestiada, Greece

ARTICLE INFO

Article history:
Received 21 December 2015
Received in revised form 29 February 2016
Accepted 14 March 2016

Keywords:
Journal *h*-index
Bayesian inference
Sensitivity analysis
Robustness
Citations
Publications

ABSTRACT

Braun, Glänzel, and Schubert (2006) recommended using the *h*-index as an alternative to the journal impact factor (IF) to qualify journals. In this paper, a Bayesian-based sensitivity analysis is performed with the aid of mathematical models to examine the behavior of the journal *h*-index to changes in the publication/citation counts of journals. Sensitivity of the *h*-index was most apparent for changes in the number of citations, revealing similar patterns of behavior for almost all models and independently to the field of research. In general, the *h*-index was found to be robust to changes in citations up to approximately the 25th percentile of the citation distribution, inflating its value afterwards.

1. Introduction

Hirsch (2005) introduced the *h*-index for the assessment of the research performance of scientists. Not only the indicator has found a wide use in a very short time, but also a series of articles were subsequently published either proposing modifications of the original *h*-index for its improvement, or new implementations of the proposed index. Increasingly, the *h*-index is proposed as an alternative to the most commonly used IF for evaluating the scientific impact of journals (see, e.g. Braun et al. 2006; Bornmann, Marx, & Gasparyan, 2012; Malesios & Arabatzis, 2012; Schubert, 2015). Despite the fact that various mathematical models for the *h*-index have been proposed, yet little is known about the mechanisms governing the relationship between the *h*-index and publications (P)/citations (C) and its robustness to the latter indicators by utilizing the aforementioned models. The general perspective is that the (journal) *h*-index is robust to changes in number of publications and citations. Franceschini, Maisano, and Mastrogiacomo (2013) for instance deduce that *h*-indices are robust to small variations in the publication/citation data and even to significant changes in the C values of the papers of interest, by investigating the robustness of the *h*-index to missing or wrong citation records. Courtault and Hayek (2008) have theoretically shown that a significant number of papers significantly cited must be published to increase the *h*-index. In the same lines, Rousseau (2007) found, by utilizing theoretical models, that a relative small number of highly cited publications have a small influence on the *h*-index. According to Minasny, Hartemink, McBratney, and Jang (2013), the *h*-index is less sensitive to the increase

* Tel.: +30 2552041133.
E-mail address: malesios@agro.duth.gr

http://dx.doi.org/10.1016/j.joi.2016.03.002
1751-1577/© 2016 Elsevier Ltd. All rights reserved.

in the number of citations and it does not penalize a journal for publishing a larger number of papers. For a more applied examination concerning the robustness of the *h*-index we refer the interested reader to Vanclay (2007).

However, the latter claims have not been examined thoroughly up to now, especially in the context of the *h*-index of a research journal. One may ask: what are small variations in *P*, *C* and how they can be quantified? This paper tries to fill this gap and answer the following question; how the *h*-index varies according to specific changes in the number of *P* and *C*? This research question cannot be addressed without specifying a mathematical relation between *h*, *P* and *C*. Hence, by relying on some of the well-established mathematical functions relating *h*-index with *P*, *C*, an empirical contribution to the issue of quantifying the sensitivity of the *h*-index by adopting a statistical modeling view is attempted, within the Bayesian paradigm. Bayesian methods permit model flexibility and appropriateness and the present study shall attempt to highlight the practical benefits of the Bayesian view of statistics.

In this context, it shall be also attempted to answer which model is more robust when compared to the others. The proposed methodology is illustrated utilizing two different datasets consisting of the *h*-indices, *P* and *C* of the journals in the fields of ecology and forestry included in the Web of Science (WoS) (Collection date: March, 2013 and November 2011 for ecology and forestry journals respectively). The total samples consisted of 264,519 and 71,683 research publications from 134 ecology and 54 forestry, scientific journals, respectively, thus constituting two diverse groups of data suitable for credible inferences. For more details on the collected data see Malesios (2015).

2. Methods

2.1. Introduction to Bayesian model-based inference

Statistics uses two major paradigms, classical (or conventional or frequentist) and Bayesian. Bayesian methods can incorporate scientific hypothesis in the analysis through the prior distribution and also have the advantage of being applied to problems with too complex structures that cannot be solved through classical statistics (Bernardo, 2003). Many statistical models are currently too complex to be fitted using classical statistical methods, but they can be fitted using Bayesian computational methods.

Inference for classical statistical modeling traditionally is based on the Maximum Likelihood (ML), where parameter estimates and corresponding confidence intervals are valid only for large samples. In contrast, Bayesian inference is exact for any sample, regardless of its size. Another distinctive characteristic of the Bayesian paradigm is that the data are treated as a fixed quantity and the parameters as random variables. Hence, in this sense, every parameter is assigned distributions, in contrast to classical statistics, where parameters are treated as fixed unknown constants.

Although Bayesian inference has been criticized for the use of the prior distribution, alternatively to utilizing an informative prior distribution it is also possible to specify ignorance in Bayesian analysis (i.e. we do not know anything about the parameters of interest) by assigning an uninformative (or vague or diffuse) prior. By the term uninformative prior, we mean assigning to the parameter a prior distribution with a very large variance.

In a general setting, under Bayesian inference we denote by $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)^t$ the vector of a set of, say *k*, unobserved parameters, and by $\mathbf{x}$ the observed data. Bayesian inference is based on Bayes' theorem, according to which the posterior distribution, denoted by $p(\boldsymbol{\theta}|\mathbf{x})$ is given by:

$$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\mathbf{x})} = \frac{p(\mathbf{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{\int_{\theta} p(\mathbf{x}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}) d\boldsymbol{\theta}}. \tag{1}$$

Eq. (1) states that the probability of parameters $\boldsymbol{\theta}$ given the data $\mathbf{x}$ is proportional to the likelihood function $L(\boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})$ and the prior distribution of $\boldsymbol{\theta}$, $p(\boldsymbol{\theta})$, i.e.:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

The latter constitutes the intuitive basis of model-based Bayesian inference combining the information that we know before (through prior distribution), updated using the likelihood function (the data) in order to obtain the posterior distribution which gives information about the parameter of interest.

The most challenging issue in Bayesian inference is – as is well known – the normalizing term $p(\mathbf{x})$ (often called the marginal likelihood) in the denominator of Eq. (1), due to that in most modeling cases $p(\mathbf{x})$ includes complex high-dimensional integrals which are analytically intractable. Due to this issue, the problem of generating samples from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$ is not straightforward. Only after the mid-1980s the implementation of simulation-based computing algorithms like Markov chain Monte Carlo (McMC) (Gelman, Meng, Stern, & Rubin, 2003) on widely accessible powerful computers helped to overcome these problems and led to an explosion of interest in Bayesian modeling (Ntzoufras, 2011).

Markov chain simulation yields a sample from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{x})$ of a parameter. One of the most widely used McMC techniques is Gibbs sampler (Gelfand, Hills, Racine-Poon, & Smith, 1990; Geman & Geman, 1984). A brief description of the Gibbs sampler iterative scheme for obtaining posterior samples for parameters $\boldsymbol{\theta}$ is presented below: