No. 2007–86

ASYMPTOTICS FOR THE HIRSCH INDEX

By Jan Beirlant, John H.J. Einmahl

October 2007

ISSN 0924-7815

Asymptotics for the Hirsch index [*]

Jan Beirlant John H.J. Einmahl

October 25, 2007

Abstract

The last decade methods for quantifying the research output of individual researchers have become quite popular in academic policy making. The h-index (Hirsch, 2005) constitutes an interesting quality measure that has attracted a lot of attention recently. It is now a standard measure available for instance on the Web of Science. In this paper we establish the asymptotic normality of the empirical h-index. The rate of convergence is non-standard: $\sqrt{h}/(1 + nf(h))$, where f is the density of the citation distribution and n the number of publications of a researcher. In case that the citations follow a Pareto-type or a Weibull-type distribution as defined in extreme value theory, our general result nicely specializes to results that are useful for constructing confidence intervals for the h-index.

JEL codes: C13, C14.

Key words: Asymptotic normality, confidence interval, extreme value theory, research output, scientometrics, tail empirical process.

1 Introduction

Since its introduction in Hirsch (2005), the h-index has been used as a measure to quantify the research output of individual scientists based on the distribution

[*] Jan Beirlant is Professor, Department of Mathematics and Leuven Statistics Research Centre, Katholieke Universiteit Leuven, Celestijnenlaan 200B, B-3001 Leuven, Belgium (Email: *jan.beirlant@wis.kuleuven.be*). John H.J. Einmahl is Professor, Department of Econometrics & OR and CentER, Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands (Email: *j.h.j.einmahl@uvt.nl*). The authors are very grateful to Goedele Dierckx for helping out with the application section.

of citation counts of the different papers of the individual. It is now a standard measure available in Citation Reports on the Web of Science. It can also be applied to higher levels of aggregation such as the different papers that appeared in a given journal over a particular period of time (Braun et al., 2005). If a researcher has n publications and each of the m most cited publications has at least m citations, then the maximal m is the empirical h-index of this researcher. In recent papers (Glänzel, 2006 and Egghe and Rousseau, 2006) the first theoretical properties of the h-index were derived. The underlying citation distribution was then assumed to be of Pareto-type. Specifically the dependence of the h-index on the basic parameters of the distribution and on the sample size was discussed. However the distribution of the empirical h-index as an estimator of a statistical functional was not discussed yet.

In order to be more precise, let us introduce some notation. Let $X_1, \ldots, X_n$ be i.i.d. random variables with common distribution function F. They denote the numbers of citations of the n articles of a certain researcher (or from a particular journal) with citation distribution F. Although the citation counts follow a discrete distribution, we will henceforth assume that F is continuous. That is of course mathematically convenient but, as shown below, it also allows to discover the main characteristics of the asymptotics. Let x^* denote the right endpoint of F: $x^* = \sup\{x \in \mathbb{R} : F(x) < 1\}$. We will assume $x^* = \infty$, since the case $x^* < \infty$ is not of interest in scientometrics, see the monograph Egghe (2005). We will also assume $F(0) = 0$, since this is natural in this context, but it is actually not needed.

From the definition above it follows that the theoretical h-index $h = h_n$ of a researcher or a journal is defined by

$$1 - F(h) = \frac{h}{n} ;$$

see also Glänzel (2006). Observe that h is unique and that $h = h_n \to \infty$ and $h/n \to 0$, as $n \to \infty$. Denoting the (right-continuous) empirical distribution function of the $X_i, i = 1, \ldots, n$, by $\hat{F}$, the empirical Hirsch index $\hat{H}$ is defined by

$$1 - \hat{F}(\hat{H}) \leq \frac{\hat{H}}{n} \text{ and } 1 - \lim_{x \uparrow \hat{H}} \hat{F}(x) \geq \frac{\hat{H}}{n} . \tag{1}$$

These two inequalities indeed have a unique solution. In case $\hat{F}$ puts only mass at non-negative integer values this definition coincides with the aforementioned definition in Hirsch (2005). $\hat{H}$ is a simple and comprehensible measure that works at any given level of aggregation. It combines citation impact with publication activity. Because of the Glivenko-Cantelli theorem we have that $\hat{H} \to \infty$ a.s. $(n \to \infty)$.

It is the aim of this paper to study the precise asymptotic behavior of $\hat{H}$ in a completely non-parametric setting. We will establish the expected asymptotic normality, but the rate of convergence is rather non-standard. In Section 2 the general asymptotic result will be stated and derived. In Section 3 we specify the general result to the case of Pareto-type and Weibull-type distributions. This allows to construct confidence intervals for h on the basis of $\hat{H}$. In Section 4 we apply these results to the publication-citation record of two well-known scientists.

2 Main results

In this section we present our general, main results. The next proposition leads to the proper consistency result and is as well the main step to asymptotic normality of $\hat{H}$.

Proposition 1 If F is continuous and $x^* = \infty$, then

$$\frac{\hat{H} - h}{\sqrt{h}} + \frac{n(F(\hat{H}) - F(h))}{\sqrt{h}} \xrightarrow{d} N(0,1) \qquad \text{as } n \to \infty. \tag{2}$$

Since the terms on the left in Proposition 1 have the same sign, we obtain consistency of $\hat{H}$. Obviously the result has to be stated in a ratio-setting; the usual consistency formulation is pointless since $h = h_n \to \infty$ as $n \to \infty$.

Corollary 1 (Consistency) Under the assumptions of Proposition 1, we have

$$\frac{\hat{H}}{h} \xrightarrow{P} 1.$$

We now state the main result, asymptotic normality of $\hat{H}$; note the unusual convergence rate.

Theorem 1 (Asymptotic normality) Assume F is continuous, the density $f = F'$

exists in h, and $x^* = \infty$. Assume also the following condition holds: for all $R > 0$,

$$\sup_{-R \le r \le R} \left| \frac{n\left(F\left(h + \frac{r\sqrt{h}}{nf(h)}\right) - F(h)\right)}{r\sqrt{h}} - 1 \right| \to 0 \quad \text{as } n \to \infty. \tag{3}$$

Then

$$\frac{1 + nf(h)}{\sqrt{h}}(\hat{H} - h) \xrightarrow{d} N(0,1) \quad \text{as } n \to \infty.$$

Proof of Proposition 1 Let α_n be the uniform empirical process based on the $F(X_i), 1 \le i \le n$, so

$$\sqrt{n}(\hat{F} - F) = \alpha_n \circ F \quad \text{a.s.}$$

We have, almost surely,

$$\begin{aligned} \hat{H} - h &\ge n(1 - \hat{F}(\hat{H})) - n(1 - F(h)) \\ &= n(F(h) - \hat{F}(\hat{H})) \\ &= -n(\hat{F}(\hat{H}) - F(\hat{H}) + F(\hat{H}) - F(h)) \\ &= -\sqrt{n}\alpha_n(F(\hat{H})) - n(F(\hat{H}) - F(h)). \end{aligned}$$

Hence

$$\hat{H} - h + n(F(\hat{H}) - F(h)) \ge -\sqrt{n}\alpha_n(F(\hat{H})) \quad \text{a.s.}$$

Similarly we obtain

$$\hat{H} - h + n(F(\hat{H}) - F(h)) \le -\sqrt{n}\alpha_n(F(\hat{H})) + 1 \quad \text{a.s.}$$

Hence

$$\frac{\hat{H} - h}{\sqrt{h}} + \frac{n(F(\hat{H}) - F(h))}{\sqrt{h}} = -\sqrt{\frac{n}{h}}\alpha_n(F(\hat{H})) + o(1) \quad \text{a.s.} \tag{4}$$

Let w_n be the tail empirical process near 1, based on α_n, i.e.

$$w_n(s) = \sqrt{\frac{n}{h}}\alpha_n\left(1 - \frac{h}{n}s\right), \quad 0 \le s \le \frac{n}{h}\ .$$

Write

$$T_n = \frac{n(F(\hat{H}) - F(h))}{\sqrt{h}}\ .$$

Using both these definitions and the definition of h, (4) translates to

$$\frac{\hat{H} - h}{\sqrt{h}} + T_n = -w_n\left(1 - \frac{T_n}{\sqrt{h}}\right) + o(1) \quad \text{a.s.} \tag{5}$$

This yields

$$\left|\frac{\hat{H}-h}{\sqrt{h}(1\vee(1-T_n/\sqrt{h}))^{3/4}}+\frac{T_n}{(1\vee(1-T_n/\sqrt{h}))^{3/4}}\right|\leq\sup_{0\leq s\leq n/h}\frac{|w_n(s)|}{(1\vee s)^{3/4}}+o(1)\quad\text{a.s.}\tag{6}$$

It is well-known that

$$\sup_{0\leq s\leq n/h}\frac{|w_n(s)|}{(1\vee s)^{3/4}}=O_P(1)\tag{7}$$

see, e.g., Einmahl (1997). Recall that $\hat{H}-h$ and $n(F(\hat{H})-F(h))$ have the same sign. Hence using (7), we obtain from (6) that

$$\frac{T_n}{(1\vee(1-T_n/\sqrt{h}))^{3/4}}=O_P(1).$$

It follows that $T_n 1_{[T_n\geq -1]}=O_P(1)$. If $T_n<-1$, we have for large n

$$1\vee(1-T_n/\sqrt{h})=1-T_n/\sqrt{h}\leq 1-T_n\leq -2T_n.$$

Hence $T_n 1_{[T_n<-1]}=O_P(1)$, so $T_n=O_P(1)$. This yields

$$1-\frac{T_n}{\sqrt{h}}\xrightarrow{P}1.$$

Combining this with (5) and the weak convergence of w_n on $[0,2]$ to a standard Wiener process (see, e.g., Einmahl (1997)) yields (2). □

Proof of Theorem 1 From Proposition 1 we see that

$$\left(1+\frac{n(F(\hat{H})-F(h))}{\hat{H}-h}\right)\frac{\hat{H}-h}{\sqrt{h}}\xrightarrow{d}N(0,1).$$

So it suffices to show that

$$\frac{1+nf(h)}{1+\frac{n(F(\hat{H})-F(h))}{\hat{H}-h}}\xrightarrow{P}1,$$

which is implied by

$$\frac{F(\hat{H})-F(h)}{f(h)(\hat{H}-h)}\xrightarrow{P}1.\tag{8}$$

To prove (8), observe that it follows from Proposition 1 that $F(\hat{H})-F(h)=O_P(\sqrt{h}/n)$, i.e. for $\varepsilon>0$ we can find an $M_\varepsilon>0$ such that with probability greater than $1-\varepsilon$, for large n

$$|F(\hat{H})-F(h)|\leq\frac{M_\varepsilon\sqrt{h}}{n}\ .\tag{9}$$

Define $\hat{r} = nf(h)(\hat{H} - h)/\sqrt{h}$. Then (9) reads

$$\left| F\left(h + \frac{\hat{r}\sqrt{h}}{nf(h)}\right) - F(h)\right| \leq \frac{M_\varepsilon \sqrt{h}}{n} \ . \tag{10}$$

Assume $|\hat{r}| > 2M_\varepsilon$. Then from the condition, with $R = 2M_\varepsilon$, it follows that

$$\left| F\left(h + \frac{\hat{r}\sqrt{h}}{nf(h)}\right) - F(h)\right| \geq \frac{3}{2}\frac{M_\varepsilon \sqrt{h}}{n} \ .$$

This contradicts (10). Hence with probability greater than $1 - \varepsilon$, for large n, $|\hat{r}| \leq 2M_\varepsilon$. Now apply the condition again with $R = 2M_\varepsilon$, then

$$1 - \varepsilon \leq \frac{n(F(\hat{H}) - F(h))}{\hat{r}\sqrt{h}} \leq 1 + \varepsilon.$$

This is (8). □

3 Pareto-type and Weibull-type tails

From our main theorem two important corollaries follow, which immediately yield a confidence interval for h. These corollaries also show that the condition in Theorem 1 is appropriate and can be validated under popular, semiparametric models encountered in extreme value theory and scientometrics.

First assume that the distribution function F satisfies the von Mises condition for heavy-tailedness, i.e. we have

$$\lim_{x\to\infty} \frac{xf(x)}{1 - F(x)} = \alpha. \tag{11}$$

with tail index $\alpha = 1/\gamma > 0$, with γ the extreme value index. It is then immediate that

$$\lim_{n\to\infty} nf(h) = \lim_{n\to\infty} \frac{hf(h)}{1 - F(h)} = \alpha.$$

Nevertheless, it is remarkable that under this often-used, semiparametric model, the complicated quantity $nf(h)$ in the asymptotic normality statement can be replaced by α. For differentiable distribution functions, assumption (11) is equivalent to stating that the tail of the distribution is Pareto-type (or equivalently Zipfian or Lotkaian):

$$1 - F(x) = x^{-\alpha}\ell(x)$$

where ℓ is a slowly varying function:

$$\frac{\ell(tx)}{\ell(t)} \to 1 \text{ as } t \to \infty \text{ for every } x > 0.$$

This assumption is of fundamental importance in the field of informetrics (Egghe, 2005). When $\ell(x) = C(1 + o(1))$ for some positive constant C, when $x \to \infty$, one easily verifies that $h = (Cn)^{1/(1+\alpha)}(1 + o(1))$. Further we need Karamata's representation theorem (see Theorem 1.3.1 in Bingham et al., 1987):

$$\ell(x) = c(x) \exp\left(\int_1^x \delta(u)du/u\right) \tag{12}$$

with $c(x) \to c_0 > 0$ and $\delta(x) \to 0$, as $x \to \infty$.

Let $\hat{\gamma}$ be one of the many known consistent estimators of the extreme value index $\gamma > 0$ and set $\hat{\alpha} = 1/\hat{\gamma}$; see for instance Chapters 4 and 5 in Beirlant et al. (2004).

Corollary 2 When (11) and (12) hold with $c \equiv c_0$, we have

$$\frac{1+\hat{\alpha}}{\sqrt{\hat{H}}}(\hat{H} - h) \xrightarrow{d} N(0,1).$$

Proof of Corollary 2 We only need to check condition (3) of Theorem 1. With (12) it follows that

$$\begin{aligned} & F\left(h + \frac{r\sqrt{h}}{nf(h)}\right) - F(h) \\ = \; & h^{-\alpha}\ell(h)\left(1 - \left(1 + r/(\sqrt{h}nf(h))\right)^{-\alpha} \frac{\ell(h + r\sqrt{h}/(nf(h)))}{\ell(h)}\right). \end{aligned}$$

When the function c is constant we have for large n

$$\left|\log\left(\frac{\ell(h + r\sqrt{h}/(nf(h)))}{\ell(h)}\right)\right| = \left|\int_h^{h+r\sqrt{h}/(nf(h))} \delta(u)du/u\right| \leq \frac{2|r|}{\sqrt{h}nf(h)} \sup_{u \geq h/2} |\delta(u)|.$$

Hence, since $nh^{-\alpha-1}\ell(h) = 1$ by the definition of h, we obtain after a straightforward calculation

$$\begin{aligned} &\sup_{-R\le r\le R}\left|\frac{n\left(F\left(h+\frac{r\sqrt{h}}{nf(h)}\right)-F(h)\right)}{r\sqrt{h}}-1\right| \\ = \;\; & nh^{-\alpha-1}\ell(h)\left(\left|\frac{\alpha}{nf(h)}(1+o(1))-1\right|+\frac{2\sup_{u\ge h/2}|\delta(u)|}{nf(h)}(1+o(1))\right)\to 0. \quad \square \end{aligned}$$

Similarly the main theorem can be specified for the class of Weibull-type distributions, defined as

$$1 - F(x) = \exp\left(-x^{\tau}\ell(x)\right) \tag{13}$$

with Weibull parameter $\tau > 0$ and a slowly varying function ℓ. Examples are the gamma, normal and Weibull distributions. This subclass constitutes an important subclass of the Gumbel domain of max-attraction, characterized by $\gamma = 0$. Consistent estimators $\hat{\tau}$ of τ can be found for instance in Girard (2004) and Diebolt et al. (2007). Remark that under (13) with $\ell(x) = C(1 + o(1))$ for some positive constant C as $x \to \infty$, we have $h_n = (C^{-1}\log n)^{1/\tau}(1 + o(1))$ as $n \to \infty$. The following result now follows using similar calculations as for Corollary 2.

Corollary 3 When (13) holds with ℓ satisfying (12) with the function c constant, we have

$$\frac{\hat{\tau}\log n}{\sqrt{\hat{H}}}(\hat{H} - h) \xrightarrow{d} N(0, 1).$$

It is interesting to note that for $\tau = 1/2$ and $\ell(x) = C(1 + o(1))$ for some positive constant C as $x \to \infty$, this result simplifies to $\hat{H} - h \xrightarrow{d} N(0, 4/C^2)$.

4 Application

As an application we consider here the h-index of David R. Cox and of Pál Erdős, two really outstanding scientists who do not need any further introduction.

In Figure 1 we see the functions $1 - \hat{F}(\cdot)$ and $\cdot/n$, cf. (1), for the $n = 450$ papers of Cox. We find an empirical Hirsch index $\hat{H}_{\text{Cox}} = 57$. Figure 2 shows the Pareto QQ-plot, plotting on a log-log scale the ordered data against their theoretical values under the standard Pareto model, jointly with four well-known estimators

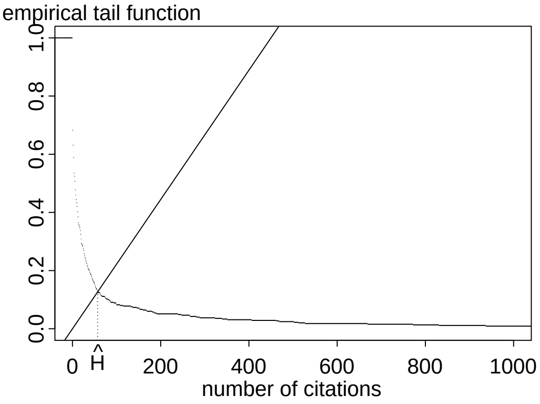

Figure 1: The empirical Hirsch index for Cox.

of the extreme value index γ (see Chapters 4 and 5 in Beirlant et al. (2004)). The QQ-plot is based on the 307 papers with non-zero citations. It shows linear

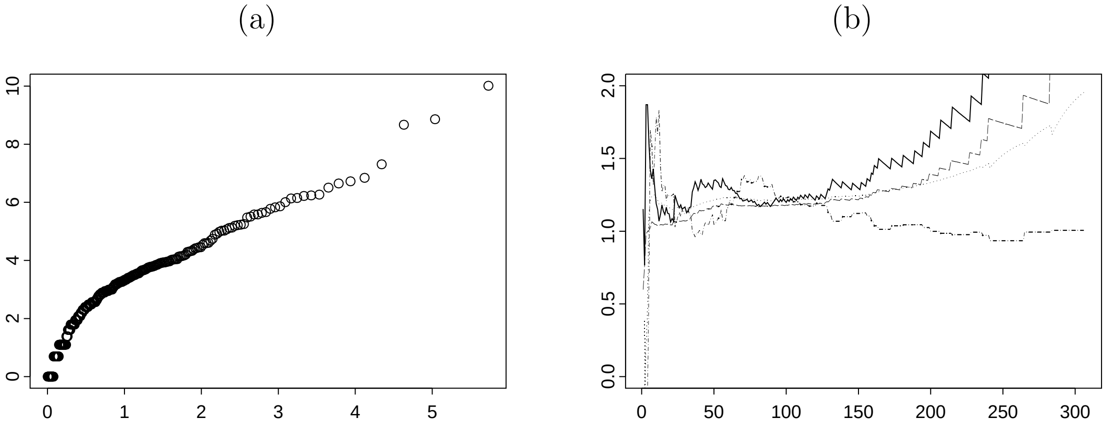

Figure 2: For the citation sizes of Cox: (a) Pareto QQ-plot (b) Hill (full line), Moment (dotted line), ML (dashed-dotted), UH (dashed line) estimator.

behavior with positive slope for the large citations numbers, indicating that the citation distribution is heavy tailed. In fact, the Hill estimator can be considered as the slope of the QQ plot when using only the largest k observations. A careful inspection of the plot of the estimators of the extreme value index leads to an estimate $\hat{\gamma} = 1.2$; certainly for k running from 25 to 125 the four estimators show stable behavior. Using $\hat{\alpha} = 1/\hat{\gamma}$ in Corollary 2, we obtain $(50.2, 63.8)$ an asymptotic

90%-confidence interval for h_{Cox} .

Based on the $n = 519$ papers of Erdős, we find an empirical Hirsch index $\hat{H}_{\text{Erdős}} = 22$. The Pareto QQ-plot (based on the 393 papers with non-zero citations) and the four estimators of γ are shown in Figure 3. We see again linear behavior in the QQ-plot for the large citation numbers. The plot of the estimators of the extreme value index leads to an estimate $\hat{\gamma} = 0.6$. This yields $(19.1, 24.9)$ as an asymptotic 90%-confidence interval for $h_{\text{Erdős}}$.

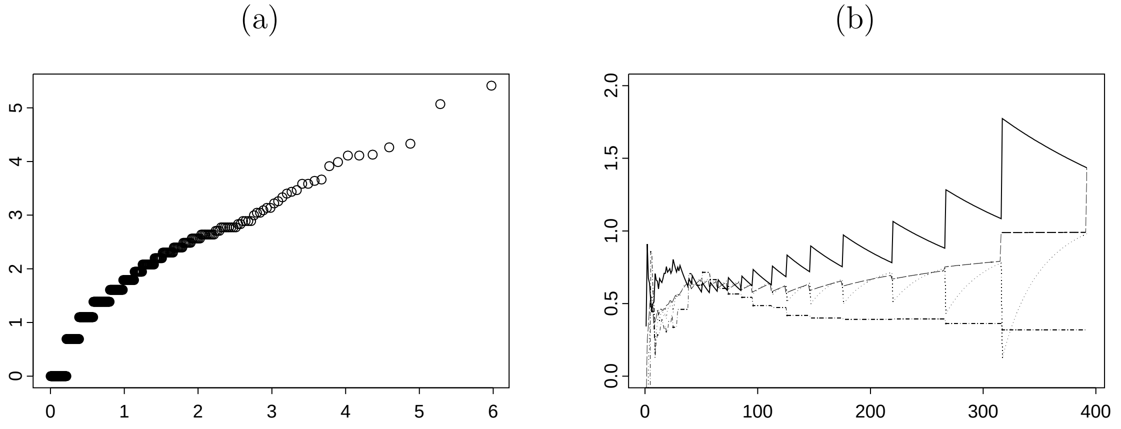

Figure 3: For the citation sizes of Erdős: (a) Pareto QQ-plot (b) Hill (full line), Moment (dotted line), ML (dashed-dotted), UH (dashed line) estimator.

References

[1] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). *Statistics of Extremes, Theory and Applications.* Wiley, New York.

[2] Bingham, N.H., Goldie, C.M. and Teugels, J.L. (1987). *Regular Variation.* Encyclopedia of Mathematics and its Applications, 27. Cambridge University Press.

[3] Braun, T., Glänzel, W. and Schubert, A. (2005). A Hirsch-type index for journals. *The Scientist*, **19**, 8.

[4] Diebolt, J., Gardes, L., Girard S. and Guillou A. (2007). Bias-reduced estimators of the Weibull tail-coefficient, *Test*, to appear.

[5] Egghe, L. (2005). *Power Laws in the Information Production Process: Lotkaian Informetrics.* Wiley, New York.

[6] Egghe, L. and Rousseau, R. (2006). An informetric model for the h-index. *Scientometrics*, **69**, 121-129.

[7] Einmahl, J.H.J. (1997). Poisson and Gaussian approximation of weighted local empirical processes. *Stochastic Process. Appl.*, **70**, 31-58.

[8] Girard, S. (2004). A Hill type estimate of the Weibull tail-coefficient. *Comm. in Statist. Theory and Methods*, **33**, 205-234.

[9] Glänzel, W. (2006). On the h-index - A mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, **67**, 315-321.

[10] 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.