Three Dimensions of Impact (3DI)

Marek Gagolewski

Deakin University, Melbourne–Burwood, VIC, AU
Faculty of Mathematics and Information Science, Warsaw University of Technology, PL
Systems Research Institute, Polish Academy of Sciences, PL

March 2023

Introduction

The Hirsch (2005) index of a sequence with $x_1 \geq x_2 \geq \cdots \geq x_N$:

$$h(x_1, \ldots, x_n) = \max\{H : x_H \geq H\} = \bigvee_{H=1}^{N} H \wedge \lfloor x_H \rfloor.$$

Torra and Narukawa[1] noted that this is a discrete Sugeno integral wrt the counting measure.

[1] Torra V., Narukawa Y., The h-index and the number of citations: Two fuzzy integrals, *IEEE Transactions on Fuzzy Systems* 16(3), 795–797, 2008.

Introduction

We've been playing with the h-index and its "fuzzy" generalisations/counterparts for quite a while [2] [3] [4] [5].
For instance, studying theoretical properties, construction methods, proposing new variants, introducing algorithms for fitting monotone measures to data.

[2] Beliakov G., James S., Citation-based journal ranks: The use of fuzzy measures, *Fuzzy Sets and Systems* 167, 101–119, 2011.

[3] Mesiar R., Gagolewski M., H-index and other Sugeno integrals: Some defects and their compensation, *IEEE Transactions on Fuzzy Systems* 24(6), 1668–1672, 2016, DOI:10.1109/TFUZZ.2016.2516579

[4] Gagolewski M., Mesiar R., Monotone measures and universal integrals in a uniform framework for the scientific impact assessment problem, *Information Sciences* 263, 166–174, 2014, DOI:10.1016/j.ins.2013.12.004

[5] ... and many others.

Introduction

For instance, in a recent paper[6], we propose a new generalisation of the classical Sugeno integral motivated by the Hirsch, Woeginger, and other geometrically-inspired indices of scientific impact.
The new integral adapts to the rank-size curve better as it allows for putting more emphasis on highly-valuated items and/or the tail of the distribution (level measure).
We study its fundamental properties and give the conditions guaranteeing the fulfilment of subadditivity as well as the Jensen-, Liapunov-, Hardy-, Markov-, and Paley-Zygmund-type inequalities.

[6] Boczek M., Gagolewski M., Kaluszka M., Okolewski A., A benchmark-type generalization of the Sugeno integral with applications in bibliometrics, *Fuzzy Sets and Systems*, 2023, in press, DOI:10.1016/j.fss.2023.01.014.

Introduction

For instance, in a recent paper[6], we propose a new generalisation of the classical Sugeno integral motivated by the Hirsch, Woeginger, and other geometrically-inspired indices of scientific impact.
The new integral adapts to the rank-size curve better as it allows for putting more emphasis on highly-valuated items and/or the tail of the distribution (level measure).
We study its fundamental properties and give the conditions guaranteeing the fulfilment of subadditivity as well as the Jensen-, Liapunov-, Hardy-, Markov-, and Paley-Zygmund-type inequalities.

Applications of the h-index go beyond bibliometrics.

[6] Boczek M., Gagolewski M., Kaluszka M., Okolewski A., A benchmark-type generalization of the Sugeno integral with applications in bibliometrics, *Fuzzy Sets and Systems*, 2023, in press, DOI:10.1016/j.fss.2023.01.014.

Introduction

For instance, in a recent paper[6], we propose a new generalisation of the classical Sugeno integral motivated by the Hirsch, Woeginger, and other geometrically-inspired indices of scientific impact.
The new integral adapts to the rank-size curve better as it allows for putting more emphasis on highly-valuated items and/or the tail of the distribution (level measure).
We study its fundamental properties and give the conditions guaranteeing the fulfilment of subadditivity as well as the Jensen-, Liapunov-, Hardy-, Markov-, and Paley-Zygmund-type inequalities.

Applications of the h-index go beyond bibliometrics.

. . . but I will not talk about these results today.

[6] Boczek M., Gagolewski M., Kaluszka M., Okolewski A., A benchmark-type generalization of the Sugeno integral with applications in bibliometrics, *Fuzzy Sets and Systems*, 2023, in press, DOI:10.1016/j.fss.2023.01.014.

3DI

The "aggregation community" usually studies different "integrals" without assuming any data model.
Instead, I'll present some of my recent results [7] [8] [9] [10] [11] related to complex networks, statistics, and economics.

[7] Siudem G., Żogała-Siudem B., Cena A., Gagolewski M., Three dimensions of scientific impact, *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* 117, 13896–13900, 2020, DOI:10.1073/pnas.2001064117

[8] Siudem G., Nowak P., Gagolewski M., Power laws, the Price Model, and the Pareto type-2 distribution, *Physica A: Statistical Mechanics and its Applications* 606, 128059, 2022, DOI:10.1016/j.physa.2022.128059

[9] Żogała-Siudem B., Siudem G., Cena A., Gagolewski M., Agent-based model for the bibliometric h-index – Exact solution, European Physical Journal B 89(21), 2016, DOI:10.1140/epjb/e2015-60757-1

[10] Bertoli-Barsotti L., Gagolewski M., Siudem G., Żogała-Siudem B., *Equivalence of Inequality Indices*, 2023 – in preparation

[11] Bertoli-Barsotti L., Gagolewski M., Siudem G., Żogała-Siudem B., *Gini-Stable Leimkuhler Curves*, 2023 – in preparation

3DI

The 3DI (three dimensions of /scientific/ impact) model is a variation of the classical preferential attachment rule proposed by Barabási and Albert[12].

Consider a process where in every time step, a system (e.g., a citation network) grows by one entity (e.g., a new paper).
In each iteration, we distribute m wealth units (e.g., citations) amongst the already existing entities:

- $a = (1 - \rho)m$ units at random,
- $p = \rho m$ units according to the preferential attachment rule,

with ρ representing the extent to which the rich-get-richer rule dominates over pure luck.

[12] A. L. Barabási, R. Albert, Science 286, 509 (1999)

Let $X_k(t)$ denote the wealth of k-th wealthiest entity at time step t. We assume $X_k(k-1) = 0$ for every k, i.e., the k-th entity enters the system with no wealth units. Then:

$$X_k(t) = \underbrace{X_k(t-1)}_{\text{previous value}} + \underbrace{\frac{a}{t}}_{\text{accidental income}} + \underbrace{p\,\frac{X_k(t-1)+\frac{a}{t}}{(t-1)m+a}}_{\text{preferential gain or loss}},$$

First, if we assume $\rho = 0$, then we are only left with the accidental component and our model reduces to the harmonic one:

$$X_k(t) = m \sum_{i=k}^{t} \frac{1}{i} = m \left(H_t - H_{k-1} \right).$$

Note that "purely accidental" does not mean that every entity ends up with the same amount of wealth, as still older agents have had more opportunity to become impactful ("old get richer").

For $\rho < 1$ and $\rho \neq 0$, the wealth is distributed according to a mixture[13] of the accidental and the preferential component. In such a case, the solution is:

$$X_k(t) = m\frac{1-\rho}{\rho}\left(\frac{\Gamma(t+1)}{\Gamma(t+1-\rho)}\frac{\Gamma(k-\rho)}{\Gamma(k)} - 1\right).$$

[13] Let us note that $\rho < 0$ is not only possible but also has a nice interpretation: we initially distribute more than the assumed m citations at random, but then we take away from those who are already rich (rich get less).

For $\rho < 1$ and $\rho \neq 0$, the wealth is distributed according to a mixture[13] of the accidental and the preferential component. In such a case, the solution is:

$$X_k(t) = m\frac{1-\rho}{\rho}\left(\frac{\Gamma(t+1)}{\Gamma(t+1-\rho)}\frac{\Gamma(k-\rho)}{\Gamma(k)} - 1\right).$$

Denote the normalised version of the above with:

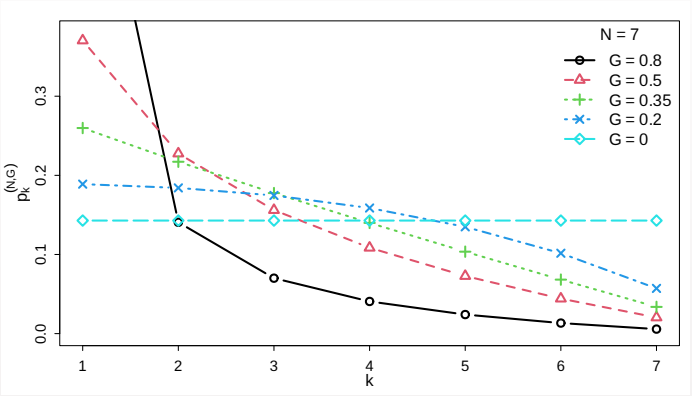
$$p_k^{(t)} = X_k(t)/mt$$

(ordered, sums to 1)

[13] Let us note that $\rho < 0$ is not only possible but also has a nice interpretation: we initially distribute more than the assumed m citations at random, but then we take away from those who are already rich (rich get less).

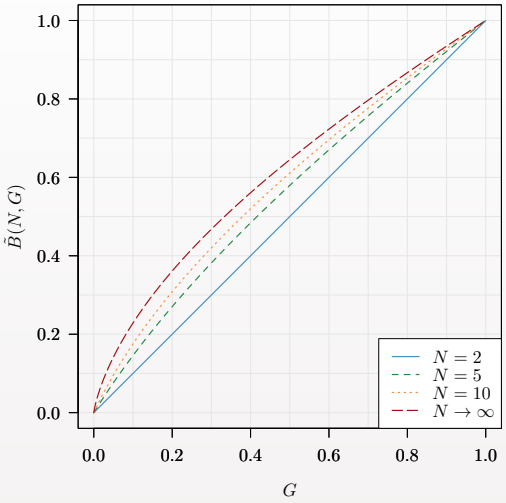
Interestingly, the Gini index of the sequence $p_1^{(N)}, \ldots, p_N^{(N)}$ only depends on ρ:

$$\rho(G) = 2 - \frac{1}{G} = \frac{2G-1}{G}, \text{ or, equivalently, } G(\rho) = \frac{1}{2-\rho}.$$



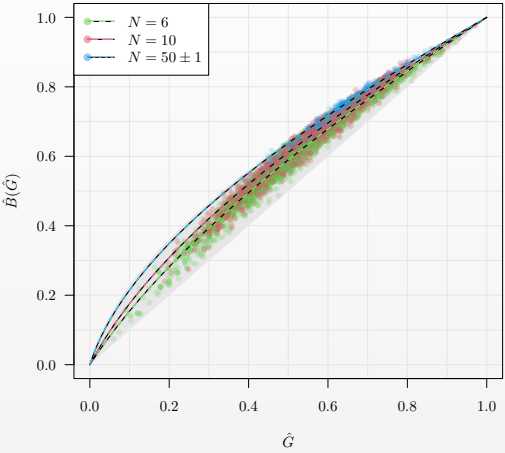
We've derived formulae for many sample statistics as functions of N, m, ρ (or G), including the h-index and many indices of economic inequality.

In particular, the Bonferroni index:
$\tilde{B}(N,G) = \frac{N}{N-1} \sum_{k=2}^{N} \frac{1}{k(k+1/G-2)}$



This model is very flexible and fits empirical data from many domains reasonably well (citations, sizes of cities, word counts, node degrees in communication networks, impact of natural disasters, . . .).

RePEc data (Research Papers in Economics; see https://citec.repec.org/), which features 66,347 authors and 1,843,967 papers:



We can also study the limiting behaviour of $p_{yN}^{(N,G)}$ as $t \to \infty$:
It holds for $G \neq \frac{1}{2}$:

$$X(y) = \lim_{N\to\infty} p_{yN}^{(N,G)} = \frac{1-\rho}{\rho}\left(y^{-\rho} - 1\right) = \frac{1-G}{2G-1}\left(y^{1/G-2} - 1\right).$$

Its inverse is:

$$S(x) = X^{-1}(x) = \left(1 + \frac{2G-1}{1-G}x\right)^{-\frac{G}{2G-1}}$$

which can be treated as a complementary cumulative distribution function (CCDF); $x > 0$.

Introducing a scale parameter $\nu > 0$, the CDF can be generalised as:

$$F(x) = 1 - \left(1 + \frac{x}{\sigma}\right)^{\alpha},$$

where $\sigma = \nu \frac{1-G}{2G-1}$.
This corresponds the CDF of the Pickands Generalized Pareto Distribution (GPD) (Pickands, 1975), but with a new parametrisation (depending on G).

It is known this distribution family unifies three different models (Hosking and Wallis, 1987; Arnold, 2015, p.11; Johnson et al, 1994, p.614).

- $G > 0.5$ – Pareto Type-II distribution,
- $G = 0.5$ – Exponential distribution,
- $G < 0.5$ – Scaled Beta distribution.

It is known this distribution family unifies three different models (Hosking and Wallis, 1987; Arnold, 2015, p.11; Johnson et al, 1994, p.614).

- $G > 0.5$ – Pareto Type-II distribution,
- $G = 0.5$ – Exponential distribution,
- $G < 0.5$ – Scaled Beta distribution.

The study of aggregation tools under the assumed (and other) models shall continue...

Fin.