

# A Guide to the agop 0.01-devel Package for R

## Aggregation Operators in R

Marek Gagolewski<sup>1,2</sup>, Anna Cena<sup>1,2</sup>

<sup>1</sup> Systems Research Institute, Polish Academy of Sciences  
ul. Newelska 6, 01-447 Warsaw, Poland

<sup>2</sup> Rexamine, Email: {gagolews,cena}@rexamine.com  
[www.rexamine.com/resources/agop/](http://www.rexamine.com/resources/agop/)

June 1, 2013

*The package, as well as this tutorial, is still in its early days – any suggestions are welcome!*

---

## Contents

<b>1</b>	<b>Getting started</b>	<b>1</b>
<b>2</b>	<b>Theoretical Background</b>	<b>2</b>
2.1	Aggregation Operators and Their Basic Properties . . . . .	2
2.2	Impact Functions and The Producers Assessment Problem . . . . .	2
<b>3</b>	<b>Predefined Classes of Aggregation Operators in agop</b>	<b>3</b>
3.1	A Note on Representing Numeric Data and Applying Functions in R . . . . .	3
3.2	. . . . .	4
3.3	Bibliometric Impact Indices . . . . .	4
<b>4</b>	<b>Visualization</b>	<b>5</b>
4.1	Depicting producers . . . . .	5
<b>5</b>	<b>Pre-orders</b>	<b>6</b>
<b>6</b>	<b>Aggregation Operators from the Probabilistic Perspective</b>	<b>9</b>
6.1	Some Notable Probability Distributions . . . . .	10
6.1.1	Pareto-Type II Distribution . . . . .	10
6.2	Stochastic Properties of Aggregation Operators . . . . .	11
<b>7</b>	<b>NEWS/CHANGELOG</b>	<b>11</b>
	<b>Bibliography</b>	<b>12</b>

## 1 Getting started

intro..... aggregation.... [19]

R [24] is a free, open sourced software environment for statistical computing and graphics, which includes an implementation of a very powerful and quite popular high-level language called S. It runs on all major operating systems, i.e. Windows, Linux, and MacOS X. To install R and/or find some information on the S language please visit R Project's Homepage at [www.R-project.org](http://www.R-project.org).

Perhaps you may also wish to install RStudio, a convenient development environment for R. It is available at [www.rstudio.org](http://www.rstudio.org).

**agop** is an Open Source (licensed under GNU LGPL 3) package for  $R \geq 2.12$  to which anyone can contribute. It started as a fork of the CITAN (Citation Analysis Toolpack) package for R.

Each session with **agop** should be preceded by a call to:

```
library("agop") # Load the package
```

To view the main page of the manual we type:

```
library(help="agop")
```

For more information please visit the package's homepage [18]. In case of any problems, comments, or suggestions feel free to contact the authors. Good luck!

## 2 Theoretical Background

Let  $\mathbb{I} = [a, b]$ , possibly with  $a = -\infty$  or  $b = \infty$  (in many practical situations we choose  $\mathbb{I} = [0, 1]$  or  $\mathbb{I} = [0, \infty]$ ).

Moreover, let  $\mathbb{I}^{1,2,\dots} = \bigcup_{n=1}^{\infty} \mathbb{I}^n$  (set of vectors with elements in  $\mathbb{I}$  – of any length).

Notational convention: For  $\mathbf{x}, \mathbf{y} \in \mathbb{I}^n$  we write  $\mathbf{x} \leq \mathbf{y}$  iff for all  $i$  it holds  $x_i \leq y_i$ .

All binary arithmetic operations on vectors  $\mathbf{x}, \mathbf{y} \in \mathbb{I}^n$  are performed element-wise, e.g.  $\mathbf{x} + \mathbf{y} = (x_1 + y_1, \dots, x_n + y_n) \in \mathbb{I}^n$ . Similarly:  $-$ ,  $\cdot$ ,  $/$ ,  $\wedge$  (min),  $\vee$  (max), etc. Additionally, each function of one variable  $f : \mathbb{I} \rightarrow \mathbb{I}$  can be extended to the vector space: we write  $f(\mathbf{x}) = (f(x_1), \dots, f(x_n))$ .

Let  $(n * c) = (c, \dots, c) \in \mathbb{I}^n$ .

Let  $x_{(i)}$  denote the  $i$ th order statistic, i.e. the  $i$ th smallest value in  $\mathbf{x}$ . Moreover, for convenience, let  $x_{\{i\}} = x_{|\mathbf{x}|-i+1}$  denote the  $i$ th greatest value in  $\mathbf{x}$ .

### 2.1 Aggregation Operators and Their Basic Properties

**Definition 1.**  $F : \mathbb{I}^{1,2,\dots} \rightarrow \mathbb{I}$  is called an (extended) aggregation operator (cf. [19]) if it is at least *nondecreasing* in each variable, i.e. for all  $n$  and  $\mathbf{x}, \mathbf{y} \in \mathbb{I}^n$  if  $\mathbf{x} \leq \mathbf{y}$ , then  $F(\mathbf{x}) \leq F(\mathbf{y})$ .

Note that each aggregation operator is a mapping into  $\mathbb{I}$ , thus for all  $n$  we have  $\inf_{\mathbf{x} \in \mathbb{I}^n} F(\mathbf{x}) \geq a$  and  $\sup_{\mathbf{x} \in \mathbb{I}^n} F(\mathbf{x}) \leq b$ . By nondecreasingness, however, these conditions reduce to  $F(n * a) \geq a$  and  $F(n * b) \leq b$ .

**Definition 2.** We call  $F$  *symmetric* if ....

**Definition 3.** We call  $F$  *idempotent* if ....

**Definition 4.** We call  $F$  *additive* if ....

**Definition 5.** We call  $F$  *minitive* if ....

**Definition 6.** We call  $F$  *maxitive* if ....

**Definition 7.** We call  $F$  *modular* if ....

### 2.2 Impact Functions and The Producers Assessment Problem

.....

Let  $\mathbb{I} = [0, \infty]$  represent the set of values that some a priori chosen paper quality measure may take. These may of course be non-integers, for example when we consider citations normalized with respect to the number of papers' authors.

It is widely accepted, see e.g. [30, 29, 31, 25, 22, 23, 8, 10, 6], that each aggregation operator  $J : \mathbb{I}^{1,2,\dots} \rightarrow \mathbb{I}$  to be applied in the impact assessment process should at least be:

- (a) nondecreasing in each variable (additional citations received by a paper or an improvement of its quality measure does not result in a decrease of the authors' overall evaluation),
- (b) arity-monotonic (by publishing a new paper we never decrease the overall valuation of the entity),
- (c) symmetric (independent of the order of elements' presentation, i.e. we may always assume that we aggregate vectors that are already sorted).

Conditions (a) and (b) imply that each impact function is able – at least potentially – to describe two “dimensions” of the author's output quality: (a) his/her ability to write eagerly-cited or highly-valuated papers and (b) his/her overall productivity.

More formally, condition (a) holds if and only if for each  $n$  and  $\mathbf{x}, \mathbf{y} \in \mathbb{I}^n$  such that  $(\forall i) x_i \leq y_i$  we have  $J(\mathbf{x}) \leq J(\mathbf{y})$ . On the other hand, axiom (b) is fulfilled iff for any  $\mathbf{x} \in \mathbb{I}^{1,2,\dots}$  and  $y \in \mathbb{I}$  it holds  $J(\mathbf{x}) \leq J(x_1, \dots, x_n, y)$ . Lastly, requirement (c) holds iff for all  $n$  and  $\mathbf{x} \in \mathbb{I}^n$  we have  $J(\mathbf{x}) = J(x_{\{1\}}, \dots, x_{\{n\}})$ , where  $x_{\{i\}}$  denotes the  $i$ th largest value from  $\mathbf{x}$ , i.e. its  $(n-i+1)$ th order statistic.

### 3 Predefined Classes of Aggregation Operators in **agop**

#### 3.1 A Note on Representing Numeric Data and Applying Functions in R

Generally, in our implementation we most often deal with numeric vectors. Recall how we create them in R:

```
(x1 <- c(5, 2, 3, 1, 0, 0))
## [1] 5 2 3 1 0 0
class(x1)
## [1] "numeric"
(x2 <- rep(10, 3))
## [1] 10 10 10
(x3 <- 10:1) # the same as seq(10, 1)
## [1] 10 9 8 7 6 5 4 3 2 1
(x4 <- seq(1, 5, length.out=6))
## [1] 1.0 1.8 2.6 3.4 4.2 5.0
(x5 <- seq(1, 5, by=1.25))
## [1] 1.00 2.25 3.50 4.75
```

Sometimes we will store the vectors of the same length in a matrix (column-/row-wise.... col/rownames....) **apply()**....

```
expertopinions <- matrix(c(
  6,7,2,3,1, # this will be the first COLUMN
  8,3,2,1,9, # 2nd
  4,2,4,1,6 # 3rd
),
ncol=3,
```

```

    dimnames=list(NULL, c("A", "B", "C")) # only column names set
)
class(expertopinions)
## [1] "matrix"
print(expertopinions) # or print(authors)
##      A B C
## [1,] 6 8 4
## [2,] 7 3 2
## [3,] 2 2 4
## [4,] 3 1 1
## [5,] 1 9 6
apply(expertopinions, 2, mean) # on each COLUMN apply the mean() function
##      A      B      C
## 3.8 4.6 3.4

```

...or in a list, especially when they are not of the same length.... `lapply()`.... `sapply()`..... possibly named elements...

```

authors <- list(
  "John S." = c(7,6,2,1,0),
  "Kate F." = c(9,8,7,6,4,1,1,0)
)
class(authors)
## [1] "list"
str(authors) # or print(authors)
## List of 2
## $ John S.: num [1:5] 7 6 2 1 0
## $ Kate F.: num [1:8] 9 8 7 6 4 1 1 0
index_h(authors[[1]]) # the h-index (see below) for 1st author
## [1] 2
sapply(authors, index_h) # calculate the h-index for all vectors in a list
## John S. Kate F.
##      2      4
index_h(authors) # index_h() expects an numeric vector on input
## Error: argument 'x' should be a numeric vector (or an object coercible to)

```

## 3.2 ...

weighted average, OWA (L-statistics), qL

WMin, OWMMin, qI

WMax, OWMMax (S-statistics), qS

OM3

## 3.3 Bibliometric Impact Indices

Below we assume that  $\mathbb{I} = [0, \infty]$ .

**The  $h$ -index.** Given a sequence  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{I}^{1,2,\dots}$ , the *Hirsch index* [20] of  $\mathbf{x}$  is defined as  $H(\mathbf{x}) = \max\{i = 1, \dots, n : x_{\{i\}} \geq i\}$  if  $n \geq 1$  and  $x_{\{1\}} \geq 1$ , or  $H(\mathbf{x}) = 0$  otherwise. It may be shown that the  $h$ -index is an zero-insensitive OM3 aggeration operator, see [7], with:

$$H(\mathbf{x}) = \bigvee_{i=1, \dots, n}^n i \wedge \lfloor x_{\{i\}} \rfloor.$$

Interpretation: “an author has  $h$ -index of  $H$  if  $H$  of his/her  $n$  most cited papers have at least  $H$  citations each, and the other  $n - H$  papers are cited no more that  $H$  times each”. The  $h$ -index may also be expressed as a Sugeno integral [26] w.r.t. to a counting measure, cf. [27].

agop implementation: `index_h()`.

```
index_h(c(6,5,4,2,1,0,0,0,0,0))
## [1] 3
index_h(c(-1,3,4,2)) # only for x>=0
## Error: all elements in 'x' should be in [0,Inf]
```

**The  $g$ -index.** Egghe’s  $g$ -index [5]

The index is interesting from the computational point of view – it is calculated on the nondecreasing vector of cumulative sums, `cumsum(sort(x, decreasing=TRUE))`

**The  $w$ -index.**

**The MAXPROD-index.**

**The  $r_p$ -indices.**

**The  $l_p$ -indices.**

## 4 Visualization

### 4.1 Depicting producers

The `plot_producer()` function may be used to draw a graphical representation of a given numeric vector, i.e. what is sometimes called a citation function in scientometrics.

A given vector  $\mathbf{x} = (x_1, \dots, x_n)$  can be represented by a step function defined for  $0 \leq y < n$  and given by:

$$\pi(y) = x_{(n-\lfloor y \rfloor + 1)}.$$

This function may be obtained by setting `type == 'right.continuous'` argument in `plot_producer()`. Recall that  $x_{(i)}$  denotes  $i$ -th smallest value in  $\mathbf{x}$ .

On the other hand, for `type == 'left.continuous'` (the default), we get

$$\pi(y) = x_{(n-\lfloor y \rfloor + 1)}$$

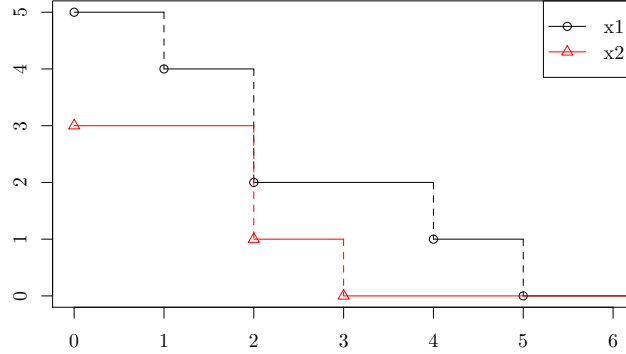
for  $0 < y \leq n$ .

Moreover, this function may depict the curve joining the sequence of points  $(0, x_{(n)}), (1, x_{(n)}), (1, x_{(n-1)}), (2, x_{(n-1)}), \dots, (n, x_{(1)})$ .

The `plot_producer()` function behaves much like the well-known R's `plot.default()` and allows for passing all its graphical parameters.

For example, let us depict the state of two given producers,  $\mathbf{x}^{(1)}$  and  $\mathbf{x}^{(2)}$ .

```
x1 <- c(5, 4, 2, 2, 1)
x2 <- c(3, 3, 1, 0, 0, 0, 0)
plot_producer(x1, extend=TRUE)
plot_producer(x2, add=TRUE, col=2, pch=2, extend=TRUE)
legend('topright', c('x1', 'x2'), col=c(1, 2), lty=1, pch=c(1, 2))
```



## 5 Pre-orders

.....

Let us consider the following relation on  $\mathbb{I}^{1,2,\dots}$ . For any  $\mathbf{x} \in \mathbb{I}^n$  and  $\mathbf{y} \in \mathbb{I}^m$  we write  $\mathbf{x} \trianglelefteq \mathbf{y}$  if and only if  $n \leq m$  and  $x_{\{i\}} \leq y_{\{i\}}$  for all  $i \in \min\{n, m\}$ . Of course,  $\trianglelefteq$  is a pre-order – it would have been a partial order, if we had defined it on the set of *sorted* vectors.

In other words, we say that an author  $X$  is (weakly) dominated by an author  $Y$ , if  $X$  has no more papers than  $Y$  and each the  $i$ th most cited paper of  $X$  has no more citations than the  $i$ th most cited paper of  $Y$ . Not that the  $m - n$  least cited  $Y$ 's papers are not taken into account here. Most importantly, however, there exist pairs of vectors that are *incomparable* with respect to  $\trianglelefteq$  (see the illustration below).

This pre-order is agop as `pord_weakdom()`.

```
c(pord_weakdom(5:1, 10:1), pord_weakdom(10:1, 5:1)) # 5:1 <= 10:1
## [1] TRUE FALSE
c(pord_weakdom(3:1, 5:4), pord_weakdom(5:1, 3:1)) # 3:1 ?? 5:4
## [1] FALSE FALSE
```

We have the following result (Gagolewski, Grzegorzewski, [8]). Let  $F \in \mathcal{E}(\mathbb{I})$ . Then  $F$  is symmetric, nondecreasing in each variable and arity-monotonic if and only if for any  $\mathbf{x}, \mathbf{y}$  if  $\mathbf{x} \trianglelefteq \mathbf{y}$ , then  $F(\mathbf{x}) \leq F(\mathbf{y})$ . Therefore, the class of impact functions may be equivalently defined as all the aggregation operators that are nondecreasing with respect to this preorder.

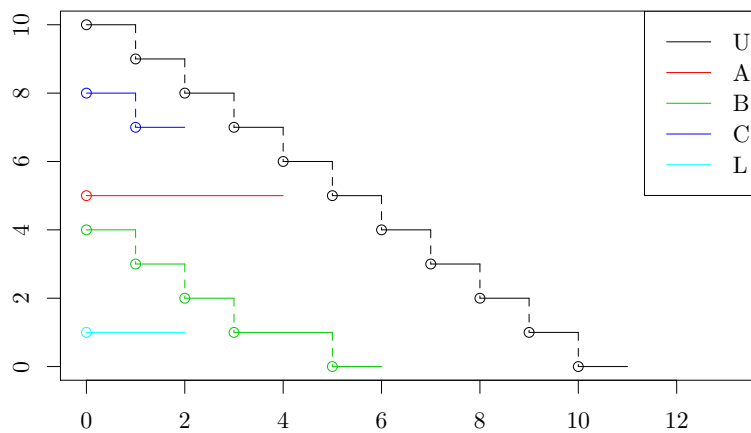
Additionally, we will write  $\mathbf{x} \triangleleft \mathbf{y}$  if  $\mathbf{x} \trianglelefteq \mathbf{y}$  and  $\mathbf{x} \neq \mathbf{y}$  (strict dominance).

**Example.** Let us consider the 5 following vectors.

```
ex1 <- list(
  U = 10:0,          # some upper bound
  A = c(5,5,5,5),    # moderate productivity & quality
  B = c(4,3,2,1,1,0), # high productivity
  C = c(8,7),         # high quality
  L = c(1,1)         # some lower bound
)
```

Plot of “citation” curves:

```
for (i in seq_along(ex1))
plot_producer(ex1[[i]], add=(i>1), col=i)
legend("topright", legend=names(ex1), col=1:length(ex1), lty=1)
```



get adjacency matrix for  $(\{A, B, C, L, U\}, \trianglelefteq)$ ....

```
ord <- rel_graph(ex1, pord_weakdom)
print(ord)

## 5 x 5 sparse Matrix of class "dtCMatrix"
##   U A B C L
## U 1 . . . .
## A 1 1 . . .
## B 1 . 1 . .
## C 1 . . 1 .
## L 1 1 1 1 1

is_reflexive(ord) # is reflexive
## [1] TRUE

is_transitive(ord) # is transitive
## [1] TRUE

is_total(ord)      # not a total preorder...
## [1] FALSE
```

We see that we have  $A??B$ ,  $A??C$ ,  $B??C$  (no pair from  $\{A, B, C\}$  is comparable w.r.t.  $\trianglelefteq$ ):

```
incomp <- get_incomparable_pairs(ord)
incomp <- incomp[incomp[,1]<incomp[,2],] # remove permutations: ((1,2), (2,1)) -> (1,2)
```

```

incomp[, ] <- rownames(ord)[incomp]
print(incomp) # all incomparable pairs
##      [,1] [,2]
## [1,] "A"  "B"
## [2,] "A"  "C"
## [3,] "B"  "C"

# the other way: generate maximal independent sets
lapply(get_independent_sets(ord), function(set) rownames(ord)[set])
## [[1]]
## [1] "A" "B" "C"

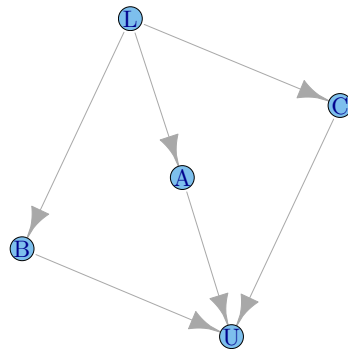
```

To draw the Hasse diagram, it will be good to de-transitivize the graph (for aesthetic reasons)....

```

require(igraph)
hasse <- graph.adjacency(de_transitive(ord))
set.seed(1234567) # igraph's drawing facilities are far from perfect
plot(hasse, layout=layout_fruchterman_reingold(hasse, dim=2))

```



$(\{A, B, C, L, U\}, \preceq)$  is not totally ordered, let's apply fair totalization (set  $x \preceq'' y$  and  $y \preceq'' x$  whenever  $\neg(x \preceq y \text{ or } y \preceq x)$ ) + calculate transitive closure

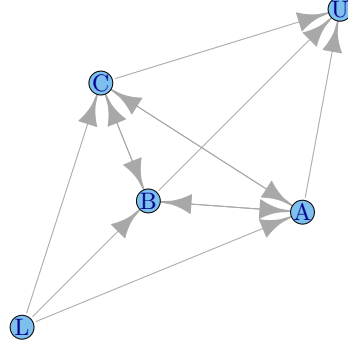
```

ord_total <- closure_transitive(closure_total_fair(ord)) # a total preorder
print(ord_total)
## 5 x 5 sparse Matrix of class "dgCMatrix"
##   U A B C L
## U 1 . . . .
## A 1 1 1 1 .
## B 1 1 1 1 .
## C 1 1 1 1 .
## L 1 1 1 1 1

hasse <- graph.adjacency(de_transitive(ord_total))
set.seed(1234)
plot(hasse, layout=layout_fruchterman_reingold(hasse, dim=2))

```





...Note that each total preorder  $\preceq''$  induces an equivalence relation ( $x \simeq y$  iff  $x \preceq'' y$  and  $y \preceq'' x$ ; the equivalence classes may be ordered with  $\preceq''$ ). These may be explored with the `get_equivalence_classes()` function...

```
supply(get_equivalence_classes(ord_total), function(set) rownames(ord)[set])
## [[1]]
## [1] "L"
##
## [[2]]
## [1] "A" "B" "C"
##
## [[3]]
## [1] "U"
```

Thus, we've obtained  $L \prec (A \simeq B \simeq C) \prec U$ .

## 6 Aggregation Operators from the Probabilistic Perspective

Theory of aggregation looks on the aggregation operators from the algebraic/calculus perspective. Of course, we should always be interested in their probabilistic properties, e.g. in i.i.d. RVs models (the simplest and the most “natural” ones in statistics), cf. [10] for discussion.

In such case we assume that input data are in fact realizations of some random samples.

In probability, an aggregation operator is simply called a *statistic* (formalism.....)

Let  $(X_1, \dots, X_n)$  i.i.d.  $F$ , where  $\text{supp } F = \mathbb{I}$ .

....

In social phenomena modeling, if  $F$  is continuous, we often assume that the underlying density  $f$  is decreasing and convex on  $\mathbb{I}$ , possibly with heavy-tails. E.g. in the bibliometric impact assessment problem, this assumption reflect the fact that a high paper valuation is more difficult to obtain than the lower one, most of the papers have very small valuation (near 0), and the probability of attaining a high note decreases no slower than linearly.

## 6.1 Some Notable Probability Distributions

### 6.1.1 Pareto-Type II Distribution

Many generalizations of the Pareto distribution have been proposed (GPD, *Generalized Pareto Distributions*, cf. e.g. [28, 32]). Here we will introduce the so-called Pareto-Type II (Lomax) distribution, which has support  $\mathbb{I} = [0, \infty]$ .

Formally,  $X$  follows the Pareto-II distribution with shape parameter  $k > 0$  and scale parameter  $s > 0$ , denoted  $X \sim \text{P2}(k, s)$ , if its density is of the form

$$f(x) = \frac{ks^k}{(s+x)^{k+1}} \quad (x \geq 0). \quad (1)$$

The cumulative distribution function of  $X$  is then:

$$F(x) = 1 - \frac{s^k}{(s+x)^k} \quad (x \geq 0). \quad (2)$$

TO DO: agop: `dpareto2()` – (1), `ppareto2()` – (2), and `qpareto2()`... `rpareto2()`.....

**Properties.** The expected value of  $X \sim \text{P2}(k, s)$  exists for  $k > 1$  and is equal to

$$\mathbb{E}X = \frac{s}{k-1}.$$

Variance exists for  $k > 2$  and is equal to

$$\text{Var } X = \frac{ks^2}{(k-2)(k-1)^2}.$$

More generally, the  $i$ -th raw moment for  $k > i$  is given by:

$$\mathbb{E}X^i = \frac{\Gamma(i+1)\Gamma(k-i)}{\Gamma(k+1)} ks^i.$$

For a fixed  $s$ , if  $X \sim \text{P2}(k_x, s)$  and  $Y \sim \text{P2}(k_y, s)$ ,  $k_x < k_y$ , then  $X$  stochastically dominates  $Y$ , denoted  $X \succ Y$ . On the other hand, for a fixed  $k$ ,  $X \sim \text{P2}(k, s_x)$ ,  $Y \sim \text{P2}(k, s_y)$ ,  $s_x > s_y$ , implies  $X \succ Y$ .

Interestingly, if  $X \sim \text{P2}(k, s)$ , then the conditional distribution of  $X - t$  given  $X > t$ , is  $\text{P2}(k, s+t)$   $t \geq 0$ .

Additionally, it might be shown that if  $X \sim \text{P2}(k, s)$ , then  $\ln(s+X)$  has c.d.f.  $F(x) = 1 - s^k e^{-kx}$  and density  $f(x) = ks^k e^{-kx}$  for  $x \geq \ln s$ , i.e. has the same distribution as  $Z + \ln s$ , where  $Z \sim \text{Exp}(k) \equiv \Gamma(1, 1/k)$ .

**Parameter estimation.** Let  $\mathbf{x} = (x_1, \dots, x_n)$  be a realization of the Pareto-Type II i.i.d. sample with known  $s > 0$ . The unbiased (corrected) maximum likelihood estimator for  $k$ :

$$\hat{k}(\mathbf{x}) = \frac{n-1}{\sum_{i=1}^n \ln\left(1 + \frac{1}{s}x_i\right)}.$$

It may be shown that for  $n > 2$  it holds  $\text{Var } \hat{k}(\mathbf{x}) = k^2 \frac{1}{n-2}$ .

TO DO: agop: `pareto2.mleestimate()`

For both unknown  $k$  and  $s$  we have:

$$\begin{cases} \hat{k} = \frac{n}{\sum_{i=1}^n \ln(1+x_i/\hat{s})}, \\ 1 + \frac{1}{n} \sum_{i=1}^n \ln(1+x_i/\hat{s}) - \frac{n}{\sum_{i=1}^n (1+x_i/\hat{s})^{-1}} = 0. \end{cases} \quad (3)$$

The second equation must be solved, unfortunately, numerically. The estimation procedure has been implemented in **agop** as TO DO: **pareto2.mleksestimate()**.... It is worth noting that the above system of equations may sometimes have no solution (as the local minimum of the likelihood function may not exist, see [4] for discussion).

In this case one of the estimators worth noting (and often better than MLE) was proposed in [33]. The Zhang-Stevens MMS (*minimum mean square error*) (Bayesian) estimator has relatively small bias (often positive) and mean squared error. In **agop** it is available as TO DO: **pareto2.zsestimate()**.

**Goodness-of-fit tests.** TO BE DONE....

**Applications.** TO DO

**Two-sample  $F$ -test.** The following simple test was introduced in [10]. Let  $(X_1, X_2, \dots, X_{n_1})$  *i.i.d.*  $P2(k_1, s)$  and  $(Y_1, Y_2, \dots, Y_{n_2})$  *i.i.d.*  $P2(k_2, s)$ , where  $s$  is an a-priori known scale parameter. We are going to verify the null hypothesis  $H_0 : k_1 = k_2$  against the two-sided alternative hypothesis  $K : k_1 \neq k_2$ .

It might be shown that  $\sum_{i=1}^n \ln(s + X_i) - n \ln s \sim \Gamma(n, 1/k)$ . This implies that under  $H_0$ , the following test statistic follows the Snedecor  $F$  distribution:

$$R(\mathbf{X}, \mathbf{Y}) = \frac{n_1 \sum_{i=1}^{n_2} \ln\left(1 + \frac{Y_i}{s}\right)}{n_2 \sum_{i=1}^{n_1} \ln\left(1 + \frac{X_i}{s}\right)} \stackrel{H_0}{\sim} F[2n_2, 2n_1]. \quad (4)$$

The null hypothesis is accepted iff

$$R(\mathbf{x}, \mathbf{y}) \in [\mathbf{qf}(\frac{\alpha}{2}, 2n_2, 2n_1), \mathbf{qf}(1 - \frac{\alpha}{2}, 2n_2, 2n_1)],$$

where  $\mathbf{qf}(q, d_1, d_2)$  denotes the  $q$ -quantile of  $F^{[d_1, d_2]}$

The  $p$ -value may be determined as follows:

$$p = 2 \left( \frac{1}{2} - \left| \mathbf{pf}(R(\mathbf{x}, \mathbf{y}), 2n_2, 2n_1) - \frac{1}{2} \right| \right), \quad (5)$$

where  $\mathbf{pf}(x, d_1, d_2)$  is the c.d.f. of  $F^{[d_1, d_2]}$ .

TO DO: **pareto2.ftest()**.

## 6.2 Stochastic Properties of Aggregation Operators

OWA, L-statistics

OWMax, S-statistics

$h$ -index and its distribution

## 7 NEWS/CHANGELOG

```
Warning: cannot open file '../NEWS': No such file or directory
Error: cannot open the connection
```

**Acknowledgments.** This document has been generated with L<sup>A</sup>T<sub>E</sub>X, knitr and the tikzDevice package for R. Their authors' wonderful work is fully appreciated.

The contribution of Marek Gagolewski was partially supported by the European Union from resources of the European Social Fund, Project PO KL "Information technologies: Research and their interdisciplinary applications", agreement UDA-POKL.04.01.01-00-051/10-00 (March-June 2013), and by FNP START Scholarship from the Foundation for Polish Science (2013).

## Bibliography

- [1] Cena A., Gagolewski M., OM3: ordered maxitive, minitive, and modular aggregation operators – Part I: Axiomatic analysis under arity-dependence (AGOP 2013), in-press, 2013.
- [2] Cena A., Gagolewski M., OM3: ordered maxitive, minitive, and modular aggregation operators – Part II: A simulation study (AGOP 2013), in-press, 2013.
- [3] Choquet G., Theory of capacities, *Ann. Inst. Fourier* 5, pp. 131–292, 1954.
- [4] Del Castillo J., Daoudi J., Estimation of the Generalized Pareto Distribution. *Statistics and Probability Letters*, 79, pp. 684–688, 2009.
- [5] Egghe L., An improvement of the  $H$ -index: the  $G$ -index, *ISSI Newsletter* 2(1), pp. 8–9, 2006.
- [6] Franceschini F., Maisano D.A., Structured evaluation of the scientific output of academic research groups by recent  $h$ -based indicators. *Journal of Informetrics*, 5, pp. 64–74, 2011.
- [7] Gagolewski M., On the Relationship Between Symmetric Maxitive, Minitive, and Modular Aggregation Operators, *Information Sciences* 221, pp. 170–180, 2013.
- [8] Gagolewski M., Grzegorzewski P., Possibilistic Analysis of Arity-Monotonic Aggregation Operators and Its Relation to Bibliometric Impact Assessment of Individuals, *International Journal of Approximate Reasoning* 52(9), pp. 1312–1324, 2011.
- [9] Gagolewski M., Mesiar R., Aggregating Different Paper Quality Measures with a Generalized  $h$ -index, *Journal of Informetrics* 6(4), pp. 566–579, 2012.
- [10] Gagolewski M., *Aggregation Operators and Their Application in a Formal Model for Quality Evaluation System of Scientific Research* (Wybrane operatory agregacji i ich zastosowanie w modelu formalnym systemu jakości w nauce), PhD Thesis, Systems Research Institute, Polish Academy of Sciences, 2011 (In Polish).
- [11] Gagolewski M., Bibliometric Impact Assessment with R and the CITAN Package, *Journal of Informetrics* 5(4), pp. 678–692, 2011.
- [12] Gagolewski M., Grzegorzewski P., A Geometric Approach to the Construction of Scientific Impact Indices, *Scientometrics* 81(3), 2009, pp. 617–634.

- [13] Gagolewski M., Statistical Hypothesis Test for the Difference between Hirsch Indices of Two Pareto-Distributed Random Samples, In: Kruse R. et al (Eds.), *Synergies of Soft Computing and Statistics for Intelligent Data Analysis (AISC 190)*, Springer-Verlag, Heidelberg, pp. 359-367, 2013.
- [14] Gagolewski M., On the Relation Between Effort-Dominating and Symmetric Minitive Aggregation Operators, In: Greco S. et al (Eds.), *Advances in Computational Intelligence, Part III (CCIS 299)*, Springer-Verlag, Heidelberg, pp. 276-285, 2012.
- [15] Gagolewski M., Grzegorzewski P., Axiomatic Characterizations of (quasi-) L-statistics and S-statistics and the Producer Assessment Problem, In: Galichet S., Montero J., Mauris G. (Eds.), *Proc. EUSFLAT/LFA 2011*, Atlantic Press, pp. 53-58, 2011.
- [16] Gagolewski M., Grzegorzewski P., S-Statistics and Their Basic Properties, In: Borgelt C. et al (Eds.), *Combining Soft Computing and Statistical Methods in Data Analysis (AISC 77)*, Springer-Verlag, Heidelberg, pp. 281-288, 2010.
- [17] Gagolewski M., Grzegorzewski P., Arity-Monotonic Extended Aggregation Operators, In: Hüllermeier E., Kruse R., Hoffmann F. (Eds.), *Information Processing and Management of Uncertainty in Knowledge-Based Systems (CCIS 80)*, Springer-Verlag, Heidelberg, pp. 693-702, 2010.
- [18] Gagolewski M., Cena A., `agop`: Aggregation Operators in R, [www.rexamine.com/resources/agop/](http://www.rexamine.com/resources/agop/), 2013.
- [19] Grabisch M., Marichal J.-L., Mesiar R., Pap E., *Aggregation functions*, Cambridge University Press, 2009.
- [20] Hirsch J.E., An index to quantify individual's scientific research output, *Proceedings of the National Academy of Sciences* 102(46), pp. 16569–16572, 2005.
- [21] Kosmulski M., MAXPROD – A new index for assessment of the scientific output of an individual, and a comparison with the  $h$ -index, *Cybermetrics* 11(1), 2007.
- [22] Quesada A., Monotonicity and the Hirsch index. *Journal of Informetrics*, 3(2), pp. 158–160, 2009.
- [23] Quesada A., More axiomatics for the Hirsch index. *Scientometrics*, 82, pp. 413–418, 2010.
- [24] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. <http://www.R-project.org>.
- [25] Rousseau R., Woeginger's axiomatisation of the  $h$ -index and its relation to the  $g$ -index, the  $h(2)$ -index and the  $r^2$ -index. *Journal of Informetrics*, 2(4), pp. 335–340, 2008.
- [26] Sugeno M., *Theory of fuzzy integrals and its applications*, PhD dissertation, Tokyo Institute of Technology, Meguro-ku, Tokyo, 1974.
- [27] Torra V., Narukawa Y., The  $h$ -index and the number of citations: Two fuzzy integrals, *IEEE Transactions on Fuzzy Systems* 16(3), pp. 795–797, 2008.

- [28] Villasenor-Alva J., Gonzalez-Estrada E., A bootstrap goodness of fit test for the Generalized Pareto Distribution. *Computational Statistics and Data Analysis*, 53(11), pp. 3835–3841, 2009.
- [29] Woeginger G.J., An axiomatic analysis of Egghe’s  $g$ -index. *Journal of Informetrics*, 2(4), pp. 364–368, 2008.
- [30] Woeginger G.J., An axiomatic characterization of the Hirsch-index. *Mathematical Social Sciences*, 56(2), pp. 224–232, 2008.
- [31] Woeginger G.J., A symmetry axiom for scientific impact indices. *Journal of Informetrics*, 2, pp. 298–303, 2008.
- [32] Zhang J., Improving on estimation for the Generalized Pareto Distribution. *Technometrics*, 52(3), pp. 335–339, 2010.
- [33] Zhang J., Stephens M.A., A new and efficient estimation method for the Generalized Pareto Distribution. *Technometrics*, 51(3), pp. 316–325, 2009.