A Guide to the agop 0.01-devel Package for R

Aggregation Operators in R

Marek Gagolewski[1,2], Anna Cena[1,2]
[1] Systems Research Institute, Polish Academy of Sciences
ul. Newelska 6, 01-447 Warsaw, Poland
[2] Rexamine, Email: {gagolews,cena}@rexamine.com
www.rexamine.com/resources/agop/

May 20, 2013

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

Contents

1 Getting started

intro..... aggregation.... [2]

R 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. Perhaps you may also wish to install RStudio, a convenient development environment for R. It is available at www.rsudio.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 [1]. 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 often 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 $\mathbf{x} \le \mathbf{y}$ iff for all i it holds $x_i \le y_i$

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

Let $x_{(i)}$ denote the ith order statistic...

2.1 Aggregation Operators and Their Basic Properties

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

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

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. (Woeginger, [**?**, **?**, **?**]; Rousseau, [**?**]; Quesada, [**?**, **?**]; Gagolewski, Grzegorzewski, [**?**]; Franceschini, Maisano, [**?**]), that each aggregation operator $\mathsf{J} : \mathbb{I}^{1,2,\dots} \to \mathbb{I}$ 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 $\mathsf{J}(\mathbf{x}) \leq \mathsf{J}(\mathbf{y})$. On the other hand, axiom (b) is fulfilled iff for any $\mathbf{x} \in \mathbb{I}^{1,2,\cdots}$ and $y \in \mathbb{I}$ it holds $\mathsf{J}(\mathbf{x}) \leq \mathsf{J}(x_1, \dots, x_n, y)$. Lastly, requirement (c) holds iff for all n and $\mathbf{x} \in \mathbb{I}^n$ we have $\mathsf{J}(\mathbf{x}) = \mathsf{J}(x_{\{1\}}, \dots, x_{\{n\}})$, where $x_{\{i\}}$ denotes the ith largest value from $\mathbf{x}$, i.e. its $(n-i+1)$th order statistic.

3 Predefined Classes of Aggregation Operators in agop

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()....

...or in a list, especially when they are not of the same length.... lapply.... sapply.....

weighted average, OWA
WMin, OWMin, qI
WMax, OWMax, qS
qL
OM3
h-index, g-, w-, r_p-, l_p-

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:

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

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

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

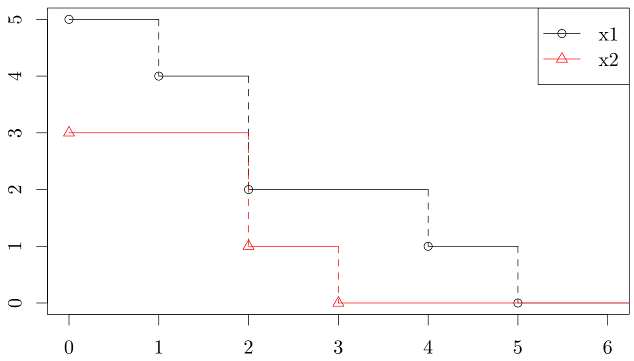
for $0 < y \leq n$.

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

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 $\mathbb{I}^{1,2,\cdots}$. 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 ith most cited paper of X has no more citations than the ith 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 in 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, [?]). Let $\mathsf{F} \in \mathcal{E}(\mathbb{I})$. 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 $\mathsf{F}(\mathbf{x}) \le \mathsf{F}(\mathbf{y})$. Therefore, the class of impact functions may be equivalently defined as all the aggregation operators that are nondecreasing with respect to this preorder.

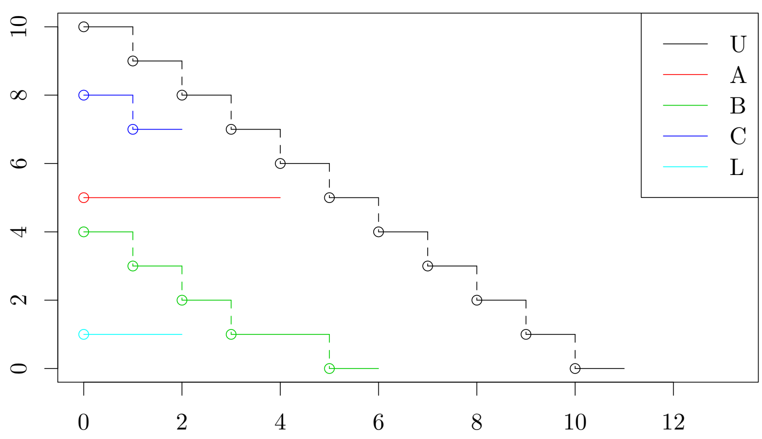
Additionally, we will write $\mathbf{x} \triangleleft \mathbf{y}$ if $\mathbf{x} \trianglelefteq \mathbf{y}$ and $\mathbf{x} \neq \mathbf{y}$ (strict dominance).

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 $(\{A, B, C, L, U\}, \trianglelefteq)$....

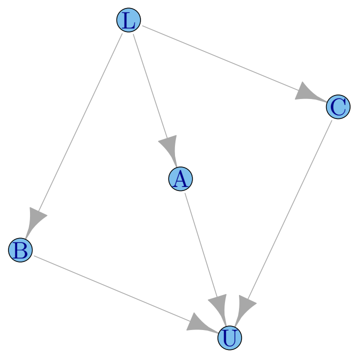
```
ord <- rel_graph(ex1, pord_weakdom, disable.check=TRUE)
print(ord)

## 5 x 5 sparse Matrix of class "dgCMatrix"
##   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. $\trianglelefteq$).

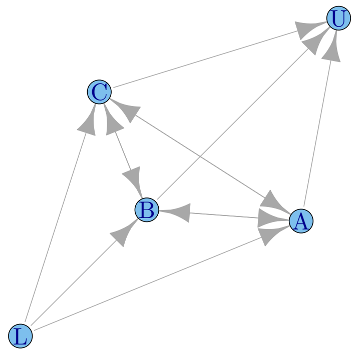
To draw the Hasse diagram, it will be good to de-transitivize the graph (for æsthetic reasons)....

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



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

```
ord_total <- make_transitive(make_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))
```



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

6 NEWS/CHANGELOG

```
agop package NEWS

****************************************************************************

0.01 /under development/

* initial release
```

Acknowledgments. This document has been generated with LaTeX, 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] Gagolewski M., Cena A., agop: Aggregation Operators in R, www.rexamine.com/resources/agop/, 2013.

[2] Grabisch M., Marichal J.-L., Mesiar R., Pap E., *Aggregation functions*, Cambridge University Press, 2009.