Lorenz theory of symmetric relative concentration and similarity, incorporating variable array length

EGGHE, Leo & ROUSSEAU, Ronald (2006) Lorenz theory of symmetric relative concentration and similarity, incorporating variable array length. In: MATHEMATICAL AND COMPUTER MODELLING, 44(7-8). p. 628-639.

DOI: 10.1016/j.mcm.2006.02.001
Handle: http://hdl.handle.net/1942/750

Lorenz theory of symmetric relative concentration and similarity, incorporating variable array length

by

L. Egghe, Limburgs Universitair Centrum (LUC), Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), IBW, Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium
Email: leo.egghe@luc.ac.be

R. Rousseau, KHBO, IWT, Zeedijk 101, B-8400 Oostende, Belgium
and
Universiteit Antwerpen (UA), IBW, Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium
Email: ronald.rousseau@khbo.be

ABSTRACT

This paper extends the Lorenz theory, developed in [L. Egghe and R. Rousseau. Symmetric and asymmetric theory of relative concentration and applications. Scientometrics 52(2), 261-290, 2001], so that it can deal with comparing arrays of variable length. We show that in this case we need to divide the Lorenz curves by certain types of increasing functions of the array length N.

[1] Permanent address

Key words and phrases: Lorenz, symmetric relative concentration, similarity, Pearson correlation coefficient, Spearman correlation coefficient.

We then prove that, in this theory, adding zeros to two arrays, increases their similarity, a property that is not satisfied by the Pearson correlation coefficient.

Among the many good similarity measures, satisfying the developed Lorenz theory, we deduce the correlation coefficient of Spearman, hence showing that this measure can be used as a good measure of symmetric relative concentration (or similarity).

I. Introduction

In Egghe and Rousseau (2001) see also Rousseau (2001), a Lorenz concentration theory is developed in order to have a framework in which good measures of symmetric relative concentration or similarity can be derived. What do we mean by this ? Suppose we have two vectors $X = (x_1,...,x_N)$ and $Y = (y_1,...,y_N)$ with $x_i, y_i \geq 0$, $i = 1,...,N$. They can be interpreted as collaboration vectors of two authors (again called X and Y) where, for each $i = 1,...,N$, x_i denotes the number of times author X collaborated (i.e. was a co-author) with author i (and similarly for y_i and Y). The vectors X and Y can also be interpreted as co-citation vectors (as done in Ahlgren, Jarneving and Rousseau (2003)): here x_i denotes the number of times author X and author i are co-cited (and similarly for y_i and Y). These are just examples: the vectors X and Y can be interpreted in many more ways. In this framework the following question is natural: How can we compare X and Y, i.e. how similar are X and Y? This information is important since it reveals mutual relations of (e.g.) authors in a certain research field, expressed by their similarity in (e.g.) collaboration with other authors or co-citedness with other authors. It is e.g. clear that authors $X = (3,5,2,1,0)$ and $Y = (3,4,2,1,0)$ are much more similar than authors $X' = (3,5,2,1,0) = X$ and $Y' = (0,1,1,0,3)$. Note the value 0 here denoting a case of no collaboration or co-citation (using again these examples) – a case that will become important in the sequel.

In Egghe and Rousseau (2001) the classical Lorenz model (see e.g. Egghe and Rousseau (1990, 2001)) was extended as follows in order to compare two vectors X and Y as above. Let

$X = (x_1,...,x_N)$, $Y = (y_1,...,y_N)$, $x_i, y_i \geq 0$ for all $i = 1,...,N$ with $X \neq 0$ and $Y \neq 0$, where $0 = (0,...,0)$ is the zero-vector of length N. Define, for all $i = 1,...,N$

$$a_i = \frac{x_i}{\sum_{j=1}^{N} x_j} \tag{1}$$

$$\alpha_i = \frac{y_i}{\sum_{j=1}^{N} y_j} \tag{2}$$

and denote $A_X = (a_1,...,a_N)$, $A_Y = (\alpha_1,...,\alpha_N)$. Note that

$$\sum_{j=1}^{N} a_j = \sum_{j=1}^{N} \alpha_j = 1 \tag{3}$$

Next we form the vector

$$A_X - A_Y = (a_1 - \alpha_1,...,a_N - \alpha_N) \tag{4}$$

which we assume to be ranked in decreasing order. Note that we do not take into account the order of the coordinates in a vector; therefore we will henceforth use the terminology “array” for X and Y as above where the order of the coordinates does not matter (but we keep on comparing the same coordinates in X and Y as expressed by (4)).

The Lorenz curve of X,Y, denoted $L_{X,Y}$ is the curve consisting of the line segments connecting the consecutive points $(0,0)$ and

$$\left(\frac{i}{N}, \sum_{j=1}^{i} (a_j - \alpha_j)\right) \tag{5}$$

$i = 1,...,N$, hence ending in $(1,0)$ because of (3). Since $A_X -\ A_Y$ is decreasing, $L_{X,Y}$ is concave.

This Lorenz curve is the basic tool in the comparison of the arrays X and Y: the more similar X and Y are, the lower $L_{X,Y}$ (verify that $L_{X,Y} = [0,1]x\{0\}$ for $X = Y$). Of course our similarity theory should comprise the fact that comparing X and Y must be the same as comparing Y and X. We showed in Egghe and Rousseau (2001) that $L_{Y,X}$ is the Lorenz curve $L_{X,Y}$ but mirrored over the vertical line $x = \frac{1}{2}$ (we denote $L_{Y,X} = R(L_{X,Y})$). Since comparing X with Y must be the same as comparing Y with X we will denote

$$D = \{X, Y\} = \{Y, X\} \tag{6}$$

as the "duo" X,Y. We can now introduce the following partial order relation. Let $D = \{X, Y\}$ and $D^{'} = \{X^{'}, Y^{'}\}$ be two duos (not necessarily of the same length: say X and Y have length N and $X^{'}, Y^{'}$ have length $N^{'}$). We define

$$D\ ^{3}\ D^{'} \tag{7}$$

and say that <u>duo D is more similar than duo D'</u> if

$$L_{X,Y}\ £\ L_{X^{'},Y^{'}} \tag{8}$$

or

$$L_{X,Y}\ £\ L_{Y^{'},X^{'}} = R\left(L_{X^{'},Y^{'}}\right) \tag{9}$$

and say that $D > D^{'}$ (duo D is strictly more similar than duo $D^{'}$) if at least one of the inequalities in (8) or (9) is strict, meaning that these Lorenz curves differ in at least one (hence infinitely many) point(s).

We say that f is a good measure of symmetric relative concentration (or inequality) if f is defined on the set of these duos and if $D > D'$ implies $f(D) < f(D')$. We say that f is a good Lorenz similarity function if f is defined on the set of duos and if $D > D'$ implies $f(D) > f(D')$. Note that f is a Lorenz similarity measure iff -f is a good measure of symmetric relative concentration. Note also that (6) forces f to be symmetric.

The requirement for f to be a good measure of symmetric relative concentration is hence such that (8) OR (9) (with strict inequality) must lead to $f(D) < f(D')$. In Egghe and Rousseau (2001) we only required that (8) (with strict inequality) implies $f(D) < f(D')$ but this is obviously the same for symmetric functions (which we assume). The relations (8) and (9) (in the strict sense) are, however, not equivalent. In Egghe and Rousseau (2004a) an example is given where $L_{X,Y} < L_{X',Y'}$ but where $L_{X,Y}$ and $L_{Y',X'}$ intersect.

In the next section we will study new similarity requirements as formulated in Ahlgren, Jarneving and Rousseau (2003) (in short: the A-J-R requirements) involving the comparison of two duos D and D' with different array length. We show that the above theory does not follow this requirement and, hence, a modification of the above model is presented such that, if the array lengths are fixed, the new model is equivalent with the one above and, in addition, such that the new model satisfies the A-J-R requirements.

In the third section, this new model is studied in case X and Y are arrays of ranks of objects. We show that the new model generates the Spearman rank correlation coefficient, hence satisfying the Lorenz order requirements and hence, since the new Lorenz theory implies the A-J-R requirements, the Spearman correlation coefficient satisfies the Ahlgren, Jarneving and Rousseau requirement (contrary to the Pearson correlation coefficient).

II. A drawback of the existing Lorenz theory of symmetric relative concentration and an improvement of this theory.

In Ahlgren, Jarneving and Rousseau (2003) the following two requirements for good similarity measures f are formulated. Let $X = (x_1,...,x_N)$, $Y = (y_1,...,y_N)$,

$X\$M = \left(x_1,...,x_N,\underbrace{0,...,0}_{M\ \text{times}}\right)$, $Y\$M = \left(y_1,...,y_N,\underbrace{0,...,0}_{M\ \text{times}}\right)$, where $M \in \mathbb{N}$. Then

(i)

$$f(X,Y) \le f(X\$M, Y\$M) \qquad (10)$$

(ii) if

$$f(X,Y) \le f(X',Y')$$

then

$$f(X\$M, Y\$M) \le f(X'\$M, Y'\$M) \qquad (11)$$

The idea behind these requirements is that X$M and Y$M are not less similar than X and Y and that the “operation” $M should not destroy existing similarity inequalities. Why should this be? In the context of collaboration or co-citation this means that if two authors do not collaborate with (or are not co-cited by) a certain group of M authors in a field (e.g. active in a subfield in which authors X and Y are both not active), this makes X and Y more similar (or at least not less similar). This model hence still incorporates the case where the $M operation has no influence on the similarity measure (i.e. having an equality in (10)) which is also of interest in the following case: suppose, for two given authors, represented by their arrays X and Y, we add M new persons but who are not at all related to the field of research of authors

X and Y (say we add the pope and the president of the USA, here M=2). Then X and Y are not more similar because they both did not publish with (or were not co-cited by) these two persons.

Obviously, when comparing X,Y with XM,YM as in (10) we consider two duos of unequal length: X,Y have length N and XM,YM have length M+N>N.

According to our Lorenz theory of relative symmetric concentration, condition (10) could be logically extended tot the requirement that

$$\{X\$M, Y\$M\}\text{³ } \{X, Y\}$$

hence that

$$L_{X\$M,Y\$M} \text{ £ } L_{X,Y} \tag{12}$$

or

$$L_{X\$M,Y\$M} \text{ £ } L_{Y,X} \tag{13}$$

However, (12) nor (13) can be true due to the following proposition.

<u>Proposition II.1</u>: In the Lorenz theory developed in Section I we have, for $X \neq Y$,

$$L_{X,Y} < L_{X\$M,Y\$M} \;, \tag{14}$$

where < is strict in every point $x\,\text{Î}\,\,[0,1]$ except for a possible horizontal maximum (situated above a closed interval $\text{Ì}\;\;]0,1[$ of $L_{X,Y}$,) where there is equality.

<u>Proof</u>:

Since $\left(a_j - \alpha_j\right)_{j=1}^{N}$ is decreasing and by (3) we have that $i_0\,\text{Î}\;\;\{1,...,N\}$ exists such that i_0 is the highest index for which $a_{i_0} - \alpha_{i_0}\text{ ³ }\;0$. So $L_{X,Y}$ is increasing (followed by a possible horizontal interval) on the interval $\left[0,\frac{i_0}{N}\right]$ after which $L_{X,Y}$ is strictly decreasing. The part of $L_{X,Y}$ above the interval $\left[0,\frac{i_0}{N}\right]$ is in $L_{X\$M,Y\$M}$ homothetically transformed to the abscissa interval $\left[0,\frac{N}{N+M}\frac{i_0}{N}\right]=\left[0,\frac{i_0}{N+M}\right]$, i.e

$$L_{X,Y}(x)=L_{X\$M,Y\$M}\left(\frac{N}{M+N}x\right)$$

Since $L_{X,Y}(x)$ is not decreasing on this interval we have that here $L_{X,Y}<L_{X\$M,Y\$M}$ except for a possible equality at the end of this interval. The maximal value $L_{X,Y}\left(\frac{i_0}{N}\right)$ of $L_{X,Y}$ is attained for $L_{X\$M,Y\$M}$ in the abscissa $\frac{N}{M+N}\frac{i_0}{N}=\frac{i_0}{M+N}$ and continued (because of the M zero values) until the abscissa $\frac{i_0+M}{M+N}$.

On the interval $\left[\frac{i_0}{N},1\right]$ $L_{X,Y}$ decreases strictly: this is for $L_{X\$M,Y\$M}$ transformed homothetically on the interval $\left[\frac{i_0+M}{M+N},1\right]$. Since $L_{X,Y}$ decreases strictly on $\left[\frac{i_0}{N},1\right]$ we hence have that $L_{X,Y}<L_{X\$M,Y\$M}$ on $\left[\frac{i_0+M}{M+N},1\right]$. This proves the proposition.

Proposition II.1 contradicts both (12) and (13). This is clear for (12). Let now $x\in]0,1[$ and consider $L_{X\$M,Y\$M}(x)$ and $L_{X\$M,Y\$M}(1-x)$. The value $x\in]0,1[$ can be chosen so that

$$L_{X\$M,Y\$M}(1-x)\geq L_{X\$M,Y\$M}(x)$$

(otherwise replace x by $1-x$) and so that (by Proposition II.1)

$$L_{X,Y}(x) < L_{X\$M,Y\$M}(x)$$

So we have

$$\begin{aligned} L_{Y,X}(1-x) &= L_{X,Y}(x) \\ &< L_{X\$M,Y\$M}(x) \\ &\le L_{X\$M,Y\$M}(1-x) \end{aligned}$$

contradicting (13).

We will now adapt the Lorenz theory above so that it is the same theory if the array length is constant but such that (12) will be valid even with strict inequality. The modification is as follows: instead of $L_{X,Y}$ consisting of line segments connecting (0,0) and the points given by equation (5) we will define the Lorenz curve (again denoted by $L_{X,Y}$ - confusion will not be possible since we will not use the previous one anymore in the sequel) as linearly connecting the consecutive points (0,0) and

$$\left(\frac{i}{N}, \sum_{j=1}^{i} \frac{a_j - \alpha_j}{\varphi(N)} \right) \tag{15}$$

, $i = 1,\ldots,N$, where φ is a certain function of N, to be determined in the next theorem.

<u>**Theorem II.2**</u>: Let φ be a function such that

$$\frac{\varphi(N+1)}{\varphi(N)} \ge \frac{N+1}{N} \tag{16}$$

(e.g. $\varphi(N) = N$). Then, using (15) for the construction of our Lorenz curves we have, for all arrays $X = (x_1, ..., x_N)$, $Y = (y_1, ..., y_N)$ and for all $M \in \mathbb{N}_0 = \{1, 2, 3, ...\}$:

(i)

$$L_{X,Y} > L_{X\$M,Y\$M} \tag{17}$$

(ii)

$$L_{X,Y} < L_{X',Y'} \tag{18}$$

$$\Rightarrow \quad L_{X\$M,Y\$M} < L_{X'\$M,Y'\$M}$$

i.e. the A-J-R requirements, as expressed by (i) and (ii), i.e. extended to conditions on Lorenz curves, are valid.

Proof:

(i) It suffices (by induction) to prove the theorem for $M = 1$.

Let $i_0 \in \{1, ..., N\}$ be the last index such that $a_{i_0} - \alpha_{i_0} \geq 0$ (this exists since the sequence $(a_j - \alpha_j)_{j=1}^{N}$ is decreasing and by (3). Let now $i \in \{1, ..., i_0\}$. Adding a zero to both X and Y forces the point $P_i = \left(\frac{i}{N}, z_i\right)$ on $L_{X,Y}$ to move to the point $P_i' = \left(\frac{i}{N+1}, z_i \frac{\varphi(N)}{\varphi(N+1)}\right)$. The equation of the straight line OP_i is $y = \frac{Nz_i}{i}x$. For $x = \frac{i}{N+1}$ we have on OP_i that $y = \frac{Nz_i}{N+1}$. So P_i' is under or on this line if

$$z_i \frac{\varphi(N)}{\varphi(N+1)} \leq \frac{Nz_i}{N+1}$$

hence if

$$\frac{\varphi(N+1)}{\varphi(N)} \geq \frac{N+1}{N}$$

which is so by (16). This also implies that P_i' is under $L_{X,Y}$ (strict if $i \neq 1$).

Let now $i \in \{i_0,\ldots,N-1\}$. Adding a zero to both X and Y forces the point $Q_i = \left(\frac{i}{N}, z_i\right)$ on $L_{X,Y}$ to move to the point $Q_i' = \left(\frac{i+1}{N+1}, z_i\frac{\varphi(N)}{\varphi(N+1)}\right)$. The equation of the straight line Q_iE $(E = (1,0))$ is

$$y = \frac{z_iN}{i-N}(x-1)$$

For $x = \frac{i+1}{N+1}$ we have on Q_iE

$$y = \frac{z_iN}{i-N}\left(\frac{i+1}{N+1} - 1\right) = z_i\frac{N}{N+1}$$

So Q_i' is under or on this line if

$$z_i\frac{\varphi(N)}{\varphi(N+1)} \leq z_i\frac{N}{N+1}$$

again satisfied because of (16). This also implies that Q_i' is under $L_{X,Y}$ (strict if $i \neq N-1$). This proves (17). Note that the point T on $L_{X,Y}$ in the abscissa $x = \frac{i_0}{N}$ is transformed to the left and to the right (and between them we have an horizontal part because of the added zero).

(ii) The addition of M zeros to X, Y, X' and Y' first shifts the decreasing part of $L_{X,Y}$ and $L_{X',Y'}$ over $\frac{M}{N}$ ("in the middle" we keep the constant maximal value of $L_{X,Y}$ and $L_{X',Y'}$) and then both new curves are homothetically transformed (over the abscissa) with a factor $\frac{N}{M+N}$ and finally both curves are multiplied by $\frac{\varphi(N)}{\varphi(N+1)}$. See Fig. 1 for an illustration (Lorenz curves are drawn smoothly for clarity). This proves (ii) and hence the theorem. ~

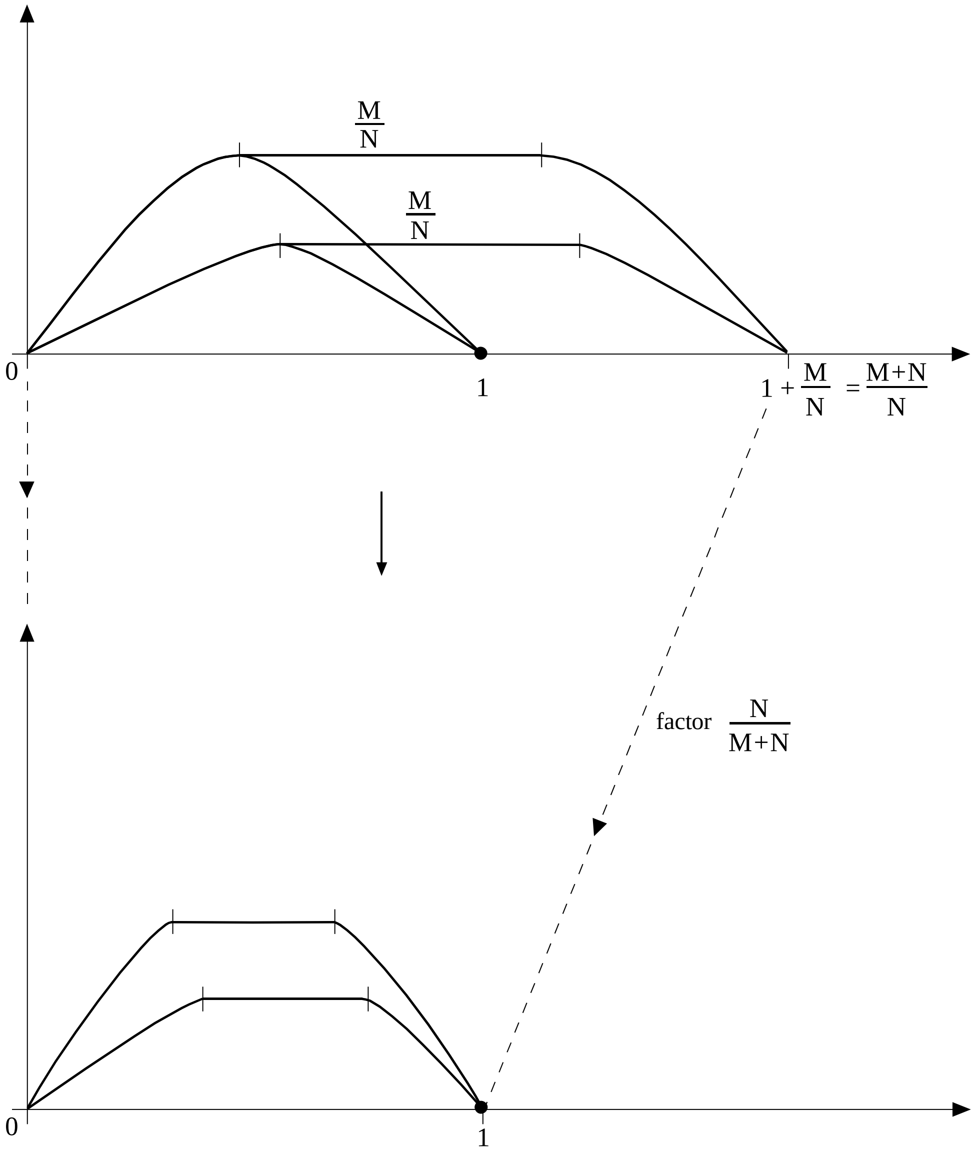

Fig. 1 Illustration of the proof of Theorem II.2 (ii).

Note that Theorem II.2 (ii) is also valid for Lorenz curves of Section I (i.e. where we do not divide by $\varphi(N)$). So the crucial part of Theorem II.2 (valid only if we divide by $\varphi(N)$) is (i).

Because of the above result we are now confident that our new approach is the right one. From the theory of Lorenz curves (see e.g. Egghe and Rousseau (2001), Egghe (2002)) we can now deduce several good measures of symmetric relative concentration. We present two of them: the Gini index and the coefficient of variation. First a Lemma (taken from Egghe and Rousseau (2004a) but adapted to Lorenz curves where we divide by $\varphi(N) = N$). Let

$$d_i = \frac{a_i - \alpha_i}{N} \quad (i = 1, \ldots, N).$$

Lemma II.3:

The expression $-\frac{1}{N}\sum_{i=1}^{N} i d_i$ is equal to the area under the Lorenz similarity curve.

Proof:

Define $c_i = \sum_{j=1}^{i} d_j$. The area under the Lorenz similarity curve is equal to:

$$\frac{1}{2N}\left(c_1 + \sum_{j=1}^{N-2}\left(c_j + c_{j+1}\right) + c_{N-1}\right)$$

$$= \frac{1}{2N}\Big(d_1 + \left(2d_1 + d_2\right) + \left(2d_1 + 2d_2 + d_3\right) +$$

$$\ldots + \left(2d_1 + \ldots + 2d_i + d_{i+1}\right) +$$

$$\ldots + \left(2d_1 + \ldots + 2d_{N-2} + d_{N-1}\right) + \left(d_1 + d_2 + \ldots + d_{N-1}\right)\Big)$$

$$= \frac{1}{N}\left(\left(N - 1\right)d_1 + \left(N - 2\right)d_2 + \ldots + d_{N-1}\right)$$

$$= \frac{1}{N}\sum_{j=1}^{N-1}(N-j)d_j - \frac{1}{N}N\sum_{j=1}^{N} d_j \qquad \text{by (3)}$$

$$= -\frac{1}{N}\sum_{i=1}^{N} id_i$$

The Gini similarity measure

One of the best known concentration measures is the Gini index (Myles, 1995). It is easy to derive a Gini similarity measure, denoted as G_s, from the Gini concentration index:

$$G_s(D) = G_s\{r,s\} = 1 + \frac{2}{N}\sum_{i=1}^{N} id_i = 1 + \frac{2}{N^2}\sum_{i=1}^{N} i(a_i - \alpha_i) \qquad (19)$$

where the d_i are ranked in decreasing order. This is, by the lemma, nothing but one minus twice the area under the Lorenz similarity curve. This normalizes the Gini similarity measure in such a way that all minimal Lorenz similarity curves correspond to a Gini-value of zero, and the equality line has a similarity value of one.

The coefficient of variation as a similarity measure

It is shown in Egghe and Rousseau (2001) that

$$V^2 = N\sum_{i=1}^{N} d_i^2 \qquad (20)$$

is a good measure of symmetric relative concentration. Here

$$d_i = \frac{a_i - \alpha_i}{\varphi(N)}.$$

This gives, for $\varphi(N) = N$, that

$$V^2 = \frac{1}{N}\sum_{i=1}^{N} (a_i - \alpha_i)^2 \qquad (21)$$

is a good measure of symmetric relative concentration. Consequently, all measures of the form

$$a + bV^2$$

with a, b constants, $b < 0$ are good measures of similarity in the sense of the A-J-R requirements.

III. Lorenz similarity theory for rank-order arrays and the correlation coefficient of Spearman

Rank-order arrays are arrays of the type

$$R = (r_1, ..., r_N), \tag{22}$$

where

$$\{r_1, ..., r_N\} = \{1, ..., N\} \tag{23}$$

Usually (but not always) rank-order arrays are derived from arrays $X = (x_1, ..., x_N)$, x_i ³ 0 $(i = 1, ..., N)$ as we studied in this article. The largest x_i value receives the rank 1, the second-largest the rank 2 and so on: the smallest value receives the rank N. If this is the case we will denote R by R_X.

Since all ranks are positive we can apply the Lorenz theory of Section II to two rank-order vectors R_X and R_Y just as we applied it to X and Y. However, the case of rank-order arrays is special since, if

$$R_X = (r_1, \ldots, r_N)$$

$$R_Y = (s_1, \ldots, s_N)$$

are two such arrays, we always have that

$$\sum_{j=1}^{N} r_j = \sum_{j=1}^{N} s_j = \frac{N(N+1)}{2} \tag{24}$$

, hence we always have (3) but for the not-normalized arrays R_X and R_Y (a fact that is not true for X and Y). It turns out that the Lorenz theory of Section II can be given also for the not-normalized vectors R_X and R_Y; this, in turn, since we do not divide by $\frac{N(N+1)}{2}$, gives us more possibilities of constructing good similarity measures (or measures of symmetric relative concentration), as we will see in the sequel.

So, given two rank-order arrays R_X and R_Y as above, we will compare them by constructing the Lorenz curve which linearly connects the consecutive points (0,0) and

$$\left(\frac{i}{N}, \sum_{j=1}^{i} \frac{r_j - s_j}{\varphi(N)}\right) \tag{25}$$

(cf. (15)) (where $(r_j - s_j)_{j=1}^{N}$ is decreasing). We will denote this Lorenz curve as

$$\mathbf{L}_{X,Y} = L_{R_X,R_Y} \tag{26}$$

as in the other case, $\mathbf{L}_{X,Y}$ concavely connects (0,0) with (1,0) and hence the classical Lorenz theory applies (cf. Egghe (2002)).

Now what is the effect of adding M zeros to the arrays X and Y ? It is clear that, since all $x_i \geq 0$ $(i = 1, \ldots, N)$ that, if $R_X = (r_1, \ldots, r_N)$,

$$R_{X\$M} = \left(r_1,\ldots,r_N,N+1,\ldots,N+M\right) \tag{27}$$

Note that the same array (27) is obtained when we add to the original array M values t such that $t \leq \min\{x_1,\ldots,x_N\}$.

We have the following Lemma.

Lemma III.1:

$$L_{R_X\$M,R_Y\$M} = L_{R_{X\$M},R_{Y\$M}} = \mathbf{L}_{X\$M,Y\$M} \tag{28}$$

Proof: The last equality follows by notation (26). Now $L_{R_{X\$M,Y\$M}}$, by (25), is constructed using the difference array $\left(r_1 - s_1,\ldots,r_N - s_N,\underbrace{0,\ldots,0}_{M}\right)$ (ordered decreasingly). Now

$$R_X\$M = \left(r_1,\ldots,r_N,\underbrace{0,\ldots,0}_{M}\right)$$

$$R_Y\$M = \left(s_1,\ldots,s_N,\underbrace{0,\ldots,0}_{M}\right)$$

So $L_{R_X\$M,R_Y\$M}$ is constructed using the same difference array

$$\left(r_1 - s_1,\ldots,r_N - s_N,\underbrace{0,\ldots,0}_{M}\right)$$

Hence both Lorenz curves are the same. ~

We explicitly prove now that Theorem II.2 is also valid for this Lorenz theory.

Theorem III.2: Let φ be a function such that

$$\frac{\varphi(N+1)}{\varphi(N)} ³ \frac{N+1}{N}. \qquad (29)$$

Then, for all arrays $X = (x_1,...,x_N)$, $Y = (y_1,...,y_N)$ with $x_i, y_i ³ 0$ $(i = 1,...,N)$ we have

(i)

$$\boldsymbol{L}_{X,Y} > \boldsymbol{L}_{X\$M,Y\$M} \qquad (30)$$

(ii)

$$\boldsymbol{L}_{X,Y} < \boldsymbol{L}_{X',Y'} \qquad (31)$$

$$Þ \; \boldsymbol{L}_{X\$M,Y\$M} < \boldsymbol{L}_{X'\$M,Y'\$M}$$

Proof: We first remark that Theorem II.2 is also valid if we do not divide by $\overset{N}{\underset{i=1}{å}} x_i$ (for X) respectively $\overset{N}{\underset{i=1}{å}} y_i$ (for Y), in case $\overset{N}{\underset{i=1}{å}} x_i = \overset{N}{\underset{i=1}{å}} y_i$. We call this "statement (*)".

(i)

$$\boldsymbol{L}_{X,Y} = L_{R_X,R_Y} \quad \text{(by (26))}$$

$$> L_{R_X\$M,R_Y\$M} \quad \text{(by(*))}$$

$$= L_{R_{X\$M},R_{Y\$M}} \quad \text{(by Lemma III.1)}$$

$$= \boldsymbol{L}_{X\$M,Y\$M} \quad \text{(by (26))}$$

(ii)

$$\boldsymbol{L}_{X,Y} < \boldsymbol{L}_{X',Y'}$$

implies, by (26))

$$L_{R_X,R_Y} < L_{R_{X'},R_{Y'}}$$

Hence, by (*)

$$L_{R_X\$M,R_Y\$M} < L_{R_{X'}\$M,R_{Y'}\$M}$$

Using Lemma III.1 we have

$$L_{R_{X\$M},R_{Y\$M}} < L_{R_{X'\$M,Y'\$M}}$$

and so, by (26),

$$\boldsymbol{L}_{X\$M,Y\$M} < \boldsymbol{L}_{X'\$M,Y'\$M}$$

completing the proof.

This shows the good properties of this rank-order Lorenz theory. We can apply the classical results of Lorenz concentration theory on the construction of good concentration measures (see e.g. Egghe (2002), Egghe and Rousseau (2001)). For the not-normalized rank-order arrays $R_X = (r_1,...,r_N)$, $R_Y = (s_1,...,s_N)$ we hence have that

$$V^2_{r,\varphi} = N\sum_{j=1}^{N}\left(\frac{r_j - s_j}{\varphi(N)}\right)^2 \qquad (32)$$

is a good measure of symmetric relative concentration. Hence any measure of the form (a,b constants, $b < 0$)

$$a + bV^2_{r,\varphi} \tag{33}$$

is a good similarity measure satisfying the Lorenz-orderings as well as the properties in Theorem III.2. Let us give two examples.

Example III.3: For $\varphi(N) = N$, (32) yields

$$V_r^2 = \frac{1}{N}\sum_{j=1}^{N}\left(r_j - s_j\right)^2 \tag{34}$$

Problem: using (33), what interesting (known ?) similarity measures can be derived from V_r^2 in (34) ?

Example III.4: Take

$$\varphi(N) = N\sqrt{N^2 - 1}\,. \tag{35}$$

We first have to check that (29) is valid:

$$\left(\frac{\varphi(N+1)}{\varphi(N)}\right)^2 = \frac{(N+1)^2\left((N+1)^2 - 1\right)}{N^2\left(N^2 - 1\right)}$$

$$= \frac{(N+1)^2(N+2)N}{N^2(N+1)(N-1)}$$

$$= \frac{(N+1)(N+2)}{N(N-1)}$$

$$> \left(\frac{N+1}{N}\right)^2.$$

Hence this function can be used in (32). We now have

$$V_{r,\varphi}^{2} = N\sum_{j=1}^{N}\left(\frac{r_j - s_j}{N\sqrt{N^2 - 1}}\right)^{2}$$

$$V_{r,\varphi}^{2} = \frac{1}{N\left(N^2 - 1\right)}\sum_{j=1}^{N}\left(r_j - s_j\right)^{2} \tag{36}$$

If we now apply (33) with $a = 1$, $b = -6$, then we have that

$$1 - 6V_{r,\varphi}^{2} = 1 - \frac{6}{N\left(N^2 - 1\right)}\sum_{j=1}^{N}\left(r_j - s_j\right)^{2}$$

$$= 1 - \frac{6}{N\left(N^2 - 1\right)}\sum_{j=1}^{N}\Delta_j^{2} \tag{37}$$

$\left(\Delta_j =: r_j - s_j, j = 1,..., N\right)$ is a good measure of similarity. But (37) is nothing else than the classical rank-order correlation coefficient of Spearman (see e.g. Liebetrau (1983)). So, unlike the correlation coefficient of Pearson, the one of Spearman satisfies the A-J-R requirements (2003) (see Theorem III.2), a fact that can also be checked directly. However, we now have the important information that the rank-order correlation coefficient of Spearman fits into the Lorenz theory of symmetric relative concentration (similarity).

Pearson's correlation coefficient does not fit into this Lorenz theory and does not satisfy the A-J-R requirements. Also Shoukry (2004) (and references therein) mention some problems with Pearson's coefficient. Hence, this classical measure should be used for what it was made: to calculate the degree of linearity of a cloud of points in the framework of the theory of linear regression. The debates on these remarks on the coefficient of Pearson (see Bensman (2004), Ahlgren, Jarneving and Rousseau (2004a,b)) could have been avoided if this had been realized earlier.

Remark III.5:

Note that the correlation coefficient of Spearman was found in this Lorenz theory using φ as in (35). If we had used the normalized Lorenz theory as in Section II we had to divide the difference vector by $\frac{N(N+1)}{2}$ and then to divide again by $\varphi(N)$ where (29) must be valid, i.e. $\varphi(N)$ must at least be of the order N. So in this theory we had (at least) to divide in the order N^3. This means that, in (32), we multiply $\Sigma\Delta_j^2$ by a number of the order $\frac{N}{N^6} = \frac{1}{N^5}$, hence missing many simpler measures, and especially the coefficient of Spearman (based on (36) where one multiplies $\Sigma\Delta_j^2$ by a number of the order $\frac{1}{N^3}$). Hence the extension of the Lorenz theory given in this section has evident applications.

IV. Problem and conclusions

IV.1 Problem

In this paper we modified the Lorenz theory of symmetric relative concentration by dividing the difference array by $\varphi(N)$, where (29) is valid. Hence we have

$$\varphi(N+1)^3 \ \frac{N+1}{N}\varphi(N)$$

$$^3 \ \frac{N+1}{N}\frac{N}{N-1}\varphi(N-1)$$

$$\ldots$$

$$\varphi(N+1)^3 \ (N+1)\varphi(1).$$

Hence $\lim_{N\to\infty} \varphi(N) = +\infty$. This means that, in the comparison of two arrays $X = (x_1,\dots,x_N)$, $Y = (y_1,\dots,y_N)$ and if we consider the k-replicas of X and Y:

$$k_X = \left(\underbrace{\underbrace{x_1,\dots,x_N},\dots,\underbrace{x_1,\dots,x_N}}_{k\ \text{times}}\right)$$

$$k_Y = \left(\underbrace{\underbrace{y_1,\dots,y_N},\dots,\underbrace{y_1,\dots,y_N}}_{k\ \text{times}}\right)$$

we have that

$$\lim_{k\to\infty} L_{k_X,k_Y} = 0 \tag{38}$$

no matter how different X and Y are. We do not know how to interpret (38) in practise. Note, however, that in our application in Section III on rank-order arrays, k-replicas do not occur.

IV.2 Conclusions

In this paper we showed that the Lorenz theory of symmetric relative concentration and similarity developed in Egghe and Rousseau (2001) can only be applied if the length N of arrays is kept constant. Indeed we showed that adding zeros to two arrays yields higher Lorenz curves, hence less similar arrays which is counterintuitive.

In Section II we modified this theory by dividing the Lorenz curves by $\varphi(N)$ such that

$$\frac{\varphi(N+1)}{\varphi(N)} \geq \frac{N+1}{N}$$

This condition then guarantees that, when zeros are added to two arrays, the Lorenz curves decrease, hence the similarity increases. This yields a machinery to produce good similarity measures that satisfy the A-J-R requirements.

In Section III the above theory is applied to rank-order arrays but we note that, since their coordinates always sum up to $\frac{N(N+1)}{2}$, no normalization in the construction of the Lorenz curve is needed. We show that also in this model the A-J-R requirements are valid. This yields more opportunities of constructing good similarity measures for rank-order arrays. We show that one of the good measures is the rank-order correlation coefficient of Spearman, an interesting result, certainly in connection with the fact that Pearson's correlation coefficient does not satisfy the A-J-R requirements.

References

P. Ahlgren, B. Jarneving and R. Rousseau (2003). Requirements for a cocitation similarity measure, with special reference to Pearson's correlation coefficient. Journal of the American Society for Information Science and Technology 54(6), 550-560, 2003.

P. Ahlgren, B. Jarneving and R. Rousseau (2004a). Author co-citation analysis and Pearson's *r* (letter to the editor). Journal of the American Society of Information Science and Technology 55, 843, 2004.

P. Ahlgren, B. Jarneving and R. Rousseau (2004b). Rejoinder: in defense of formal methods (letter to the editor). Journal of the American Society of Information Science and Technology 55, 936, 2004.

S.J. Bensman (2004). Pearson's *r* and author cocitation analysis: A commentary on the controversy (letter to the editor). Journal of the American Society of Information Science and Technology 55, 935, 2004.

L. Egghe (2002). Construction of concentration measures for general Lorenz curves using Riemann-Stieltjes integrals. Mathematical and Computer Modelling 35, 1149-1163, 2002.

L. Egghe and R. Rousseau (1990). Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science. Elsevier, Amsterdam, 1990.

L. Egghe and R. Rousseau (2001). Symmetric and asymmetric theory of relative concentration and applications. Scientometrics 52(2), 261-290, 2001.

L. Egghe and R. Rousseau (2004). Classical retrieval and overlap measures satisfy the requirements for rankings based on a Lorenz curve. Information Processing and Management, to appear, 2004.

A.M. Liebetrau (1983). Measures of Association. Quantitative Applications in the social Sciences 07-032. Sage Publications, Beverly Hills, USA, 1983.

G.D. Myles (1995). Public Economics. Cambridge University Press, Cambridge, 1995.

R. Rousseau (2001). Concentration and evenness measures as macro-level scientometric indicators. In: Keyan pingjia yu daxue pingjia (Research evaluation and university evaluation), (Wang, Z., Jiang, G., eds.). Beijing: Red Flag Publishing House, 72-89. (In Chinese, an English translation is available from the author).

M.M. Shoukry (2004). Measures of Interobserver Agreement. Chapman and Hall/CRC, Boca Raton (FL), USA, 2004.