Construction of concentration measures for General Lorenz curves using Riemann-Stieltjes integrals

CONSTRUCTION OF CONCENTRATION MEASURES FOR GENERAL LORENZ CURVES USING RIEMANN-STIELTJES INTEGRALS

by

L. Egghe LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
and
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium
leo.egghe@luc.ac.be

ABSTRACT

Lorenz curves were invented to model situations of inequality in real life and applied in econometrics (distribution of wealth or poverty), biometrics (distribution of species richness), informetrics (distribution of literature over their producers). Different types of Lorenz curves are hereby found in the literature and in each case a theory of good concentration measures is presented.

The present paper unifies these approaches by presenting one general model of concentration measure that applies to all these cases. Riemann-Stieltjes integrals are hereby

Acknowledgement : The author is indebted to Prof. Dr. R. Rousseau (U. Antwerp, Belgium) for formulating the problem and for interesting discussions on it.

Key words and phrases : Lorenz curve, concentration measure, Riemann-Stieltjes
MSC2000 classification : 26D10, 91F99, 62Pxx

[1] Permanent address

needed where the integrand is a convex function and the integrator a function that generalises the inverse of the derivative of the Lorenz function, in case this function is not everywhere differentiable.

Calling this general measure C we prove that, if we have two Lorenz functions f, g such that $f<g$, then $C(f)<C(g)$. This general proof contains the many partial results that are proved before in the literature in the respective special cases.

I. Introduction

The type of problem that we study in this paper can be introduced - in its most elementary form - as follows. Suppose we have a vector $X=(x_1,\ldots,x_N)$, where $x_i\in\mathbb{N}$, $\forall i=1,\ldots,N$ and where we order X decreasingly. The numbers x_i can represent incomes (econometrics), species richness (biometrics), number of articles written by author i (or published in journal i) (informetrics). The Lorenz curve of X is formed as follows) cf. Lorenz (1905) : connect (by lines) the points

$$\left(\frac{i}{N}, \sum_{j=1}^{i} a_j \right)_{i=1,\ldots,N} \tag{1}$$

and (0,0) and (1,1), where

$$a_i = \frac{x_i}{\sum_{j=1}^{N} x_j} . \tag{2}$$

Since the x_i are decreasing we obtain a concave polygonal increasing curve. Let f denote the corresponding function. If we also have a second vector $X'=(x'_1,\ldots,x'_N)$ and apply the same construction on X' as in (1) and (2), we obtain a second Lorenz curve with corresponding function g. Suppose $f\leq g$. It is clear that this is equivalent with

$$\sum_{j=1}^{i} a_j \leq \sum_{j=1}^{i} a'_j . \tag{3}$$

for all i=1,...,N. It is well-known (cf. Sen (1973), Allison (1978)) that $f \leq g$ and $f \neq g$ represents a situation in which X′ is more concentrated (more unequal) than X. In this case we say that any function C on these Lorenz curves such that $C(f) < C(g)$ is a good measure of concentration. There exist many examples of good measures of concentration, say the coefficient of variation V, where V is the positive root of

$$V^2 = \frac{\sigma^2}{\mu^2} = N \sum_{i=1}^{N} a_i^2 - 1 \tag{4}$$

where σ^2 and μ are the variance and the mean of X respectively, or Theil's measure (Theil (1967))

$$Th = \ln(N) + \sum_{i=1}^{N} a_i \ln a_i \ , \tag{5}$$

to mention just two of them.

This theory can be extended in several ways. First of all one can replace the uniform distribution $\left(\frac{1}{N},\ldots,\frac{1}{N}\right)$ (N times) as abscissae in (1) by a weight vector $W=(w_1,\ldots,w_N)$ ($w_i \geq 0$, $\forall i=1,\ldots,N$ and $\sum_{i=1}^{N} w_i = 1$). Here the weighted Lorenz curve is constructed as follows : *rearrange the vector* $X=(x_1,\ldots,x_N)$ *in such a way that*

$$\frac{x_1}{w_1} \geq \frac{x_2}{w_2} \geq \ldots \geq \frac{x_N}{w_N} \tag{6}$$

and connect the points

$$\left(\sum_{j=1}^{i} w_j, \sum_{j=1}^{i} a_j \right) \tag{7}$$

and (0,0) and (1,1). Also in this case one obtains a (weighted) Lorenz curve which is concavely increasing and polygonal. Let us again denote its corresponding function by f. If we have two such functions f,g then we say that C is a good measure of (weighted) concentration if $f \leq g$, $f \neq g \Rightarrow C(f) < C(g)$. Applications of this are given at the end of the paper.

We also refer to Rousseau (1992) and Egghe and Rousseau (2000), where, in the latter reference, a proof is given that (e.g.) the measures

$$V_w^2 = \sum_{i=1}^{N} \frac{a_i^2}{w_i} - 1 \tag{8}$$

and

$$Th_w = \sum_{i=1}^{N} a_i \ln\left(\frac{a_i}{w_i}\right) \tag{9}$$

are good measures in this case.

The functions f and g mentioned so far all have the property that $f(0)=g(0)=0$ and $f(1)=g(1)=1$. One can also construct the Lorenz curve of the difference of two vectors X and X′. Here we order the numbers a_i-a_i' decreasingly (note that these numbers can be negative). If we form the Lorenz curve (weighted or not) based on the numbers $(a_i-a_i')_{i=1,\ldots,N}$ we obtain a function f that still is 0 in 0 but for which we have $f(1)=0$. This has applications as we will see at the end of the paper.

Another generalization is obtained by studying continuous processes (as e.g. in econometrics) hereby using functions f,g which are continuously differentiable (so-called C^1 functions).

Definition I.1 : A common generalization of all the above mentioned cases is given by functions f on [0,1] such that $f(0)=0$, which are concave and piecewise C^1, i.e. $\exists\ x_1,\ldots,x_N \in\ ]0,1[$ such that $x_1<x_2<\ldots<x_N$ and such that

$$f|_{[x_k,x_{k+1}]}$$

is continuously differentiable, $\forall k=0,\ldots,N$ where we put $x_0=0$, $x_{N+1}=1$.

For such general functions f,g such that $f(0)=g(0)=0$ and $f(1)=g(1)$, we will study the problem of finding measures C such that $f \leq g$, $f \neq g \Rightarrow C(f) < C(g)$. A general solution will be given, using continuous convex functions φ as the integrand and an extension of f'^{-1} (which does not always exist) as the integrator of a Riemann-Stieltjes integral over the interval $[f'(1), f'(0)]$.

For the many results on Riemann-Stieltjes integration we refer the reader to the classical book Apostol (1957) or the more recent Douglas (1996). The following inequality of Hardy, Littlewood and Pólya (1929) will be used (in its full - less known - generality using Riemann-Stieltjes integrals) - see also Brunk (1956) :

Theorem I.2 (Hardy, Littlewood and Polya)
Let $x:t \to x(t)$ on the interval [a,b] and $y:t \to y(t)$ on the interval [c,d] be functions of bounded variation. Then we have

$$\int_a^b \varphi(t)dx(t) \leq \int_c^d \varphi(t)dy(t) \tag{10}$$

for all φ : continuous convex function on a closed interval containing [a,b] and [c,d] iff

(i) $$\int_a^b dx(t) = \int_c^d dy(t) \tag{11}$$

(ii) $$\int_a^b t\,dx(t) = \int_c^d t\,dy(t) \tag{12}$$

(iii) $$\int_a^b [t-u]^+ dx(t) \leq \int_c^d [t-u]^+ dy(t) \tag{13}$$

$\forall u \in \mathbb{R}$. Here we denote

$$[s]^+ = \max(s,0) \tag{14}$$

$\forall s \in \mathbb{R}$.

In the next section we will present the general formula for a good concentration measure C which is defined on all functions f as given in definition I.1. We also present some general properties. In section II we present then the main result : $f \leq g$ and $f \neq g \Rightarrow C(f) < C(g)$. Also the direct proof for the special case of weighted polygonal Lorenz curves - which is far more elementary - is given, for those who want a direct proof in the discrete case.

Section IV closes the paper by presenting applications.

II. General form of a concentration measure on a piecewise smooth (i.e. C^1) function.

II.1 Definitions

Let f be a piecewise smooth function on [0,1] as described in definition I.1. In order to deal with the discrete and continuous cases as described in the introduction and in view of (10) in theorem I.2, we are lead to the following definitions.

Definition II.1.1.

We denote by D_f the following function on [0,1] :

$$D_f(x) = f'(x),\ x \in \bigcup_{k=1}^{N-1}]x_k, x_{k+1}[\cup [0, x_1[\cup]x_N, 1]$$

$$D_f(x) = \lim_{\substack{x \to x_k \\ < }} f'(x) = f'(x_k-). \tag{15}$$

Since f is piecewise C^1, D_f exists. Note that D_f is not necessarily injective. We therefore define (a kind of generalized inverse of D_f).

Definition II.1.2.

Denote by $R(D_f)$ the range of D_f. Define

$$D_f^I(t) = \sup\{x \| f'(x)=t\} \tag{16}$$

if $t \in R(D_f)$. If $t \notin R(D_f)$, define

$$D_f^I(t) = D_f^I(t_0) \tag{17}$$

with $t_0 > t$, $t_0 \in R(D_f)$ and t_0 is the smallest number with this property. By definition II.1.1, D_f^I exists on every $t \in [b,a]$ such that $a = D_f(0) = f'(0)$ and $b = D_f(1) = f'(1) < f'(0)$ since f is concave (we exclude the case $f'(0) = f'(1)$ in which case f is the first bissectrice of the square $[0,1] \times [0,1]$: $f(x) = x$).

The function D_f^I is nothing else than the cumulative distribution function of the given situation (on which one builds the Lorenz curve). Hence, here our approach is opposite to the ones adopted in Gastwirth (1971, 1972) or Thistle (1989) : they start from the cumulative distribution function and build the Lorenz curve ; we start from the Lorenz curve and construct the cumulative distribution function. Both approaches are equivalent but the latter one is more logical in our framework where the task is, given two Lorenz curves, based on functions $f < g$, to construct good concentration measures C such that $C(f) < C(g)$. C will be constructed using D_f^I (see (20) further on).

Examples II.1.3

(1) The case of the weighted polygonal Lorenz curve as described by (7) :

$$D_f^I(t) = \sum_{j=1}^{i} w_j \tag{18}$$

for

$$t \in \left[\frac{a_{i+1}}{w_{i+1}} , \frac{a_i}{w_i} \right] \tag{19}$$

$(i=1,\ldots,N-1)$ and $D_f^I(t) = 1$ for $t = \frac{a_N}{w_N}$. Note that $\frac{a_{i+1}}{w_{i+1}} < \frac{a_i}{w_i}$ since f is concave.

(2) The case that f is C^1. Hence $D_f = f'$ attains all values between $f'(1)$ and $f'(0)$ and is injective (again we exclude $f(x)=x$). Hence, by definition : $D_f^I = f'^{-1}$.

Definition II.1.4

For f as in definition I.1 and with D_f^I as in definition II.1.2 we define, for every continuous convex function φ on an interval that contains $[f'(1), f'(0)]$:

$$C(f) = -\int_{f'(1)}^{f'(0)} \varphi(t)\ d[D_f^I(t)]$$

$$C(f) = \int_{f'(0)}^{f'(1)} \varphi(t)\ d[D_f^I(t)] \qquad (20)$$

in the Riemann-Stieltjes sense.

Note that D_f^I decreases and hence, if $\varphi \geq 0$, $C(f) \geq 0$ by the properties of Riemann-Stieltjes integrals (RSI). For ease of notation we write for f and g, piecewise C^1 :

$$a = f'(0) > b = f'(1),\ x = D_f^I$$

$$c = g'(0) > d = g'(1),\ y = D_g^I$$

II.2 Concrete expressions of C(f) in classical cases.

II.2.1 Weighted polygonal Lorenz curves

Since the integrator is a step function with jumps w_i (see example II.1.3 (1)), the RSI reduces to

$$C(f) = \int_a^b \varphi(t)dx(t) = \sum_{i=1}^{N} \varphi\left(\frac{a_i}{w_i}\right) w_i \qquad (21)$$

This is a convex generalization of (8) and (9) (and hence also of (4) and (5)), where we use $\varphi(t)=t^2$ and $\varphi(t)=t \ln t$ respectively for V_W^2+1 and Th_w. Alternatively one can use $\varphi(t)=t^2-1$ for V_W^2.

II.2.2 C^1 functions f.

Using example II.1.3 (2) and substitution we have

$$C(f) = \int_a^b \varphi(t)dx(t)$$

$$= \int_{f'(0)}^{f'(1)} \varphi(t)d[f'^{-1}(t)]$$

$$= \int_0^1 \varphi(f'(y))dy \qquad (22)$$

Analogous to the discrete case we hence define

$$V_f^2 = \int_0^1 f'^2(y)dy - 1 \qquad (23)$$

(hence we obtain V_f^2+1 by using $\varphi(t)=t^2$ in (22)) and

$$Th_f = \int_0^1 f'(y)\ln(f'(y))dy \qquad (24)$$

(hence we obtain Th_f by using $\varphi(t)=t \ln t$ in (22)), showing that these measures can be treated by the general formula (22) or (20).

With these expressions we feel that we are heading in the right direction in the construction of good concentration measures for piecewise C^1 functions. The general proof that indeed $f \le g$, $f \ne g \Rightarrow C(f) < C(g)$ for C as in (20) will be given in section III. We first continue section II by presenting an alternative form for formula (20) in the two special cases above.

II.3 Other forms of C(f)

Proposition II.3.1. In case f is the function of a polygonal curve or in case f is C^1 we have the following alternative formula for C(f) :

$$C(f) = \int_0^1 \frac{\varphi(D_f)}{D_f}\, df \tag{25}$$

$$C(f) = \frac{\varphi(f'(1))}{f'(1)} - \int_0^1 fd\left[\frac{\varphi(D_f)}{D_f}\right] \tag{26}$$

Proof : That (26) equals (25) follows from partial integration on the RSI (25) and using that f(0)=0.

(i) f is the function of a polygonal (i.e. weighted Lorenz) curve.

We use (26) and note that $\frac{\varphi(D_f)}{D_f}$ is a stepfunction. Hence (26) can be evaluated as

$$\frac{\varphi(f'(1))}{f'(1)} = \frac{\varphi\left(\frac{a_N}{w_N}\right)}{\frac{a_N}{w_N}} \tag{27}$$

and

$$\int_0^1 fd\left[\frac{\varphi(D_f)}{D_f}\right] = \sum_{i=1}^{N-1}\left(\sum_{j=1}^{i} a_j\right)\left[\frac{\varphi\left(\frac{a_{i+1}}{w_{i+1}}\right)}{\frac{a_{i+1}}{w_{i+1}}} - \frac{\varphi\left(\frac{a_i}{w_i}\right)}{\frac{a_i}{w_i}}\right] \tag{28}$$

(27) and (28) in (26) yield

$$\frac{\varphi\left(\frac{a_N}{w_N}\right)}{\frac{a_N}{w_N}} - \sum_{i=1}^{N-1}\left(\sum_{j=1}^{i} a_j\right)\frac{\varphi\left(\frac{a_{i+1}}{w_{i+1}}\right)}{\frac{a_{i+1}}{w_{i+1}}} + \sum_{i=1}^{N-1}\left(\sum_{j=1}^{i} a_j\right)\frac{\varphi\left(\frac{a_i}{w_i}\right)}{\frac{a_i}{w_i}}$$

$$= \frac{\varphi\left(\frac{a_N}{w_N}\right)}{\frac{a_N}{w_N}} - \sum_{i=2}^{N}\left(\sum_{j=1}^{i-1} a_j\right)\frac{\varphi\left(\frac{a_i}{w_i}\right)}{\frac{a_i}{w_i}} + \sum_{i=1}^{N-1}\left(\sum_{j=1}^{i} a_j\right)\frac{\varphi\left(\frac{a_i}{w_i}\right)}{\frac{a_i}{w_i}}$$

$$= \left[1 - \sum_{j=1}^{N-1} a_j\right]\frac{\varphi\left(\frac{a_N}{w_N}\right)}{\frac{a_N}{w_N}} + \sum_{i=2}^{N-1} a_i \frac{\varphi\left(\frac{a_i}{w_i}\right)}{\frac{a_i}{w_i}} + a_1 \frac{\varphi\left(\frac{a_1}{w_1}\right)}{\frac{a_1}{w_1}}$$

$$= C(f),$$

by (21) and the fact that $\sum_{i=1}^{N} a_i = 1$.

(ii) f is C^1

Now we use (25), yielding (by II.1.3 (2))

$$\int_0^1 \frac{\varphi(f')}{f'}\, df$$

$$= \int_0^1 \varphi(f'(y))\, dy$$

since f' is continuous. Hence this is C(f) by (22). □

III. General proof that C (formula (20)) is a good concentration measure.

Theorem III.1. Let f and g be, as in definition I.1, piecewise smooth functions such that $f(0)=g(0)=0$, $f(1)=g(1)$, f,g : concave, $f \le g$ and $f \neq g$. Then we have that

$$C(f) < C(g) ,$$

where C is given by (20).

Proof :

1. We first assume that f and g are C^1 (i.e. continuously differentiable).

We will check (11), (12), (13) for $x = D_f^I$, $y = D_g^I$, $a = f'(0)$, $b = f'(1)$, $c = g'(0)$, $d = g'(1)$. We have

$$\int_c^d dy(t) - \int_a^b dx(t)$$

$$= g'^{-1}(d) - g'^{-1}(c) - (f'^{-1}(b) - f'^{-1}(a))$$

$$= 0, \qquad (29)$$

since $D_g^I = g'^{-1}$, $D_f^I = f'^{-1}$ (cf. II.1.3(2))

Next :

$$\int_c^d t\ dy(t) - \int_a^b t\ dx(t)$$

$$= \int_0^1 g'(s)ds - \int_0^1 f'(s)ds$$

(substitute $t=g'(s)$ and $t=f'(s)$ respectively)

$$= g(1) - g(0) - (f(1) - f(0))$$

$$= 0 \qquad (30)$$

It remains to show the non-trivial inequality (13). Now

$$\int_a^b [t-u]^+ dx(t)$$

$$= [b-u]^+ f'^{-1}(b) - [a-u]^+ f'^{-1}(a) - \int_a^b f'^{-1}(t)\ d[t-u]^+ \qquad (31)$$

, using partial integration for RSI. The same is true for g. We have several cases for $u \in \mathbb{R}$.

(i) $u \in [b,a] \subset [d,c]$

Note that the last inclusion is always valid since $g'(1) \le f'(1) < f'(0) \le g'(0)$. Now $[b-u]^+ = 0$ and $[t-u]^+ \begin{cases} = 0 & , u \ge t \\ = t-u, & u \le t \end{cases}$

Hence

$$\int_a^b [t-u]^+ dx(t) = -\int_a^u f'^{-1}(t)dt \qquad (32)$$

Since also $u \in [d,c]$, we have the same for g :

$$\int_c^d [t-u]^+ dy(t) = -\int_c^u g'^{-1}(t)dt \qquad (33)$$

(13) will be proved if we can show that

$$-\int_a^u f'^{-1}(t)dt \le -\int_c^u g'^{-1}(t)dt$$

or, using substitution and partial integration,

$$\int_{f'^{-1}(u)}^{0} sdf'(s) \le -\int_{g'^{-1}(u)}^{0} sdg'(s)$$

$$-uf'^{-1}(u) + f(f'^{-1}(u)) \le -ug'^{-1}(u) + g(g'^{-1}(u)) \qquad (34)$$

To simplify the notation we put $x_1 = f'^{-1}(u)$, $x_2 = g'^{-1}(u)$. Then (34) reduces to :

$$g'(x_2)(x_2-x_1) \le g(x_2)-f(x_1) \qquad (35)$$

(a) $x_2 > x_1$. Now

$$g(x_2) - f(x_1)$$

$$\ge g(x_2) - g(x_1)$$

$$= g'(\alpha)(x_2-x_1) \ge g'(x_2)(x_2-x_1),$$

$\exists \alpha \in]x_1, x_2[$ by the mean value theorem and by the fact that $\alpha < x_2$ implies $g'(\alpha) > g'(x_2)$ (g concave).

(b) $x_1 < x_2$. Now

$$g(x_2) - f(x_1)$$

$$\ge g(x_2) - g(x_1)$$

$$= g'(\alpha)(x_2-x_1)$$

$\geq g'(x_2)(x_2-x_1)$

$\exists\alpha\in]x_2,x_1[$ by the mean value theorem and by the fact that $\alpha>x_2$ implies $g'(\alpha)<g'(x_2)$ and that $x_2-x_1<0$.

This proves (35) in case (i)

(ii) $d\leq u<b$.

Now we have

$$\int_a^b [t-u]^+ \, dx(t)$$

$$= b-u-\int_a^b f'^{-1}(t)dt$$

since $t\in[b,a] \Rightarrow t>u$. Since $u\in[d,c]$ we have (cf. (33)) :

$$\int_c^d [t-u]^+ \, dy(t)$$

$$= -\int_c^d g'^{-1}(t)dt$$

We hence have to show

$$b-u-\int_a^b f'^{-1}(t)dt \leq -\int_c^d g'^{-1}(t)dt \qquad (36)$$

As in (i) we can deduce the condition :

$$u(g'^{-1}(u) - f'^{-1}(b)) \leq g(g'^{-1}(u)) - f(f'^{-1}(b))$$

or, denoting $x_2=g'^{-1}(u)$ and the fact that $f'^{-1}(b)=1$:

$$g(x_2) - f(1) \geq g'(x_2)(x_2-1) \qquad (37)$$

(a) $x_2 > 1$. $g(x_2) - f(1)$

$\geq g(x_2) - g(1)$ (even $=$)

$= g'(\alpha)(x_2-1)$

$> g'(x_2)(x_2-1)$

$\exists \alpha \in]1, x_2[$ by the mean value theorem and $\alpha < x_2 \Rightarrow g'(\alpha) > g'(x_2)$

(b) $x_2 < 1$. $g(x_2) - f(1)$

$\geq g(x_2) - g(1)$

$= g'(\alpha)(x_2-1)$

$\geq g'(x_2)(x_2-1)$

$\exists \alpha \in]x_2, 1[$ by the mean value theorem and $\alpha > x_2 \Rightarrow g'(\alpha) < g'(x_2)$ and $x_2-1 < 0$. This proves (13) in case (ii).

<u>(iii) $u < d < b$</u>

Now we have the condition

$$b-u+(-b+f(1)) \leq d-u+(-d+g(1))$$

which is trivial. Hence (13) is checked.

<u>(iv) $a \leq u < c$</u>

Now $\forall t \in [b,a]$: $t \leq u$, hence $[t-u]^+ = 0$ on $[b,a]$. Since $\int_c^d [t-u]^+ dy(t) \geq 0$, (13) is checked.

(v) $u \geq c$

Now both sides in (13) are 0 and hence (13) is checked.

Since (13) is now checked $\forall u \in \mathbb{R}$ we have, by (10)

$$\int_a^b \varphi(t)dx(t) \leq \int_c^d \varphi(t)dy(t)$$

for every continuous convex function φ. It is clear that in this case, if $f \leq g$, $f \neq g$ that the above inequality is strict.

2. General case : f and g are piecewise C^1.

Let $\varepsilon_n > 0$, $\forall n \in \mathbb{N}$ be such that $\lim_{n\to\infty} \varepsilon_n = 0$. Since f is piecewise C^1, we have a finite number of points $x_1, \ldots x_N$ as expressed in definition I.1, where possible discontinuities of the tangent line of f occur. Define a function f_n as follows : $f_n = f$ outside $\bigcup_{k=1}^{N}]x_k - \varepsilon_N, x_k + \varepsilon_N[$, $f_n < f$ on this set and such that f_n is C^1, i.e. f_n is tangent to f in the points $x_k - \varepsilon_n$, $x_k + \varepsilon_n$, $k = 1, \ldots, N$ - see Fig. 1 as an illustration in one point x_1.

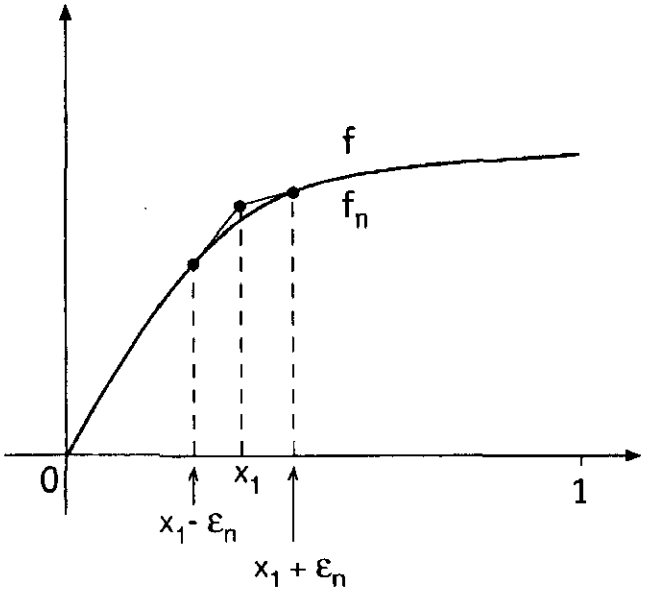

Fig. 1 : C^1 approximation of a piecewise C^1 function.

We do the same for g and, since $f \le g$ we can construct g_n such that $f_n \le g_n$, $\forall n \in \mathbb{N}$.

Note that $D_f^I(t) = D_{f_n}^I(t)$, $\forall t \in [0,1] \setminus \bigcup_{k=1}^{N}]f'(x_k+\varepsilon_n), f'(x_k-\varepsilon_n)[$.

Further

$$]f'(x_k+\varepsilon_n), f'(x_k-\varepsilon_n)[\; = \;]f_n'(x_k+\varepsilon_n), f_n'(x_k-\varepsilon_n)[$$

$\forall k = 1,\ldots,N$ and $\forall n \in \mathbb{N}$. Hence on such an interval we have

$$f_n'^{-1}(t) \; = \; D_{f_n}^I(t) \in]x_k-\varepsilon_n, x_k+\varepsilon_n[$$

and

$$D_f^I(t) \in]x_k-\varepsilon_n, x_k+\varepsilon_n[.$$

Hence, $\forall t \in [b,a]$ $= [f'(1), f'(0)]$

$= [f_n'(1), f_n'(0)]$

$\forall n \in \mathbb{N}$, we have

$$\left| D_f^I(t) \; - \; D_{f_n}^I(t) \right| \le 2\varepsilon_n$$

Hence $\lim_{n \to \infty} D_{f_n}^I = D_f^I$, uniformly on [b,a]. Hence, by the property of RSI on uniform convergence and using partial integration,

$$\int_a^b \varphi(t) d\left[D_{f_n}^I(t) \right]$$

$$= \left[\varphi(t) D_{f_n}^I(t) \right]_a^b \; - \; \int_a^b D_{f_n}^I(t) d(\varphi(t))$$

converges to

$$\varphi(t)D_f^I(t)\Big]_a^b - \int_a^b D_f^I(t)d(\varphi(t))$$

$$= \int_a^b \varphi(t)d\left[D_f^I(t)\right]$$

On $f_n \leq g_n$ we apply the C^1-case, hence

$$\int_a^b \varphi(t)d\left[D_{f_n}^I(t)\right] \leq \int_c^d \varphi(t)d\left[D_{g_n}^I(t)\right]$$

for all $n \in \mathbb{N}$. We hence can conclude that

$$\int_a^b \varphi(t)d\left[D_f^I(t)\right] \leq \int_c^d \varphi(t)d\left[D_g^I(t)\right] \tag{38}$$

with strict inequality if $f \neq g$. This proves theorem III.1. □

Notes III.2 :

1. As noted already, (38) reduces to the following in case of vectors $X=(x_1,\ldots,x_N)$, weighted by $W=(w_1,\ldots,w_N)$ and $X'=(x'_1,\ldots,x'_M)$, weighted by $W'=(w'_1,\ldots,w'_M)$, such that the weighted Lorenz curve of X' is above the one of X : then we have

$$\sum_{i=1}^{N} \varphi\left(\frac{a_i}{w_i}\right) w_i < \sum_{j=1}^{M} \varphi\left(\frac{a_j'}{w_j'}\right) w_j' \tag{39}$$

, where $a_i = \dfrac{x_i}{\sum_{k=1}^{N} x_k}$ and $a_j' = \dfrac{x_j'}{\sum_{\ell=1}^{M} x_\ell'}$.

In appendix A we present a direct proof of this, which is more elementary than the one of theorem III.1.

2. Also as noted already, (38) reduces to

$$\int_0^1 \varphi(f'(y))dy < \int_0^1 \varphi(g'(y))dy \tag{40}$$

in the C^1-case.

3. In the general case that f and g are piecewise C^1, (38) can be evaluated as follows (explained on f) :

$$D_f^I(t) = x_k$$

$\forall t \in [f'(x_k+), f'(x_k-)]$, $k=1,\ldots,N$. On the other values of $t \in [b,a]$ we have

$$D_f^I(t) = f'^{-1}(t)$$

which exists there. So D_f^I is a piecewise continuous function so that $\int_a^b \varphi(t)d[D_f^I(t)]$ for continuous φ can be evaluated using the following result on RSI, derivable from the theorem 9-9 (p. 198-199) in Apostol (1957) (we omit the simple proof) :

Proposition III.3 : Let h be RS-integrable w.r.t. α on the interval [a,b] and that α is continuous on [a,b] except in the points $c_1 < \ldots < c_n$ in]a,b[. Then

$$\int_a^b g d\alpha = \sum_{i=1}^{N+1} \int_{c_{i-1}}^{c_i} g d\alpha_i + \sum_{i=1}^{N} g(c_i)[\alpha(c_i+) - \alpha(c_i-)] \tag{41}$$

where $c_0=a$, $c_{N+1}=b$ and α_1 is defined on $[a,c_1]$ as

$$\alpha_1 = \alpha \text{ on } [a,c_1[$$
$$\alpha_1(c_1) = \alpha(c_1-),$$

α_i $(i=2,\ldots,N)$ on $[c_{i-1}, c_i]$ as

$\alpha_i = \alpha$ on $]c_{i-1}, c_i[$

$\alpha_i(c_{i-1}) = \alpha(c_{i-1}+)$

$\alpha_i(c_i) = \alpha(c_i-)$,

α_{N+1} on $[c_N, b]$ as

$\alpha_{N+1} = \alpha$ on $]c_N, b]$

$\alpha_{N+1}(c_N) = \alpha(c_N+)$.

IV. Applications.

This section reviews some applications of concentration theory, known so far in the literature, where separate proofs have been given for the fact that the used concentration measures (all special cases of the ones given here) are good. Hence our present theory comprises all these theories.

IV.1 Discrete non-weighted Lorenz case.

This is the "historical" case and it is impossible to mention all applications. They are found in econometrics, biometrics, informetrics and sociometrics or virtually in any situation where inequality, elitarism, ... occurs. We will suffice by giving some key references : Rothschild and Stiglitz (1973), Kanbur (1984) in econometrics, Allison (1978) in sociometrics, Patil and Taillie (1982) in biometrics, Egghe and Rousseau (1990a,b) in informetrics.

IV.2 C^1 case.

This case is often used by econometricians in order to model experimental data. The main references here are Gastwirth (1971, 1972) and Atkinson (1970). In sociometrics we can refer to Lambert (1985). Also in information science this model is often used, mainly to

describe the entropy measure $H=-Th_f$ (formula (24)). Here, entropy measures diversity, rather than concentration. We refer to Cover and Thomas (1991) and Ihara (1993) for a thorough study of the entropy (hence Th) measure. It is remarkable that in econometrics, one almost always uses the Gini index (which is twice the area between the Lorenz curve and the first bissectrice) or Theil's measure (formula (24)) rather than using the variation coefficient V or V^2 (formula (23)).

IV.3 Discrete weighted Lorenz case.

An application of this in econometrics is found in Kanbur (1984) p.428-430, were one works with grouped data : if one only has income data of a whole group or country (rather than individual income data) one has to weight each group or country, leading to weighted Lorenz curves.

In Rousseau (1992) one uses weighted Lorenz curves in order to be able to deal with N-dependence, where N denotes the total number of items.

In Egghe and Rousseau (2000) one applies weighted Lorenz curves (and corresponding concentration measures) to describe asymmetric relative concentration. Relative concentration intends to compare two vectors $X=(x_1,..,x_N)$ and $Y=(y_1,...,y_N)$. A relative concentration measure gives zero if $X=Y$ and the highest values for very different vectors X,Y (such as $X \perp Y$ for instance). Case IV.1 is contained here by taking for Y a constant vector. In general asymmetric relative concentration compares many vectors X_j with one "reference" vector Y (used as weight vector W in this article), such as in the case of the distribution of several topics over fixed sites (e.g. in distributed documentary systems, i.e. a set of autonomous distinct document collections (sites)). These sites form the fixed reference frame with which comparisons are made. This goes as follows : for an arbitrary topic one checks how many documents on this topic exist in the different sites. Then one compares these relative scores with the relative sizes of these sites. Using the model of weighted Lorenz curves, Egghe and Rousseau were able to improve an earlier solution of Viles and French (1999) for the problem of measuring content locality.

In Rousseau (2000) weighted Lorenz curves are applied as follows : suppose one wants to compare two vectors $X=(x_1,\ldots,x_N)$ and $Y=(y_1,\ldots,y_n)$. One can then compare X w.r.t. Y, where Y is considered as the reference vector, leading to one weighted Lorenz curve, or, one can compare Y w.r.t. X, where now X is considered as the reference vector. Also here we have a weighted Lorenz curve. One then takes the minimum of both curves, leading to a kind of "symmetric" comparison of X and Y. It was this paper that presented the idea of studying the problem as done in the present paper.

Another way to compare X and Y in a symmetric way is given in subsection IV.4.

IV.4 An application of the case in which f(1)=0.

So far all our applications were dealing with functions f such that f(0)=0, f(1)=1. In Egghe and Rousseau (2000) an application is given to Lorenz curves such that f(0)=f(1)=0. It is an application on symmetric relative concentration theory. Here, as already described in IV.3, one wants to compare two vectors $X=(x_1,\ldots,x_N)$ and $Y=(y_1,\ldots,y_N)$ but now we want X and Y to play an equivalent role : there is no reference vector and the comparison of X with Y should be the same as the comparison of Y with X, hence the name symmetric relative concentration.

If $A_X=(a_1,\ldots,a_N)$ with a_i as in (2) and if $A_Y=(b_1,\ldots,b_N)$ is defined in a similar way for Y, we now take A_X-A_Y and construct the Lorenz curve based on this vector (as in (1) but replacing a_j by a_j-b_j). Hence we obain a Lorenz curve that ends in (1,0) rather than (1,1). In Egghe and Rousseau (2000) one develops the theory of symmetric relative concentration hereby also defining good concentration measures for it. An example is the measure

$$V_r^2 = N\sum_{i=1}^{N}(a_i-b_i)^2 \qquad (42)$$

which is contained in our model (20) and even in (21), where we take $W=\left(\frac{1}{N},\ldots,\frac{1}{N}\right)$ (N coordinates), replace a_i by a_i-b_i and use $\varphi(x)=x^2$. It is contained there since in our general theorem III.1 we only required f(0)=g(0)=0, f(1)=g(1) (any value allowed here).

A concrete application of this is also given in Egghe and Rousseau (2000) were a good measure of symmetric relative concentration is used to compare two documents or a query and a document, using the vector model for queries and documents, hereby improving the classical "cosine" formula for comparison. (see Egghe (1990) or Salton and Mc Gill (1987)).

In Egghe and Rousseau (2001) the theory of symmetric relative concentration is extended to vectors $X=(x_1,\ldots,x_N)$, $Y=(y_1,\ldots,y_M)$, where possibly $M\neq N$. Also in this case the general theory as developed in this paper is applied. This theory is then applied to truncation of bibliographies where vectors of the type $(x_1,\ldots,x_i)$ and $(x_1,\ldots,x_i,x_{i+1})$, $i=1,\ldots,N$ are compared (the x_j s represent the number of items in source j).

IV.5 A new application.

As far as we are aware of it, we think the following application is new : in IV.4 one compares two vectors, unweighted and a good measure of symmetric relative concentration then compares two such situations. These relative differences might be related to a reference vector, which can be variable in time (as e.g. the relative sizes of sites in a documentary system). So here we want to compare a situation X-Y w.r.t. W (reference vector) at $t=t_1$ with the analogue situation X'-Y' w.r.t. W' (reference vector) at $t=t_2$. Also this is included in (21). We remind the reader that in appendix A a separate proof of the fact that C in (20) is a good measure is given. It is simpler than the general proof of theorem III.1 and hence the reader who is only interested in the discrete case, can avoid the calculations with RSI !

Note : In Thistle (1989) one deals with general Lorenz curves but only the generalized Gini index is studied there.

Appendix A

Proof of theorem III.1 in the case of the discrete weighted Lorenz curve. We give a formulation in this case.

Theorem :

Let $X=(x_1,\ldots,x_N)$, weighted by $W=(w_1,\ldots,w_n)$ and $X'=(x'_1,\ldots,x'_M)$, weighted by $W'=(w'_1,\ldots,w'_M)$ such that the weighted Lorenz curve of X' is above (and not equal to) the one of X. Then we have

$$\sum_{i=1}^{N} \varphi\left(\frac{a_i}{w_i}\right) w_i < \sum_{j=1}^{M} \varphi\left(\frac{a_j'}{w_j'}\right) w_j' \tag{A1}$$

for all convex functions φ and where

$$a_i = \frac{x_i}{\sum_{k=1}^{N} x_k} \tag{A2}$$

$$a_j' = \frac{x_j'}{\sum_{\ell=1}^{M} x_\ell'} \tag{A3}$$

$i=1,\ldots,N$, $j=1,\ldots,M$.

Proof :

1. It suffices to prove (A1) for $W=W'$. Indeed, each break point of the Lorenz curve f of X (i.e. jump of f') can be considered as one of the Lorenz curve g of X' and vice versa. In this way one subdivises the weights into several smaller ones. This has no

influence on the left or right hand side of (A1). Let us prove this for the left hand side, by cutting w_i into p pieces with weight α_r ($r=1,\ldots,p$, $\sum_{r=1}^{p} \alpha_r = 1$) - see Fig. A1, where $p=2$.

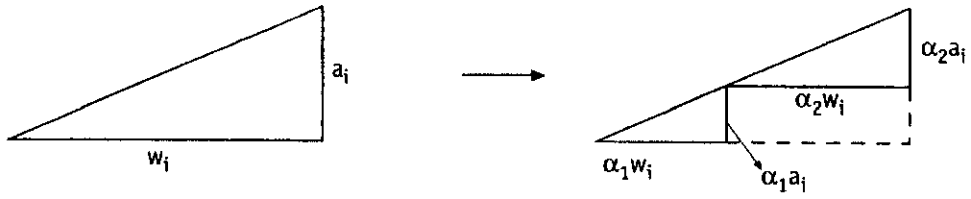

Fig. A1 Cutting w_i into two pieces

In this transformation, $\varphi\left(\frac{a_i}{w_i}\right) w_i$ is replaced by

$$\sum_{r=1}^{p} \varphi\left(\frac{\alpha_r a_i}{\alpha_r w_i}\right) \alpha_r w_i$$

$$= \sum_{r=1}^{p} \alpha_r \varphi\left(\frac{a_i}{w_i}\right) w_i$$

$$= \varphi\left(\frac{a_i}{w_i}\right) w_i$$

2. Given X, X′ and W, W′ as above we have to prove (11), (12) and (13) of theorem I.2. Now

$$\int_c^d dy(t) - \int_a^b dx(t) = \sum_{i=1}^{N} w_i - \sum_{i=1}^{N} w_i = 0$$

$$\int_c^d t\,dy(t) - \int_a^b t\,dx(t) = \sum_{i=1}^{N} \frac{a_i'}{w_i} w_i - \sum_{i=1}^{N} \frac{a_i}{w_i} w_i = 0$$

showing (11) and (12). For (13) we have

$$\int_a^b [t-u]^+ dx(t) = \sum_{i=1}^{N} w_i \left[\frac{a_i}{w_i} - u \right]^+$$

(i) Let $\frac{a_{i+1}}{w_{i+1}} \le u < \frac{a_i}{w_i}$

Noting that the $\frac{a_j}{w_j}$ decrease, we have

$$\int_a^b [t-u]^+ dx(t) = \sum_{j=1}^{i} w_j \left(\frac{a_j}{w_j} - u \right)$$

$$= \sum_{j=1}^{i} a_j - \left(\sum_{j=1}^{i} w_j \right) u$$

$$\le \sum_{j=1}^{i} a_j' - \left(\sum_{j=1}^{i} w_j \right) u$$

$$= \sum_{j=1}^{i} w_j \left(\frac{a_j'}{w_j} - u \right) \qquad \text{(A4)}$$

Since

$$\frac{a_N'}{w_N} \le \frac{a_N}{w_N} < \frac{a_1}{w_1} \le \frac{a_1'}{w_1} \qquad \text{(A5)}$$

(since $f<g$, f,g concave and polygonal) we have $\exists i' \in \{1,\ldots,N\}$ such that

$$\frac{a'_{i'+1}}{w_{i'+1}} \le u < \frac{a'_{i'}}{w_{i'}} \qquad \text{(A6)}$$

(a) Case $i=i'$

Then

$$(A4) = \int_c^d [t-u]^+ \, dy(t).$$

(b) Case $i' > i$

Now by (A6) and since also the $\frac{a_j'}{w_j}$ decrease we have $u \le \frac{a_j'}{w_j}$, $\forall j=i+1,\ldots,i'$ and hence

$$(A4) \le \sum_{j=1}^{i'} w_j \left(\frac{a_j'}{w_j} - u \right) = \int_c^d [t-u]^+ dy(t)$$

(c) Case $i' < i$

Now $u \ge \frac{a_j'}{w_j}$, $\forall j=i'+1,\ldots,i$, so

$$(A4) \le \sum_{j=1}^{i'} w_j \left(\frac{a_j'}{w_j} - u \right) = \int_c^d [t-u]^+ dy(t)$$

This shows (13) in case (i).

(ii) Let $\frac{a_N'}{w_N} \le u < \frac{a_N}{w_N}$

Now

$$\int_a^b [t-u]^+ dx(t)$$

$$= \sum_{j=1}^{N} w_j\left(\frac{a_j}{w_j}-u\right)$$

$$= \sum_{j=1}^{N} a_j - \left(\sum_{j=1}^{N} w_j\right)u$$

$$= \sum_{j=1}^{N} a_j' - \left(\sum_{j=1}^{N} w_j\right)u$$

$$= \sum_{j=1}^{N} w_j\left(\frac{a_j'}{w_j}-u\right) \qquad \text{(A7)}$$

By (A5) we have : $\exists i' \in \{1,\ldots,N\}$ such that

$$\frac{a'_{i'+1}}{w_{i'+1}} \le u < \frac{a'_{i'}}{w_{i'}}$$

Then $\frac{a_j'}{w_j} - u \le 0$, $\forall j = i'+1,\ldots,N$ and hence

$$\text{(A7)} \le \sum_{j=1}^{i'} w_j\left(\frac{a_j'}{w_j}-u\right) = \int_c^d [t-u]^+ dy(t)$$

(iii) Let $u < \frac{a_N'}{w_N}$

Hence u is smaller than all $\frac{a_i}{w_i}$ and $\frac{a_j'}{w_j}$ $(i,j=1,\ldots,N)$.

Hence

$$\int_a^b [t-u]^+ dx(t)$$

$$= \sum_{j=1}^{N} w_j\left(\frac{a_j}{w_j}-u\right)$$

$$= \sum_{j=1}^{N} w_j \left(\frac{a_j'}{w_j} - u \right)$$

$$= \int_c^d [t-u]^+ dy(t)$$

(iv) Let $\frac{a_1}{w_1} < u \leq \frac{a_1'}{w_1}$

Now

$$\int_a^b [t-u]^+ dx(t)$$

$$= 0 \leq \int_c^d [t-u]^+ dy(t).$$

(v) Let $u > \frac{a_1'}{w_1}$

Now u is larger than all $\frac{a_i}{w_i}$, $\frac{a_j'}{w_j}$ (i,j=1,...,N), hence both sides of (13) are 0.

This concludes the proof of (13) and hence of this theorem. □

References

P.D. Allison (1978). Measures of inequality. American Sociological Review 43, 865-880.

T.M. Apostol (1957). Mathematical Analysis. A modern Approach to advanced Calculus. Addison-Wesley, Reading, Massachusetts, USA.

A.B. Atkinson (1970). On the measurement of inequality. Journal of Economic Theory 2, 244-263.

H.D. Brunk (1956). On an inequality for convex functions. Proceedings of the American Mathematical Society 7, 817-824.

T.M. Cover and J.A. Thomas (1991). Elements of Information Theory. J. Wiley and Sons, New York, USA.

S.A. Douglas (1996). Introduction to Mathematical Analysis. Addison-Wesley, Reading, Massachusetts, USA.

L. Egghe (1990). A new method for information retrieval based on the theory of relative concentration. Proceedings of the 13th international Conference on Research and Development in Information Retrieval (SIGIR), Brussels (Belgium), 469-493.

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

L. Egghe and R. Rousseau (1990b). Elements of concentration theory. Proceedings of the second International Conference on Bibliometrics, Scientometrics and Informetrics, London (Ontario, Canada), 97-137.

L. Egghe and R. Rousseau (2000). Symmetric and asymmetric theory of relative concentration. Scientometrics, to appear.

L. Egghe and R. Rousseau (2001). Theory of i-truncation of a bibliography. Preprint.

J.L. Gastwirth (1971). A general definition of the Lorenz curve. Econometrica 39(6) (November 1971), 1037-1039.

J.L. Gastwirth (1972). The estimation of the Lorenz curve and Gini index. The Review of Economics and Statistics 54(3), 306-316.

G.H. Hardy, J.E. Littlewood and G. Pólya (1928). Some simple inequalities satisfied by convex functions. Messenger of Mathematics 58, 145-152.

S. Ihara (1993). Information Theory for continuous Systems. World Scientific, Singapore.

S.M.R. Kanbur (1984). The measurement and decomposition of inequality and poverty. IN : F. Van Der Ploeg (editor). Mathematical Methods in Economics, J. Wiley and Sons, 16, 403-432.

P.J. Lambert (1985). Social welfare and the Gini coefficient revisited. Mathematial Social Sciences 9, 19-26.

M.O. Lorenz (1905). Methods of measuring concentration of wealth. Journal of the American Statistical Association 9, 209-219.

G.P. Patil and C. Taillie (1982). Diversity as a concept and its measurement. Journal of the American Statistical Association 77(379), 548-561.

M. Rothschild and J.E. Stiglitz (1973). Some further results on the measurement of inequality. Journal of Economic Theory 6, 188-204.

R. Rousseau (1992). Concentration and Diversity in informetric Research. Doctorate Thesis, University of Antwerp (UIA) (Belgium).

R. Rousseau (2000). Concentration and evenness measures as scientometric indicators. Preprint.

G. Salton and M.J. Mc Gill (1987). Introduction to modern Information Retrieval. Mc Graw-Hill, Singapore.

A. Sen (1973). On economic Inequality. Clarendon Press, Oxford (UK).

H. Theil (1967). Economics and Information Theory. North-Holland, Amsterdam (the Netherlands).

P.D. Thistle (1989). Ranking distributions with generalized Lorenz curves. Southern Economic Journal 50(1), 1-12.

C.L. Viles and J.C. French (1999). Content locality in distributed digital libraries. Information Processing and Management 35, 317-336.