ISIT'2013 1569719321

Information Theoretic Measures of Distances and their Econometric Applications

Ferdinando Cicalese
Dipartimento di Informatica
Università di Salerno, Salerno, Italy
Email: cicalese@dia.unisa.it

Luisa Gargano
Dipartimento di Informatica
Università di Salerno, Salerno, Italy
Email: lg@dia.unisa.it

Ugo Vaccaro
Dipartimento di Informatica
Università di Salerno, Salerno, Italy
Email: uv@dia.unisa.it

***Abstract*—We introduce two new information theoretic measures of distances among probability distributions and we discuss their possible applications to Econometrics.**

I. Introduction

Measures of distances between probability distributions play an important role in several areas of pure and applied mathematics [1], [9]. A notable area of the applied sciences where such a kind of measures find extensive applications is the field of Econometrics. In fact, metrics on probability distributions are routinely used to solve problems of inference and discrimination that arise therein [26].

Two particularly interesting econometrics problems are the following: How to compare income/wealth distributions of different populations and how to measure the degree of income/wealth inequality among them? Dagum [5] and Shorrocks [22] first considered the questions and proposed measures of "economic distance" to deal with the consequential mathematical problems. A very general approach was put forward by Ebert [7], who proposed a set of reasonable axioms that *bona fide* measures of distance among income distributions should satisfy. Since then, the problem has been extensively studied, see [2], [15] and references therein quoted.

In bare mathematical terms, the question can be phrased as follows. Suppose there is a population of n income/wealth receivers and that person i has income equal to x_i. The income distribution is the vector $\mathbf{x} = (x_1, \ldots, x_n)$. There is no loss of generality in assuming that the x_i's have been normalized so that $0 \leq x_i \leq 1$ and $\sum_{i=0}^{n} x_i = 1$. In this case, the x_i's represent the income share of person i with respect to the total population wealth,[1] Now assume that the components of each possible income distribution $\mathbf{x} = (x_1, \ldots, x_n)$ are ordered in decreasing fashion, so that $x_1 \geq x_2 \geq \ldots \geq x_n$. A basic principle of economic analysis, due to Lorenz [14], postulates that $\mathbf{x} = (x_1, \ldots, x_n)$ represents a more even distribution of wealth than does $\mathbf{y} = (y_1, \ldots, y_n)$ if and only if

$$\sum_{i=1}^{k} x_i \leq \sum_{i=1}^{k} y_i, \quad \text{for each } k = 1, \ldots n. \tag{1}$$

In mathematical parlance [16], (1) is equivalent to say that vector $\mathbf{x}$ is *majorized* by $\mathbf{y}$, and the notation $\mathbf{x} \preceq \mathbf{y}$ is used.

[1] Actually, in real data analysis, data are grouped in percentiles

In this view, the vector $\mathbf{u}_n = (1/n, 1/n, \ldots, 1/n)$ is majorized by every income distribution $\mathbf{y}$, and clearly represents the most even wealth distribution, while the vector $\mathbf{1}_n = (1, 0, \ldots, 0)$ majorizes every vector $\mathbf{y}$ and therefore represents the most uneven wealth distribution.

The majorization order $\preceq$ among income vectors is clearly a partial order relation, and arbitrary vectors $\mathbf{x}$ and $\mathbf{y}$ might be incomparable by majorization. Economists have long sought reasonable indicators to compare pairs of income distributions $\mathbf{x}$ and $\mathbf{y}$, when neither $\mathbf{x} \preceq \mathbf{y}$ nor $\mathbf{y} \preceq \mathbf{x}$ holds. Such indicators, usually called *measures of inequality*, associate to each income distribution $\mathbf{x}$ a real number that is taken to represent the amount of inequality exhibited by $\mathbf{x}$. Among the many proposed measures of inequality (see, for example [16], chapt. 13.F), it is fair to say that the two most important ones are the Theil Index [25] and the Gini Index [10]. The Theil Index of $\mathbf{x}$ is defined as $\log n - H(\mathbf{x})$, where $H(\mathbf{x}) = -\sum_{i=1}^{n} x_i \log x_i$ is the Shannon entropy of $\mathbf{x}$, and the Gini Index of $\mathbf{x}$ is defined as $\frac{n+1}{n} - \frac{2}{n} \sum_{i=1}^{n} i x_i$. A moment of reflection reveals that both Theil and Gini Index aspire to measure how much a given income distribution is "distant" from the uniform distribution $\mathbf{u}_n = (1/n, \ldots, 1/n)$. In the present paper we aim at constructing *information theoretic* measures of distance to compare arbitrary pairs of income distributions $\mathbf{x}$ and $\mathbf{y}$. In the particular case in which either $\mathbf{x}$ or $\mathbf{y}$ is the uniform distribution, our measures coherently reduce to the Theil or Gini Index. It is interesting to note that recent research in Econometrics [15], [4] aims at revisiting the classical approach that measure the amount of inequality exhibited by an income distribution $\mathbf{x}$ by its distance from the "ideal" uniform distribution $\mathbf{u}_n = (1/n, \ldots, 1/n)$. Indeed, for policy or social motives, the target distribution could be some particular $\mathbf{y}^* \neq \mathbf{u}_n$, and therefore the index of inequality of $\mathbf{x}$ should be measured by its distance from $\mathbf{y}^*$. This adds further incentives to studying reasonable distance measures among income distributions.

We would like to remark that we have chosen to cast our problem in the econometric scenario only for sake of conciseness and concreteness. Our information theoretic distance measures are likely to be relevant to other areas as well, like Biology, Ecology, and Social Sciences [24]. For instance, the measurement of species diversity in Ecology is essentially equivalent to the econometric inequality measurement [20].

We conclude this introductory section by emphasizing that it should come as no surprise that information theoretic concepts play an important role in Econometrics. Indeed, starting from the seminal work by Theil [25], there is nowadays a vast literature witnessing the fruitful application of information theoretic concepts to Econometrics; we refer the reader to [11], [13] for excellent overviews of the area.

II. Majorization and Schur–Convex Functions

We recall the basic notions of majorization theory that are relevant to our context; the monograph [16] gives an extensive treatment of the area.

Definition 1: Given two probability distributions $\mathbf{p} = (p_1, \ldots, p_n)$ and $\mathbf{q} = (q_1, \ldots, q_n)$ with $p_1 \geq \ldots \geq p_n$ and $q_1 \geq \ldots \geq q_n$, we say that $\mathbf{p}$ is *majorized* by $\mathbf{q}$, and write $\mathbf{p} \preceq \mathbf{q}$, if and only if

$$\sum_{i=1}^{k} p_i \leq \sum_{i=1}^{k} q_i, \quad \text{for all} \quad k = 1, \ldots, n.$$

The relation $\preceq$ is a partial ordering on the set $\mathcal{P}_n = \{(p_1, \ldots, p_n) \ : \sum_{i=1}^{n} p_i = 1, \ p_1 \geq \ldots \geq p_n\}$ of all ordered probability vectors of n elements [16], that is, for each $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{P}$ it holds that

- $\mathbf{x} \preceq \mathbf{x}$,
- $\mathbf{x} \preceq \mathbf{y}$ and $\mathbf{y} \preceq \mathbf{z}$ implies $\mathbf{x} \preceq \mathbf{z}$,
- $\mathbf{x} \preceq \mathbf{y}$ and $\mathbf{y} \preceq \mathbf{x}$ implies $\mathbf{x} = \mathbf{y}$.

It was proved in [3] that the poset $(\mathcal{P}_n, \preceq)$ is a lattice, i.e., for all $\mathbf{x}, \mathbf{y} \in \mathcal{P}_n$ there exists a unique *least upper bound* $\mathbf{x} \vee \mathbf{y}$ and a unique *greatest lower bound* $\mathbf{x} \wedge \mathbf{y}$. We recall that $\mathbf{x} \vee \mathbf{y}$ is the vector in $\mathcal{P}_n$ such that: $\mathbf{x} \preceq \mathbf{x} \vee \mathbf{y}$, $\mathbf{y} \preceq \mathbf{x} \vee \mathbf{y}$, and for each $\mathbf{z} \in \mathcal{P}_n$ it holds that $\mathbf{x} \preceq \mathbf{z}$, $\mathbf{y} \preceq \mathbf{z}$ implies $\mathbf{x} \vee \mathbf{y} \preceq \mathbf{z}$. Analogously, $\mathbf{x} \wedge \mathbf{y}$ is the vector in $\mathcal{P}_n$ such that $\mathbf{x} \wedge \mathbf{y} \preceq \mathbf{x}$, $\mathbf{x} \wedge \mathbf{y} \preceq \mathbf{y}$, and for each $\mathbf{z} \in \mathcal{P}_n$ it holds that $\mathbf{z} \preceq \mathbf{x}$, $\mathbf{z} \preceq \mathbf{y}$ implies $\mathbf{z} \preceq \mathbf{x} \wedge \mathbf{y}$. Moreover,

$$(1/n, \ldots, 1/n) \preceq \mathbf{p} \preceq (1, \ldots, 0), \qquad \text{for all } \mathbf{p} \in \mathcal{P}_n. \quad (2)$$

It follows that vectors $(1/n, \ldots, 1/n)$ and $(1, 0, \ldots, 0)$ are the minimum and the maximum, respectively, of the lattice $\mathcal{P}_n$. In paper [3] the authors also gave a simple algorithm to explicitly compute $\mathbf{x} \vee \mathbf{y}$ and $\mathbf{x} \wedge \mathbf{y}$, given arbitrary vectors $\mathbf{x}, \mathbf{y} \in \mathcal{P}_n$.

Functions that preserve (resp., reverse) the majorization ordering are called *Schur–convex (resp., concave) functions*.

Definition 2: A real valued function $\phi : \mathcal{P}_n \to \mathbb{R}$ is said to be Schur–convex if

$$\mathbf{p} \preceq \mathbf{q} \ \text{ implies } \ \phi(\mathbf{p}) \leq \phi(\mathbf{q}).$$

Moreover, if $\phi(\mathbf{p}) < \phi(\mathbf{q})$ whenever $\mathbf{p} \preceq \mathbf{q}$ and $\mathbf{p}$ is not a permutation of $\mathbf{q}$, then ϕ is said to be *strictly* Schur–convex. A function ϕ is said to be (strictly) Schur–concave if $-\phi$ is (strictly) Schur–convex.

It is well known that the entropy function $H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \log p_i$ is strictly Schur-concave ([16], p. 101). This implies the following results that will be used in the sequel.

Fact 1: [16] Let $\mathbf{x}, \mathbf{y} \in \mathcal{P}_n$, with $\mathbf{x} \preceq \mathbf{y}$. One has $H(\mathbf{x}) \geq H(\mathbf{y})$, with equality if and only if $\mathbf{x} = \mathbf{y}$. □

The equality condition in the above statement follows from the strict Schur-concavity of the entropy function and from the fact that the components of the probability vectors in $\mathcal{P}_n$ are ordered in non–decreasing fashion. A strengthening of above fact has been provided in [12]. There, the authors prove that $\mathbf{x} \preceq \mathbf{y}$ implies $H(\mathbf{x}) \geq H(\mathbf{y}) + D(\mathbf{x}||\mathbf{y})$, where $D(\mathbf{x}||\mathbf{y})$ is the relative entropy between $\mathbf{x}$ and $\mathbf{y}$.

Another result we shall use is the supermodularity of the entropy function $H(\mathbf{p})$ on the lattice $(\mathcal{P}_n, \preceq)$.

Theorem 1: [3] For all $\mathbf{x}, \mathbf{y} \in \mathcal{P}_n$, it holds that

$$H(\mathbf{x} \vee \mathbf{y}) + H(\mathbf{x} \wedge \mathbf{y}) \geq H(\mathbf{x}) + H(\mathbf{y}).$$

III. An Entropy Based Distance Measure on $\mathcal{P}_n$

In this section we introduce a new distance measure on $\mathcal{P}_n$. Subsequently, we shall apply it to the study of inter-country inequality measure.

Definition 3: $\forall\ \mathbf{x}, \mathbf{y} \in \mathcal{P}_n$ we define $\mathtt{d}(\mathbf{x}, \mathbf{y})$ as

$$\mathtt{d}(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}) + H(\mathbf{y}) - 2H(\mathbf{x} \vee \mathbf{y}), \quad (3)$$

where H is the entropy function.

Before proving that the function $\mathtt{d}(\mathbf{x}, \mathbf{y})$ is a metric on $\mathcal{P}_n$, we observe that $\mathtt{d}$ is a proper generalization of the famous Theil Index [25], widely used in econometrics. Indeed, let $\mathbf{u}_n = (1/n, \ldots, n)$ be the uniform vector of length n, that is, the minimum in the lattice $\mathcal{P}_n$. By definition, for each $\mathbf{x} \in \mathcal{P}_n$ one has that $\mathbf{x} \vee \mathbf{u}_n = \mathbf{x}$. Therefore

$$\mathtt{d}(\mathbf{x}, \mathbf{u}_n) = H(\mathbf{x}) + H(\mathbf{u}_n) - 2H(\mathbf{x}) = \log n - H(\mathbf{x}), \quad (4)$$

that is exactly the Theil index of vector $\mathbf{x}$. Above equality is intuitively pleasant, since it states that the Theil Index corresponds to the distance $\mathtt{d}(\mathbf{x}, \mathbf{u}_n)$ of vector $\mathbf{x}$ from the uniform vector $\mathbf{u}_n$, thus confirming the (admittedly) already quite natural interpretation of Theil Index of $\mathbf{x}$ as a measure of how much $\mathbf{x}$ is far from being uniform.

Theorem 2: Function $\mathtt{d}$ is a metric on $\mathcal{P}_n$.

Proof. In order to show that $\mathtt{d}$ is a metric we need to prove that for all $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{P}_n$

i) $\mathtt{d}(\mathbf{x}, \mathbf{y}) = \mathtt{d}(\mathbf{y}, \mathbf{x})$;
ii) $\mathtt{d}(\mathbf{x}, \mathbf{y}) \geq 0$, with equality if and only if $\mathbf{x} = \mathbf{y}$;
iii) $\mathtt{d}(\mathbf{x}, \mathbf{y}) \leq \mathtt{d}(\mathbf{x}, \mathbf{z}) + \mathtt{d}(\mathbf{z}, \mathbf{y})$.

The symmetry property $d(\mathbf{x}, \mathbf{y}) = \mathtt{d}(\mathbf{y}, \mathbf{x})$ trivially holds.
We show now ii). Let us first observe that, by definition, $\mathbf{x} \preceq \mathbf{x} \vee \mathbf{y}$ and $\mathbf{y} \preceq \mathbf{x} \vee \mathbf{y}$ both hold. The strict Schur–concavity of the entropy function $H(\cdot)$ tells us that $H(\mathbf{x}) \geq H(\mathbf{x} \vee \mathbf{y})$ and $H(\mathbf{y}) \geq H(\mathbf{x} \vee \mathbf{y})$ (and the two equalities simultaneously hold only if $\mathbf{x} = \mathbf{x} \vee \mathbf{y} = \mathbf{y}$). This implies that $\mathtt{d}(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}) + H(\mathbf{y}) - 2H(\mathbf{x} \vee \mathbf{y}) \geq 0$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{P}_n$, with equality if and only if $\mathbf{x} = \mathbf{y}$.
We prove now the triangle inequality iii). Let us define the vectors $\mathbf{a} = \mathbf{x} \vee \mathbf{z}$ and $\mathbf{b} = \mathbf{y} \vee \mathbf{z}$. By definition we have $\mathbf{z} \preceq \mathbf{a} \wedge \mathbf{b}$ and, by Fact 1, we get

$$H(\mathbf{z}) \geq H(\mathbf{a} \wedge \mathbf{b}). \quad (5)$$

We have then

$$
\begin{aligned}
\mathtt{d}(\mathbf{x},\mathbf{z})+\mathtt{d}(\mathbf{z},\mathbf{y})-\mathtt{d}(\mathbf{x},\mathbf{y}) &= 2H(\mathbf{z})-2H(\mathbf{x}\vee\mathbf{z})\\
&\quad -2H(\mathbf{y}\vee\mathbf{z})+2H(\mathbf{x}\vee\mathbf{y})\\
&= 2[H(\mathbf{z})+H(\mathbf{x}\vee\mathbf{y})\\
&\quad -(H(\mathbf{a})+H(\mathbf{b}))]\\
&\geq 2[H(\mathbf{z})+H(\mathbf{x}\vee\mathbf{y})-H(\mathbf{a}\vee\mathbf{b})\\
&\quad -H(\mathbf{a}\wedge\mathbf{b})] \text{ (by Theorem 1)}\\
&\geq 2[H(\mathbf{x}\vee\mathbf{y})-H(\mathbf{a}\vee\mathbf{b})] \qquad (6)\\
&\qquad \text{(by (5))}
\end{aligned}
$$

We notice that $\mathbf{x}\vee\mathbf{y} \preceq (\mathbf{x}\vee\mathbf{z})\vee(\mathbf{y}\vee\mathbf{z}) = \mathbf{a}\vee\mathbf{b}$, which implies $H(\mathbf{x}\vee\mathbf{y}) \geq H(\mathbf{a}\vee\mathbf{b})$. Therefore, by (6) we get that $\mathtt{d}(\mathbf{x},\mathbf{z})+\mathtt{d}(\mathbf{z},\mathbf{y})-\mathtt{d}(\mathbf{x},\mathbf{y}) \geq 0$ and iii) holds. □

In order to get some intuition about formula (3) one can proceed as follows. As said before, in the lattice $(\mathcal{P}_n, \preceq)$ vector $\mathbf{u}_n = (1/n,\ldots,1/n)$ is at the bottom and vector $\mathbf{1}_n = (1,0,\ldots,0)$ is at the top. For any vector $\mathbf{x} \in \mathcal{P}_n$, formula (3) postulates that the distance of $\mathbf{x}$ from the bottom of the lattice is $\mathtt{d}(\mathbf{x},\mathbf{u}_n) = \log n - H(\mathbf{x})$, while the distance of $\mathbf{x}$ from the top of the lattice is $\mathtt{d}(\mathbf{x},\mathbf{1}_n) = H(\mathbf{x})$.[2] Moreover, if $\mathbf{x},\mathbf{y}$ are directly comparable in $\mathcal{P}_n$, for instance $\mathbf{x} \preceq \mathbf{y}$, then we can use the fact that $\mathbf{x} \preceq \mathbf{y} \preceq \mathbf{1}_n$, employ vector $\mathbf{1}_n$ as a "reference point", and compute the distance between $\mathbf{x}$ and $\mathbf{y}$ as $\mathtt{d}(\mathbf{x},\mathbf{1}_n) - \mathtt{d}(\mathbf{y},\mathbf{1}_n) = H(\mathbf{x}) - H(\mathbf{y})$, that is exactly equal to what (3) says in the case $\mathbf{x} \preceq \mathbf{y}$. If $\mathbf{x}$ and $\mathbf{y}$ are not directly comparable in $\mathcal{P}_n$, (i.e., neither $\mathbf{x} \preceq \mathbf{y}$ nor $\mathbf{y} \preceq \mathbf{x}$ holds) then one first compute vector $\mathbf{x}\vee\mathbf{y}$, that represents the "lowest" vector in $\mathcal{P}_n$ that is "above" $\mathbf{x}$ and $\mathbf{y}$ and it is also *directly comparable* both to $\mathbf{x}$ and $\mathbf{y}$. Subsequently, using $\mathbf{x}\vee\mathbf{y}$ as a reference point, one computes the distance from $\mathbf{x}$ to $\mathbf{y}$ as

$$
\begin{aligned}
\mathtt{d}(\mathbf{x},\mathbf{x}\vee\mathbf{y})+\mathtt{d}(\mathbf{y},\mathbf{x}\vee\mathbf{y}) &= \mathtt{d}(\mathbf{x},\mathbf{1}_n)-\mathtt{d}(\mathbf{x}\vee\mathbf{y},\mathbf{1}_n)+\\
&\quad \mathtt{d}(\mathbf{y},\mathbf{1}_n)-\mathtt{d}(\mathbf{x}\vee\mathbf{y},\mathbf{1}_n)\\
&= [H(\mathbf{x})-H(\mathbf{x}\vee\mathbf{y})]+\\
&\quad [H(\mathbf{y})-H(\mathbf{x}\vee\mathbf{y})] = \mathtt{d}(\mathbf{x},\mathbf{y})
\end{aligned}
$$

Equivalently, we are defining $\mathtt{d}(\mathbf{x},\mathbf{y})$ as being equal to $\min\{d(\mathbf{x},\mathbf{z})+d(\mathbf{z},\mathbf{y})\}$, where the min is taken among all vectors $\mathbf{z}$ such that $\mathbf{x} \preceq \mathbf{z}$ and $\mathbf{y} \preceq \mathbf{z}$.

IV. Gini Index based distance measure

In this section, we introduce a new distance measure on $\mathcal{P}_n$ based on the classical Gini Index.

Definition 4: [10], [16] Given $\mathbf{x} \in \mathcal{P}_n$ the Gini Index of $\mathbf{x}$ is defined as

$$
G(\mathbf{x}) = \frac{n+1}{n} - \frac{2}{n}A(\mathbf{x}), \text{ where } A(\mathbf{x}) = \sum_{i=1}^{n} i x_i.
$$

The reader will notice an analogy between the Theil Index $\log n - H(\mathbf{x})$ and the Gini Index $\frac{n+1}{n} - \frac{2}{n}\sum_{i=1}^n i x_i$ of $\mathbf{x}$. Indeed, one can recognize in $\sum_{i=1}^n i x_i$ the "guessing entropy" defined by Massey in [17], and in $(n+1)/2$ its maximum value. Therefore, just as the Theil Index of vector $\mathbf{x}$ is the *difference* between the maximum entropy value (i.e., $\log n$) and the Shannon entropy of $\mathbf{x}$, likewise the Gini Index of $\mathbf{x}$ is the *difference* (times $2/n$) between the maximum guessing entropy value and the guessing entropy of $\mathbf{x}$.

It is known [16] that functions G and A are strictly Schur–convex and Schur–concave, respectively. Namely, for each $\mathbf{x},\mathbf{y} \in \mathcal{P}_n$ it holds that

$$
\mathbf{x} \preceq \mathbf{y} \quad \text{implies} \quad G(\mathbf{x}) \leq G(\mathbf{y}) \quad \text{and} \quad A(\mathbf{x}) \geq A(\mathbf{y}).
$$

Moreover, above inequalities are strict whenever $\mathbf{x} \neq \mathbf{y}$.

Definition 5: $\forall\, \mathbf{x},\mathbf{y} \in \mathcal{P}_n$ we define $\mathtt{D}(\mathbf{x},\mathbf{y})$ as

$$
\begin{aligned}
\mathtt{D}(\mathbf{x},\mathbf{y}) &= G(\mathbf{x})+G(\mathbf{y})-2G(\mathbf{x}\wedge\mathbf{y})\\
&= \frac{2}{n}\left[2A(\mathbf{x}\wedge\mathbf{y})-A(\mathbf{x})-A(\mathbf{y})\right]
\end{aligned}
$$

Again, first we would like to prove that the Gini Index of a vector $\mathbf{x}$ can be seen as the distance, measured according to function $\mathtt{D}$, of vector $\mathbf{x}$ from the uniform vector $\mathbf{u}_n$. Indeed, we have $\mathbf{x}\wedge\mathbf{u}_n = \mathbf{u}_n$ and $G(\mathbf{u}_n) = \frac{n+1}{n} - \frac{2}{n}\sum_{i=1}^n \frac{i}{n} = \frac{n+1}{n} - \frac{2}{n^2}\frac{n(n+1)}{2} = 0$. Hence, $G(\mathbf{x}\wedge\mathbf{u}_n) = G(\mathbf{u}_n) = 0$ and, finally, $D(\mathbf{x},\mathbf{u}_n) = G(\mathbf{x}) + G(\mathbf{u}_n) - 2G(\mathbf{x}\wedge\mathbf{u}_n) = G(\mathbf{x})$.

Theorem 3: Function $\mathtt{D}$ is a metric on $\mathcal{P}_n$.

Proof. In order to prove that $\mathtt{D}$ is a metric we need to prove that for all $\mathbf{x},\mathbf{y},\mathbf{z} \in \mathcal{P}_n$

i) $\mathtt{D}(\mathbf{x},\mathbf{y}) = \mathtt{D}(\mathbf{y},\mathbf{x})$;
ii) $\mathtt{D}(\mathbf{x},\mathbf{y}) \geq 0$ with equality if and only if $\mathbf{x} = \mathbf{y}$;
iii) $\mathtt{D}(\mathbf{x},\mathbf{y}) \leq \mathtt{D}(\mathbf{x},\mathbf{z}) + \mathtt{D}(\mathbf{z},\mathbf{y})$.

Equality i) trivially holds by the definition of $\mathtt{D}(\mathbf{x},\mathbf{y})$.

We now prove ii). Since $\mathbf{x}\wedge\mathbf{y} \preceq \mathbf{x}$ and $\mathbf{x}\wedge\mathbf{y} \preceq \mathbf{y}$, the strict Schur–convexity of the G function implies that $G(\mathbf{x}\wedge\mathbf{y}) \leq G(\mathbf{x})$ and $G(\mathbf{x}\wedge\mathbf{y}) \leq G(\mathbf{y})$, where the two equalities simultaneously hold only if $\mathbf{x} = \mathbf{x}\wedge\mathbf{y} = \mathbf{y}$. Therefore, $\mathtt{D}(\mathbf{x},\mathbf{y}) = G(\mathbf{x}) + G(\mathbf{y}) - 2G(\mathbf{x}\wedge\mathbf{y}) \geq 0$ with equality if and only if $\mathbf{x} = \mathbf{y}$.

Finally, we prove that $\mathtt{D}$ satisfies the triangle inequality iii). We have

$$
\begin{aligned}
\mathtt{D}(\mathbf{x},\mathbf{z})+\mathtt{D}(\mathbf{z},\mathbf{y})-\mathtt{D}(\mathbf{x},\mathbf{y}) &= 2G(\mathbf{z})-2G(\mathbf{x}\wedge\mathbf{z})\\
&\quad -2G(\mathbf{y}\wedge\mathbf{z})+2G(\mathbf{x}\wedge\mathbf{y})\\
&= 2[G(\mathbf{z})-G(\mathbf{x}\wedge\mathbf{z})-G(\mathbf{y}\wedge\mathbf{z})\\
&\quad +G(\mathbf{x}\wedge\mathbf{y})]\\
&= -\frac{2}{n}[A(\mathbf{z})+A(\mathbf{x}\wedge\mathbf{y})\\
&\quad -A(\mathbf{x}\wedge\mathbf{z})-A(\mathbf{y}\wedge\mathbf{z})]
\end{aligned}
$$

Therefore, $\mathtt{D}(\mathbf{x},\mathbf{y}) \leq \mathtt{D}(\mathbf{x},\mathbf{z}) + \mathtt{D}(\mathbf{z},\mathbf{y})$ if and only if

$$
A(\mathbf{z})+A(\mathbf{x}\wedge\mathbf{y})-A(\mathbf{x}\wedge\mathbf{z})-A(\mathbf{y}\wedge\mathbf{z}) \leq 0, \qquad (7)
$$

where we recall that $A(\mathbf{a}) = \sum_{i=1}^n i a_i$. We notice now that for any $\mathbf{a},\mathbf{b} \in \mathcal{P}_n$, the vector $(\mathbf{a}+\mathbf{b})/2 = ((a_1+b_1)/2,\ldots,(a_n+b_n)/2)$ satisfies

$$
A\left(\frac{\mathbf{a}+\mathbf{b}}{2}\right) = \frac{A(\mathbf{a})+A(\mathbf{b})}{2}.
$$

[2]Schur-concave functions of $\mathbf{x}$ that can be seen as suitable measure of distance of $\mathbf{x}$ from $\mathbf{1}_n$ are shown, in [19], to satisfy all of a series of properties that qualify them as "uncertainty functions". Again, no surprise here.

This implies that

$$\begin{aligned} &A(\mathbf{z}) + A(\mathbf{x}\wedge\mathbf{y}) - [A(\mathbf{x}\wedge\mathbf{z}) + A(\mathbf{y}\wedge\mathbf{z})] \\ &= \quad 2A\left(\frac{\mathbf{z}+(\mathbf{x}\wedge\mathbf{y})}{2}\right) - 2A\left(\frac{(\mathbf{x}\wedge\mathbf{z})+(\mathbf{y}\wedge\mathbf{z})}{2}\right). \end{aligned} \tag{8}$$

Therefore, by (7) and (8) we have that iii) holds if and only if

$$A\left(\frac{\mathbf{z}+(\mathbf{x}\wedge\mathbf{y})}{2}\right) \le A\left(\frac{(\mathbf{x}\wedge\mathbf{z})+(\mathbf{y}\wedge\mathbf{z})}{2}\right). \tag{9}$$

Since $A(\cdot)$ is a Schur–concave function, to prove (9) it suffices to prove that

$$(\mathbf{x}\wedge\mathbf{z}) + (\mathbf{y}\wedge\mathbf{z}) \preceq \mathbf{z} + (\mathbf{x}\wedge\mathbf{y}). \tag{10}$$

Here we need to recall the following result from [3]: Given arbitrary vectors $\mathbf{a} = (a_1, \dots, a_n), \mathbf{b} = (b_1, \dots, b_n) \in \mathcal{P}_n$, the vector $\mathbf{a}\wedge\mathbf{b} = (\alpha_1, \dots, \alpha_n)$ satisfies

$$\sum_{i=1}^{j} \alpha_i = \min\left\{\sum_{i=1}^{j} a_i, \sum_{i=1}^{j} b_i\right\}, \tag{11}$$

for each $j = 1, \dots, n$. Therefore, by Definition 1 one can see that formula (10) holds if and only if for each $j = 1, \dots, n$

$$\begin{aligned} \min\left\{\sum_{i=1}^{j} x_i, \sum_{i=1}^{j} z_i\right\} + \min\left\{\sum_{i=1}^{j} y_i, \sum_{i=1}^{j} z_i\right\} \\ \le \sum_{i=1}^{j} z_i + \min\left\{\sum_{i=1}^{j} x_i, \sum_{i=1}^{j} y_i\right\} \end{aligned}$$

that holds true no matter what the ordering among the values $\sum_{i=1}^{j} x_i$, $\sum_{i=1}^{j} y_i$, and $\sum_{i=1}^{j} z_i$ is. $\square$

V. Case study

We have applied our two distances measures $d(\cdot,\cdot)$ and $D(\cdot,\cdot)$ to the comparison of wealth distribution among different countries. Our data set comprises data of 25 countries extracted from the 2010 Credit Suisse Global Wealth Databook [23, Table 3-3] and from the report of World Institute for Development Economics Research of the United Nations on the wealth distributions in 2000 [6, Table10b]. For each one of the 25 countries we extracted the distribution of the country population among the wealth deciles. On these distributions we used our measures to compute pairwise distances among the countries. For the visualization of the results we employed the Neighbor Joining algorithm [21], [18] as implemented in the package PHYLIP v. 3.69 [8]. Neighbor Joining is a hierarchical clustering algorithm which tries to map the input data into the leaves of a tree in such a way that the tree distance for any pair of leaves approximates the actual distance of the corresponding input items associated to those leaves. The Neighbor Joining algorithm is generally used to produce phylogenies assuming that the underlying distance matrix is *nearly additive*, meaning that it is close to represent tree distances. However, Neighbor Joining is known to produce good results also when the above condition is not satisfied. The algorithm starts with a star-tree configuration and iteratively joins the closest two elements and substitutes them with a new node, whose distance from the other nodes is computed on the basis of the distances of the joined nodes. When only one node is left, the process is reversely unfolded to build a binary tree. The branches' lengths are computed in order to represent the distance from the joined elements to the newly created node. Results are depicted in Figures 1-4 below. Discussions of the results and comparison with those obtainable using classical distances among wealth distributions are deferred to the extended journal version of the paper.

References

[1] J. Abrahams, "On the selection of measures of distances between probability distributions", *Information Sci.* vol. **26** (1982), 109–113.

[2] Q. L. Burrell, "Measuring relative equality of concentration between different income/wealth distributions", in: *Adv. on Income Inequality and Concentration Measures*, Routledge Chapman & Hall (2008).

[3] F. Cicalese, U. Vaccaro, "Supermodularity and subadditivity properties of the entropy on the majorization lattice", *IEEE Transactions on Information Theory* vol. **48** (4) (2002), 933–938.

[4] S. Bandyopadhyay,, F. Cowell, E. Flachaire, "Reference Distributions and Inequality Measurement", *J. of Economic Inequality*, forthcoming.

[5] C. Dagum, "Inequality measures between income distributions with applications", *Econometrica* vol. **48** (1980), 1191–1803.

[6] J.B. Davies, S. Sandstrom, A. Shorrocks, and E. N. Wolff, "The World Distribution of Household Wealth", World Institute for Development Economics Research of the United Nations Univ., WIDER Discussion Paper Volume: 2008/03. Helsinki, Finland.

[7] U. Ebert, "Measures of distance between income distributions", *Journal of Economic Theory* vol. **32** (1984), 266–274.

[8] J. Felsenstein, PHYLIP - Phylogeny Inference Package v. 3.69, available at `http://www.phylip.com`

[9] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics", *International Statistical Review* vol. **70**, (2002), 419–435.

[10] C. Gini, "Measurement of inequality of incomes", *The Economic Journal* vol. **31**, (1927), 124–126.

[11] A. Golan, *Information and Entropy Econometrics - A Review and Synthesis*, Now Publishers Inc., (2008)

[12] S. W. Ho and S. Verdu, "On the interplay between conditional entropy and error probability", *IEEE Trans. on Information Theory*, vol. **56** (2010), 5930–5942.

[13] G. G. Judge, R. C. Mittelhammer, *An Information Theoretic Approach to Econometrics*, Cambridge University Press, (2011).

[14] M.O. Lorenz, "Methods of measuring concentration of wealth", *Journal of the American Statistical Association*, vol. **9**, 2019–219, (1905).

[15] B. Magdalou and R. Nock, "Income Distributions and Decomposable Divergence Measures", *J. of Economic Theory*, **146**,2440–2454, (2011)

[16] A.W. Marshall, I. Olkin, B. C. Arnold, *Inequalities: Theory of Majorization and Its Applications*, Springer, New York, (2009).

[17] J.L. Massey, "Guessing and entropy", in: *Proc. of ISIT 1994*, 27 June – 01 July 1994, Trondheim, p. 204.

[18] R. Mihaescu, D. Levy, L. Pachter, "Why neighbor-joining works", *Algorithmica*, vol. **54** (1), pp. 124, 2009.

[19] D. Morales, L. Pardo, and I. Vajda, "Uncertainty of Discrete Stochastic Systems: General Theory and Statistical Inference", *IEEE Transactions on System, Man, and Cybernetics – Part A*, **26**, 681–697, (1997).

[20] C.E. Pielou, *Ecological Diversity*, Wiley, NY, 1975.

[21] N. Saitou and M. Nei, "The neighbor joining method: a new method for reconstructing phylogenetic trees", *Molecular Biology and Evolution*, vol. **4** (4), 406425, (1987).

[22] A. F. Shorrocks, "On the distance between income distributions", *Econometrica* vol. **50** (1982), 1337–1339.

[23] A. Shorrocks, J. B. Davies, R. Lluberas, *Credit Suisse Global Wealth Databook*, Credit Suisse Research Institute, 2010.

[24] I. T. Tabner, "A Review of Concentration, Diversity and Entropy Metrics in Economics, Finance, Ecology and Communication Sciences", *The Intern'l J. of Interdisciplinary Soc. Sci.*, vol. **2** (4), (2007), 53-60.

[25] H. Theil, *Economics and Information Theory*, North-Holland, 1967.

[26] A. Ullah, "Entropy, divergence and distance measures with econometric applications", *Journal of Statistical Planning and Inferences* vol. **49** (1) (1996), 137–162.

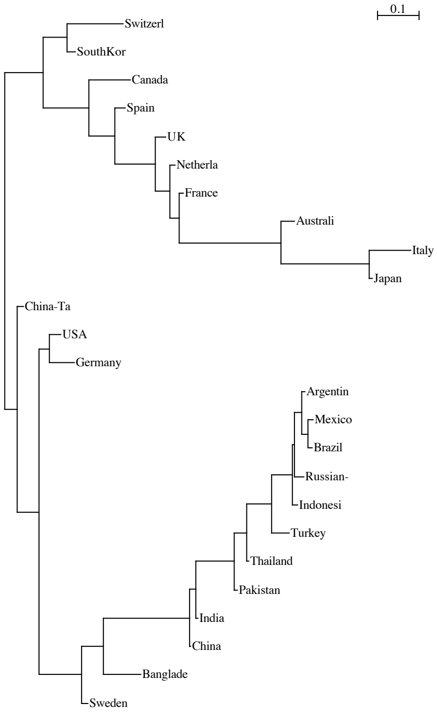

Fig. 1. NJ based hierarchical clustering of the distances among wealth distribution of the selected 25 countries in year 2010, using the distance $d(\cdot,\cdot)$ of section III.

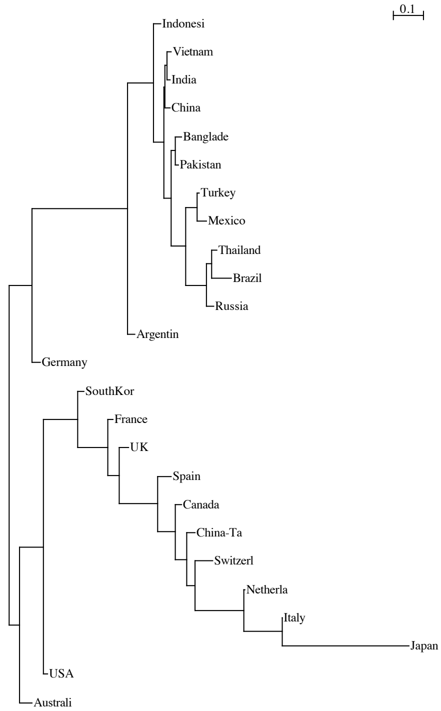

Fig. 3. NJ based hierarchical clustering of the distances among wealth distribution of the selected 25 countries in year 2000, using the distance $d(\cdot,\cdot)$ of section III.

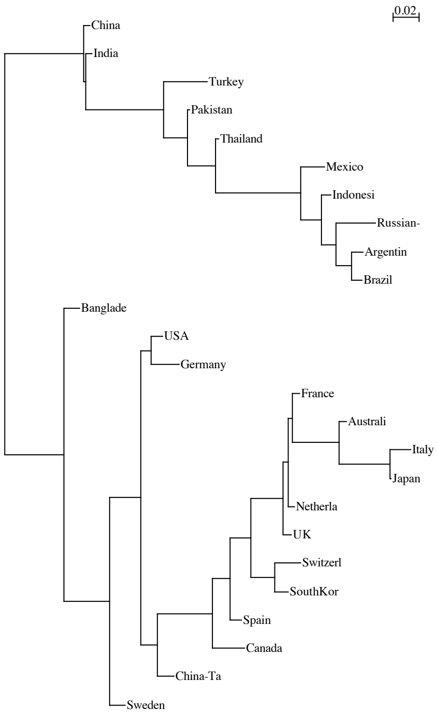

Fig. 2. NJ based hierarchical clustering of the distances among wealth distribution of the selected 25 countries in year 2010, using the distance $D(\cdot,\cdot)$ of section IV.

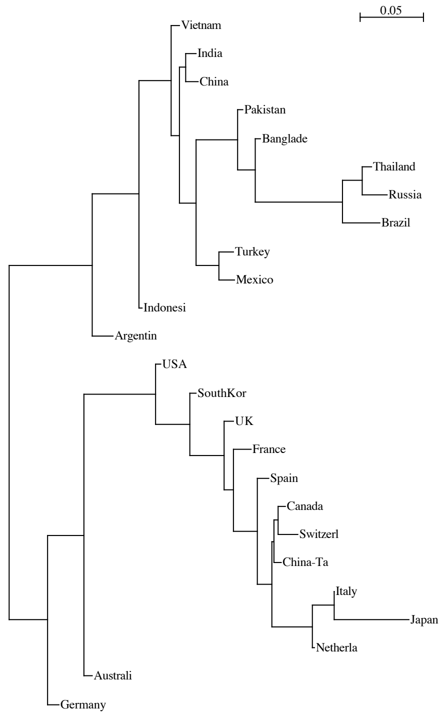

Fig. 4. NJ based hierarchical clustering of the distances among wealth distribution of the selected 25 countries in year 2000, using the distance $D(\cdot,\cdot)$ of section IV.