Statistical Functionals Consistent with a Weak Relative Majorization Ordering: Applications to the Mimimum Divergence Estimation

[1]TOMMASO LANDO, [2]LUCIO BERTOLI-BARSOTTI

[1]Department of Finance 70121
VŠB-Technical University of Ostrava
Sokolskà 33, Ostrava
CZECH REPUBLIC
tommaso.lando@vsb.cz

[2]Dipartimento di Scienze aziendali, economiche e metodi quantitativi
University of Bergamo
Via Dei Caniana 2, 24127, Bergamo
ITALY
lucio.bertoli-barsotti@unibg.it

Abstract: - Most of the statistical estimation procedures are based on a quite simple principle: find the distribution that, within a certain class, is as similar as possible to the empirical distribution, obtained from the sample observations. This leads to the minimization of some statistical functionals, usually interpreted ad measures of distance or divergence between distributions. In this paper we study the majorization pre-order of the distance between distributions. This concept, known in literature as relative majorization, is extended to the weak definition of majorization, which is more relevant in many practical contexts such as estimation problem. Providing mathematical proofs, we study under which conditions statistical functionals are consistent with respect to the relative weak majorization (from above) pre-order.

Key-Words: - Majorization; statistical functional; divergence measure; estimation; convex function; inequality; dissimilarity

1 Introduction

Within the field of statistical inference, most of the estimation procedures consist in minimizing an appropriate measure of distance between a theoretical and an empirical (observed) distribution. Statistical functionals which are suitable to measure distance between distributions are the well known divergence measures [1], [8]. In this paper we propose to analyze divergence measures and their properties by using majorization theory [18], [12]. Majorization is a pre-order on vectors used to analyze f the components of a vector are more or less "uniform", compared to another vector. Majorization can be generalized from vectors to measureable functions, as summarized in section 2. To be more specific, in the paper we refer to a generalization known as relative majorization [10] (r-majorization) which can be used to compare any couple of theoretical distributions (say H, G) with respect to a benchmark distribution (say F) in terms of dissimilarity. In particular, if G is smaller than a H relative to the ordering, then G is closer to F than H. With the assumption that the theoretical distribution is discrete and that our benchmark distribution (for estimation purpose) is the empirical distribution, we classify divergence measures into two main classes, according to the duality in their formulation: we shall call them form A and form B. Then, we analyze two different situations, viz.: i) the empirical and theoretical distributions are defined on the same set; ii) the support of the empirical distribution is included in the support of the theoretical distribution. In case i) we provide conditions under which divergence measures of the form A or B preserve r-majorization. This is done respectively in section3.2 (form A) and section 3.3 (form B). Nevertheless, in case ii) the conditions for (strong) r-majorizations are not fulfilled, as frequent in many practical situations (such as the case of small samples): thus we extend the study to the more general case of weak r-majorization (section 3.3), providing conditions that a statistical functional must satisfy in order to preserve also the weak pre-order. We find that, among the most popular divergence measures, some of them only

preserve the strong pre-order but are not consistent with respect to the weak one. This is confirmed by empirical results, summarized in section 3.4.

2 Notation and preliminary results

Majorization [12] is a pre-order on vectors aimed to determine whether the components of a vector are more (or less) "spread out" or more (or less) "equal" than the components of another vector. Functions of the vectors that are consistent with this pre-order are called Schur-convex [18]. Majorization and Schur-convexity can be generalized from vectors to integrable functions (see, among others [24], [10]): for this purpose, the definition of the re-arrangement of a function [22] is required.
In this Section we shall refer to a positive measure μ on a set $I \subseteq \Re$.

Definition 1. Let u be a real valued nonnegative μ-integrable function on I. We introduce the notation $u_{\downarrow}^{(\mu)}(t)$ to denote the *decreasing rearrangement* of u with respect to μ, that is

$$u_{\downarrow}^{(\mu)}(t) = \sup\{z \mid a(z) > t\},$$

for every $t \in [0,\mu(I)]$, where $a(z)=\mu(A_z)$, $A_z=\{x \in I \mid u(x) > z\}$.

Note that both $a(z)$ and $u_{\downarrow}^{(\mu)}(t)$ are right continuous functions. Similarly, we define the *increasing rearrangement* of u with respect to μ as:

$$u_{\uparrow}^{(\mu)}(t) = \inf\{z \mid b(z) > t\},$$

for every $t \in [0,\mu(I)]$, where $b(z) = \mu(B_z)$, $B_z=\{x \in I \mid u(x) \leq z\}$. In this case also, $b(z)$ and $u_{\downarrow}^{(\mu)}(t)$ are right continuous functions.

As special cases, the superscript (μ) may be dropped, when it is understood from the context, that is: 1) if μ is the counting measure, for any vector $u = (u_1, u_2, \dots, u_k) \in \Re^k$, $f_i \geq 0$, let denote $u_{\downarrow}$ ($u_{\uparrow}$) the decreasing (increasing) rearrangement of vector u; 2) if μ is the Lebesgue measure, and u is a real valued non-negative μ-integrable function on I, let denote $u_{\downarrow}$ ($u_{\uparrow}$) the decreasing (increasing) rearrangement of u.

Definition 2. Let u, v be real valued nonnegative μ-integrable functions on a set I .We say that v *μ-majorizes* u [3] and we write $u \prec_{(\mu)} v$ if:

i) $\int_0^z u_{\uparrow}^{(\mu)}(t)dt \geq \int_0^z v_{\uparrow}^{(\mu)}(t)dt, \ \forall z \in [0, \mu(I)]$.

ii) $\int_I u d\mu = \int_I v d\mu$.

Alternative definitions of generalized majorization with respect to a measure are given by Joe [10], van Evren and Harremoes [25]. It is worth noting that this definition is related to the concept of second order stochastic dominance [9], the usefulness of stochastic orderings and their applications (e.g. in Finance) have been discussed so fare (see, among others, [20], [21]).
When conditions for μ-majorization are not completely satisfied, there are also weak definitions of μ- majorization. In particular, we will need the following one.

Definition 3. Let u, v be real valued nonnegative be μ-integrable functions on a set I . We say that v *weakly μ-majorizes* u from above and write $u \prec_{(\mu)}^{w} v$ when only i) in Definition 2 holds.

A generalization of Karamata's theorem (independently studied by [14], [11] and proved also in [15]) allows us to characterize a class of functionals which are consistent with respect to strong and weak μ-majorization. This result, proved in Chong [1974, theorem 1.6] and Joe [1987, theorem 2.1] can be summarized as follows.

Theorem 1. Let u, v be real valued nonnegative be μ-integrable functions on a set I.

i) $u \prec_{(\mu)} v$ if and only if: $\int_I \phi(u)d\mu \leq \int_I \phi(v)d\mu$, for all continuous convex functions ϕ.

ii) $u \prec_{(\mu)}^{w} v$ if and only if: $\int_I \phi(u)d\mu \leq \int_I \phi(v)d\mu$, for all decreasing continuous convex functions ϕ.

The theorem says that the functional $\int_I \phi(u)d\mu$ is consistent with strong μ-majorization (ϕ convex) or weak μ-majorization from above (ϕ convex and decreasing). It is worth noting that point i) (*characterization theorem* of μ-majorization) may be interpreted as a definition of the generalized majorization.

3 Minimum divergence and majorization

3.1 Divergence measures

Let $\mathcal{D}$ be a family of discrete probability distributions with a common support, say S; then, $S=\{x_1,x_2,...\}$ is an at most countably infinite set. For every distribution G belonging to $\mathcal{D}$, G is a step function; let g the function defined on the set S as follows: $g(x)=G(x)-G(x^-)$, where $G(x^-)=\lim_{y\uparrow x}G(y)$. Let us write for short $g(x_i)=g_i$ for every $x_i\in S$. Otherwise said, g_i represents the jump of G at the point x_i. There are two cases: 1) $\#S=m<\infty$; 2) $\#S=\infty$. Correspondingly, there are two cases for g: 1) g is a finite vector $g=(g_1,g_2,...,g_m)$; 2) g is an infinite sequence $g=(g_1,g_2,...)$ ([18], p.25).

It is well known that the *empirical distribution function* (*EDF*) is a consistent estimator of $G(x)$. Now, if $X_1,X_2,...,X_n$ are iid with distribution $G\in\mathcal{D}$, let us denote by F the EDF corresponding to the observed sample $x_1,x_2,...,x_n$. More formally: $F(x)=\frac{1}{n}\sum_{i=1}^{n}I_{x_i}(x)$, where $I_a(x)$ is the function defined as $I_a(x)=1$ if $x\le a$, and 0 otherwise. Finally, let S_F be the support of the EDF, and let f be the function defined on the set S as follows: $f(x)=F(x)-F(x^-)$. Otherwise said, $f(x_i)$ represents the empirical relative frequency of the point x_i.

Define μ_G (or μ_F) to be the (probability) measure corresponding to the distribution function G (or F); μ_G and μ_F are finite-valued nonnegative measures with μ_F dominated by μ_G, $\mu_F\ll\mu_G$. Note that, by construction, $S_F\subseteq S$ (in particular, the inclusion is always strict in the case 2), then $f\ge 0$, while g is always greater than 0.

In a nonparametric approach to statistical inference, the estimand is sometimes the distribution G itself. In that case, an estimation problem may be solved with the minimum distance method. The minimum distance estimator for a distribution, with respect to a given "distance" function d, is defined as the distribution (if it exists and it is unique) from $\mathcal{D}$ which is "closest" to F in the sense of the distance d. Here the distance is intended to be a functional $d(P,Q)$ of two distributions P and Q with the following properties: as a function of Q, d attains its minimum d_0 for $P=Q$, and $d(P,Q)>d_0$ for $P\neq Q$ ([5], p.65). It is worth noting that d is not a distance in the strict sense. Indeed, for instance, d is not a symmetrical functional of P and Q. To emphasize this fact, we will refer to $d(P,Q)$ as *divergence* of Q with respect to P (or divergence of Q from P), and we shall speak of *minimum divergence* (*MD*) *estimation method* (see among others [6], [26]). Several divergence measures can be used to construct estimators. Perhaps the most known is the *relative entropy* (also known as Kullback-Leibler divergence, or Kullback-Leibler distance) of Q from P, defined as $\int\ln\frac{p}{q}dP=\int p\ln\frac{p}{q}d\eta$, where P and Q are dominated by a common $\sigma-$finite measure η and where $p=dP/d\eta$ and $q=dQ/d\eta$ represent the corresponding Radon-Nikodym derivatives of P and Q with respect to η.

In our context, reasonable divergence measures can be based on the ratio f/g (or g/f). Indeed, majorization can be interpreted as an ordering of "closeness to uniformity". Then, the main idea is that the more the value of the ratio (f/g or g/f) is close to one, the more F is "similar" to G (or G is "similar" to F).

We can distinguish two different main cases:

A) *Undominated case*. In general, the measure μ_G is not dominated by μ_F, then the ratio function g/f does not generally integrate to one since: $\int_{S_F}g/f\ dF\le\int_S dG=1$.

B) *Dominated case*. As $\mu_F\ll\mu_G$, the function f/g is the Radon-Nikodym derivative dF/dG, then $\int_S f/g\ dG=\int_{S_F}dF=1$

Correspondingly, we distinguish two different families of divergence measures that we shall call functionals of *type-A* and functionals of *type-B* [17], that is

A) $d_A(F,G)=E_F\big(\phi(g/f)\big)=\int_{S_F}\phi(g/f)\,dF=\sum_{S_F}\phi(g/f)f,\ G\in\mathcal{D}$

B) $d_B(G,F)=E_G\big(\phi(f/g)\big)=\int_S\phi(f/g)\,dG=\sum_S\phi(f/g)g,\ G\in\mathcal{D}$

where ϕ is a continuous and convex function. Note that the functional d_A yields a divergence measure of G with respect to F, while the functional d_B yields a divergence measure of F with respect to G, actually d_A and d_B are generally not symmetric. Moreover, for a given convex function ϕ, d_A is

simply the "dual" version of d_B. Many well known estimation methods are based on the minimization of divergence measures which belong to these classes, such as the Kullback-Leibler divergence [16], the Chi square divergence ([4], [13]) and the Hellinger divergence ([2], [23]). However, according to the function ϕ, a divergence measure can be appropriate or inappropriate for estimation purpose. We will discuss it in the sequel.

It is immediate to see under which conditions functionals of type A could be reasonable divergence measures. We consider reasonable divergence measures all the statistical functionals which are consistent with respect to a particular majorization ordering, defined below, which indicates the "relative dissimilarity" between distributions. Actually, as a direct consequence of theorem 1, the following results hold.

Proposition 1.

1) For every pair of distributions G and $H \in \mathcal{D}$, $(g/f) \prec_{(\mu_F)} (h/f) \Leftrightarrow$ $d_A(F,G) \leq d_A(F,H)$ for any convex function ϕ in d_A.
2) For every pair of distributions G and $H \in \mathcal{D}$, $(g/f) \prec^{w}_{(\mu_F)} (h/f) \Leftrightarrow$ $d_A(F,G) \leq d_A(F,H)$ for any decreasing and convex function ϕ in d_A.

Borrowing and extending to the weak case the definition by Joe [10] of *r-majorization* (*relative majorization*), the conditions $(g/f) \prec_{(F)} (h/f)$ and $(g/f) \prec^{w}_{(F)} (h/f)$ could be re-defined as follows.

Definition 4.

We say that G is *r-dominated* by H with respect to F and write $G \precsim_F H$ when $(g/f) \prec_{(F)} (h/f)$ or identically g is *r-majorized* by h with respect to f, $g \prec^r_f h$ [10].

Similarly, we say that G is *weakly r-dominated from above* by H with respect to F and write $G \precsim^w_F H$ when $(g/f) \prec^{w}_{(F)} (h/f)$.

Then proposition 1 can be rephrased as follows.

Proposton 1′.

1) For every pair of distributions G and $H \in \mathcal{D}$, $G \precsim_F H \Leftrightarrow d_A(F,G) \leq d_A(F,H)$ for any convex function ϕ in d_A.
2) For every pair of distributions G and $H \in \mathcal{D}$, $G \precsim^w_F H \Leftrightarrow d_A(F,G) \leq d_A(F,H)$ for any decreasing and convex function ϕ in d_A.

The pre-orders defined by $\precsim_F$ and $\precsim^w_F$ suggest which distribution, between G and H, should be preferred (in terms of their similarity with respect to the EDF F). In particular, the use of the strong or the weak pre-order is justified as follows. In practical estimation contexts, (strong) *r*-dominance can be verified only when $S_F = S$, which presumes that $\#S < \infty$ and that the random sample is large enough to have at least one empirical observation for any point in S. This is a quite restrictive hypothesis. In the more general case, when $S_F \subseteq S$, weak *r*-dominance should be used. Moreover, note that the strong condition $G \precsim_F H$ means that G is closer to F, compared to H. On the other hand, $G \precsim^w_F H$ means that means that G is closer to F, compared to H as well as $\mu_G(X \in S_F) \geq \mu_H(X \in S_F)$: thus, in the weak case, the distribution which maximizes the probability of the set S_F is generally preferred.

So, proposition 1′ establishes under which conditions functionals $d_A(F,G)$ preserve the *r*-majorization pre-order (strong or weak). Nevertheless, as divergence measures of the form B are based on the reciprocals of the ratio g/f, their consistency with respect to a relation of *r*-majorization cannot be immediately derived. Yet, Theorem 3 below establishes that (under certain conditions) functionals of the form B also satisfy the property of consistency with respect to weak *r*-majorization. This depends on the equivalence of weak majorization between the ratio functions g/f and h/f and a sort of (strong) majorization between their reciprocals, as it is proved in the following Theorem 2. Finally, consistency of type-B funtionals with respect to weak *r*-domnance is simply proved in corollary 1.

Theorem 2. For every pair of distributions G and $H \in \mathcal{D}$: $G \precsim^w_F H$ if and only if $(f/g)^{(G)}_{\uparrow} \prec (f/h)^{(H)}_{\uparrow}$.

Proof. It is sufficient to prove that $\int_0^t r^{(F)}_{\uparrow}(z)dz \geq \int_0^t s^{(F)}_{\uparrow}(z)dz \quad , \quad \forall z \in [0,1]$, if and only if $\int_0^t a^{(G)}_{\uparrow}(z)dz \geq \int_0^t b^{(H)}_{\uparrow}(z)dz, \quad \forall z \in [0,1], \quad \forall z \in [0,1]$, where $r = g/f, \ s = h/f, \ a = f/g$ and $b = f/h$.

First of all, observe that a and $a^{(G)}_{\uparrow}$ are equimeasurable:

$$\mu_G\{t \in S : z_1 < a(t) \leq z_2\} =$$
$$= \lambda\{t \in [0,1] : z_1 < a^G_{\uparrow}(t) \leq z_2\}$$

where λ is the Lebesgue measure. The same result holds for b and $b_{\uparrow}^{(H)}$. Thus, considering that $f = 0$ on the set $S - S_F$:

$$\int_0^1 a_{\uparrow}^{(G)}(t)dt = \int_S \frac{f}{g}dG = \int_{S_F} \frac{f}{g}dG = 1 = \int_{S_F} \frac{f}{h}dH = \int_S \frac{f}{h}dH = \int_0^1 b_{\uparrow}^{H}\,dt\,.$$

Similarly we can observe that, by assumption, $1 \geq \mu_G(S_F) \geq \mu_H(S_F)$. Denote by k the cardinality of the set S_F, hence $\nu(S_F) = k$.

We can represent S_F by the set $(w_1, w_2, ..., w_k)$, whose elements are defined by:

$$\frac{g(w_i)}{f(wi)} \geq \frac{g(w_{i-1})}{f(w_{i-1})}, \quad i = 2,...,k;$$

or similarly with the set $(z_1, z_2, ..., z_k)$, defined by:

$$\frac{h(z_j)}{f(z_j)} \geq \frac{h(z_{j-1})}{f(z_{j-1})}, \quad j = 2,...,k\,.$$

Hence, the non-decreasing rearrangements of r and s with respect to μ_F are monotone, non-decreasing and piecewise linear functions on [0,1], in particular:

$$r_{\uparrow}^{(F)} = \frac{g(w_i)}{f(w_i)}$$

$$\text{for } t \in [\sum_{m=1}^{i-1} f(w_m), \sum_{m=1}^{i} f(w_m)], i = 1,..., k;$$

$$s_{\uparrow}^{(F)} = \frac{h(z_j)}{f(z_j)}$$

$$\text{for } t \in [\sum_{m=1}^{j-1} f(z_m), \sum_{m=1}^{j} f(z_m)], \quad j = 1,...,k.$$

where $f(w_0) = f(z_0) = 0$ and obviously $\sum_{m=1}^{k} f(w_m) = \sum_{m=1}^{k} f(z_m) = 1$.

The integral between 0 and x of a monotone, non-decreasing and piecewise linear function is a continuous monotone, strictly-increasing and piecewise linear function on [0,1]. We respectively obtain:

$$U_G(x) = \int_0^x r_{\uparrow}^{(F)}(t)dt = \frac{g(w_i)}{f(w_i)}x,$$

$$\text{for } t \in [\sum_{m=1}^{i-1} f(w_m), \sum_{m=1}^{i} f(w_m)], \; i = 1,..., k.;$$

$$U_H(x) = \int_0^x s_{\uparrow}^{(F)}(t)dt = \frac{h(z_j)}{f(z_j)}x,$$

$$\text{for } t \in [\sum_{m=1}^{j-1} f(z_m), \sum_{m=1}^{j} f(z_m)], \; j = 1,...,k.,$$

where $g(w_0) = h(z_0) = 0$. Remind also that $1 \geq \mu_G(S_F) = G(w_k) \geq \mu_H(S_F) = H(z_k)$.

$U_G(x)$ is invertible on [0,1] and its inverse function $(U_G)^{-1}(y)$ is also continuous, monotone, strictly-increasing and piecewise linear on $[0, \mu_G(S_F)]$ (notice that $\mu_G(S_F) = U_G(1)$):

$$(U_G)^{-1}(y) = \frac{f(w_i)}{g(w_i)}y$$

$$\text{for } y \in [\sum_{m=1}^{i-1} g(w_m), \sum_{m=1}^{i} g(w_m)], \quad j = 1,...,k.$$

Consider the non-increasing re-arrangement of f/g with respect to μ_G. Observe that

$$\frac{g(w_i)}{f(w_i)} \geq \frac{g(w_{i-1})}{f(w_{i-1})} \iff \frac{f(w_{i-1})}{g(w_{i-1})} \geq \frac{f(w_i)}{g(w_i)}, \; i = 2,..., k\,.$$

yields:

$$a_{\downarrow}^{(G)}(t) = \begin{cases} \frac{f(w_i)}{g(w_i)} & t \in [\sum_{m=1}^{i} g(w_m), \sum_{m=1}^{i-1} g(w_m)] \\ 0 & t \in [\mu_G(S_F), 1] \end{cases},$$

$$\text{for } i = 1,..., k,$$

which is monotone, non-increasing and piecewise linear. Similarly:

$$b_{\downarrow}^{(H)}(t) = \begin{cases} \frac{f(w_i)}{h(w_i)} & t \in [\sum_{m=1}^{i} h(z_m), \sum_{m=1}^{i-1} h(z_m)] \\ 0 & t \in [\mu_H(S_F), 1] \end{cases},$$

$$\text{for } i = 1,..., k\,.$$

The integral of $a_{\uparrow}^{(G)}$ between 0 and y gives:

$$V_G(y) = \int_0^y a_{\downarrow}^{(G)}(t)dt =$$

$$= \begin{cases} \frac{f(w_i)}{g(w_i)}y & y \in [\sum_{m=1}^{i-1} g(w_m), \sum_{m=1}^{i} g(w_m)] \\ 1 & y \in [\mu_G(S_F), 1] \end{cases}, \; i = 1,..., k.$$

In other words:

$$V_G(y) = \begin{cases} (U_G)^{-1}(y) & y \in [0, \mu_G(S_F)] \\ 1 & y \in [\mu_G(S_F), 1] \end{cases}.$$

Similarly we obtain:

$$V_H(y)=\begin{cases}(U_H)^{-1}(y) & y\in[0,\mu_H(S_F)]\\ 1 & y\in[\mu_H(S_F),1]\end{cases}.$$

$U_G(x)\geq U_H(x)$, for $x\in[0,1]$, if and only if $(U_G)^{-1}(y)\leq(U_H)^{-1}(y)$, for $y\in [0, \mu_H(S_F)]$, which is equivalent to $V_G(y)\leq V_H(y)$ for $y\in$ [0, 1], as $V_H(y)=1$ for $y\in [\mu_H(S_F),\mu_G(S_F)]$ and $V_G(y)=V_H(y)=1$ for $y\in [\mu_G(S_F),1]$. But since $V_G(1)=V_H(1)=1$ we obtain:

$$\int_0^y a_{\downarrow}^{(G)}(t)dt\leq\int_0^y b_{\downarrow}^{(H)}(t)dt \quad\Leftrightarrow$$

$$\int_0^y a_{\uparrow}^{(G)}(t)\geq\int_0^y b_{\uparrow}^{(H)}(t)dt\,,\ \forall y\in[0,1],$$

which yields the thesis.

Hence, weak *r*-majorization is equivalent to continuous (strong) majorization between reciprocals. We can now prove that, under some conditions, functionals of type-*B* are also "reasonable" divergence measures.

Theorem 3. For every pair of distributions G and $H\in\mathcal{D}$: $G\precsim_F^w H \Leftrightarrow d_B(G,F)\leq d_B(H,F)$ for any function ϕ in d_B which is convex and defined in 0.

Proof. By assumption

$$\int_0^x s_{\uparrow}^{(F)}(t)dt\geq\int_0^x r_{\uparrow}^{(F)}(t)dt,\quad\forall x\in[0,1].$$

Theorem 2 establishes that this condition is equivalent to:

$$\int_0^x \frac{f}{g}^{(\mu_G)}_{\uparrow}dt\geq\int_0^x \frac{f}{h}^{(\mu_H)}_{\uparrow}dt,\quad\forall x\in[0,1],$$

$$\text{where}\int_0^1 \frac{f}{g}^{(\mu_G)}_{\uparrow}dt=\int_0^1 \frac{f}{h}^{(\mu_H)}_{\uparrow}dt.$$

The functions a and $a_{\uparrow}^{(G)}$ are equimeasurable:

$$\mu_G\{t\in S: z_1<a(t)\leq z_2\}=$$
$$=\lambda\{t\in[0,1]: z_1<a_{\uparrow}^{G}(t)\leq z_2\}$$

where λ is the Lebesgue measure. The same result holds for b and $b_{\uparrow}^{(H)}$. Moreover the functions $a_{\uparrow}^{(G)}$ and and $b_{\uparrow}^{(H)}$ are non-decreasing and continuous from the right in [0,1]. They also take value 0 in $[0,1-P_G(S_F)]$.

If ϕ is continuous, convex and defined in 0, the Karamata [14] theorem yields:

$$\int_0^1\phi\left(\frac{f}{g}^{(\mu_G)}_{\uparrow}\right)dt\geq\int_0^1\phi\left(\frac{f}{h}^{(\mu_H)}_{\uparrow}\right)dt.$$

Finally, for equimeasurability, the last inequality is equivalent to:

$$\sum_S\phi\left(\frac{f}{g}\right)g=\sum_{S_F}\phi\left(\frac{f}{g}\right)g+\sum_{S-S_F}\phi(0)g\geq$$
$$\geq\sum_S\phi\left(\frac{f}{h}\right)h=\sum_{S_F}\phi\left(\frac{f}{h}\right)h+\sum_{S-S_F}\phi(0)q$$

which proves the theorem.

Observe that $\sum_{S-S_F} g=1-\mu_G(S_F)$, hence the distribution which maximizes the probability of S_F is generally preferred.

Thus, we conclude that functionals of the Form A (with ϕ convex and decreasing) and Form B (with ϕ convex and defined in 0) are consistent with respect to the weak *r*-majorization pre-order (with respect to the empirical distribution). The following corollary shows that functionals of the form B (with ϕ convex, not necessarily defined in 0, exactly like functionals of the form A) are also consistent when (strong) *r*-majorization holds, besides weak.

Corollary 1.
For every pair of distributions G and $H\in\mathcal{D}$: $G\precsim_F H \Leftrightarrow d_B(G,F)\leq d_B(H,F)$ for any convex function ϕ in d_B.

Proof. The proof can be easily derived from proofs of theorems 2 and 3.

3.4 Applications of some well known divergence measures

To show the usefulness of our classification and properties, we now provide some examples which involve two of the main divergence measures used in statistical estimation: the Kullback-Leibler (KL) divergence and the Chi-square (Chi^2). We obtain the KL or the Chi^2 divergence if we respectively set $\phi(t)=-\ln(t)$ or $\phi(t)=(t-1)^2$ in d_A or d_B. In particular, we have the following four different formulas.

1) *KL form A: Kullback-Leibler divergence*
$$KL_A(F,G)=\sum_{S_F}-f\ln(g/f)$$
2) *KL form B: Reverse Kullback-Leibler divergence* ([8], [19])
$$KL_B(G,F)=\sum_{S}-g\ln(f/g)$$
3) *Chi^2 form A: Neyman modified Chi-Square divergence* ([8], [7])
$$\chi_A(F,G)=\sum_{S_F}(1-g/f)^2 f$$
4) *Chi^2 form B: Chi Square divergence*
$$\chi_B(G,F)=\sum_{S}(1-f/g)^2 g$$

Example 1
A random sample of dimension n=100 generated from a Binomial distribution $Bi(6,0.5)$ yields the following empirical frequencies for the points $x=0,1,...,6$:

$$(f(0),...,f(6))=(\frac{1}{100},\frac{2}{25},\frac{23}{100},\frac{1}{4},\frac{3}{10},\frac{3}{25},\frac{1}{100})$$

which define the empirical distribution F. Now, compare F to $P\sim Bi(6,0.48)$ and $Q\sim Bi(6,0.52)$. Note that $S_F=S=\{x=0,1,..,6\}$, thus we can compare P and Q by strong r-majorization. Note that:

$$(p(0),...,p(6))=$$
$$=(0.0197,0.109,0.252,0.311,0.215,0.0794,0.0122)$$
$$(q(0),...,q(6))=$$
$$=(0.0122,0.0794,0.215,0.311,0.252,0.109,0.0197)$$

The increasing re-arrangement of the ratios p/f (weighted by μ_F) seems to be more "even" and uniform compared to the re-arrangement of q/f, as shown in figure 1.

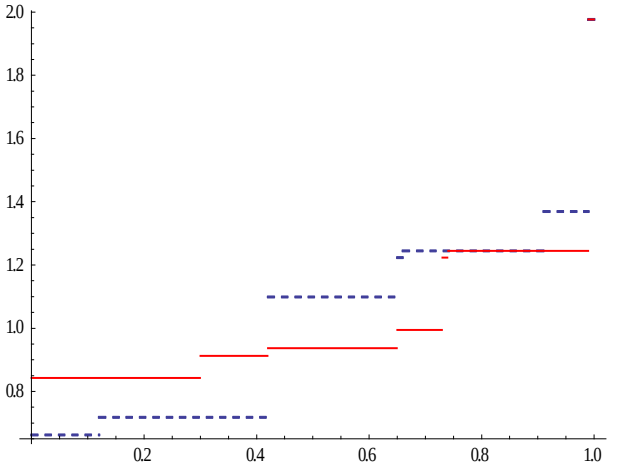

Figure 1: Increasing re-arrangements $\frac{p}{f}^{(\mu_F)\uparrow}$ and $\frac{q}{f}^{(\mu_F)\uparrow}$ (dashed)

Actually, we obtain $P\precsim_F Q$ as
$$U_P(x)=\int_0^x \frac{p}{f}^{(\mu_F)\uparrow}dt\geq\int_0^x \frac{q}{f}^{(\mu_F)\uparrow}=U_Q(x),$$
for $x\in[0,1]$. This is shown in figure 2.

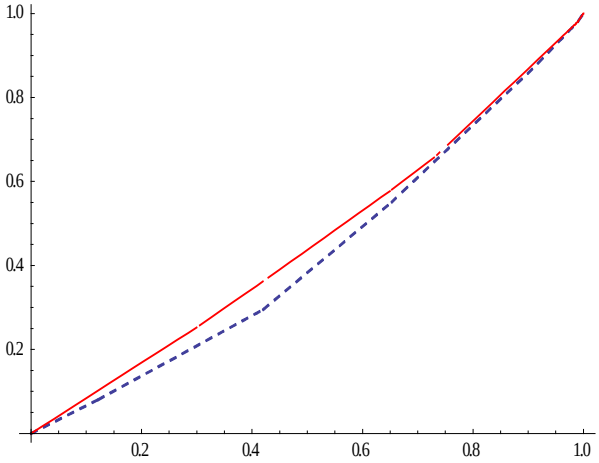

Figure 2: U_P and U_Q (dashed).

It is easy to verify that both the KL and the Chi^2 divergence measures (form A and B) are consistent with the pre-order
$KL_A(F,Q)=0.0387575>0.0147447=KL_A(F,P)$;
$KL_B(Q,F)=0.0378333>0.0157374=KL_B(P,F)$;
$\chi_A(F,Q)=0.0756226>0.0342506=\chi_A(F,P)$;
$\chi_B(F,Q)=0.0811313>0.0280714=\chi_B(F,P)$.
This property holds for theorem 1, point i) (for functionals of the form A) and corollary 1 (form B). Nevertheless, in many practical situations, condition $S_F=S$ is not satisfied. Thus, functional which do not fulfill conditions of theorem 1 point ii) (for functionals of the form A) and theorem 3 (form B)

could provide misleading results, as shown in the following example.

Example 2

A random sample of dimension *n*=20 generated from a Poisson distribution $P(4)$ yields the following empirical frequencies for the points $x=0,1,...,6$:

$$(f(0),..., f(6))=(\frac{1}{20},\frac{3}{20},\frac{1}{20},\frac{1}{4},\frac{7}{20},\frac{1}{20},\frac{1}{10})$$

which define the empirical distribution F. Now, compare F to $P\sim P(3.8)$ and $Q\sim P(4.2)$. Observe that $f(x)=0$ for $x=7,8,...$: actually we have $S_F=\{x=0,1,..,6\}\subset S=\{x=0,1,..\}$, thus we can compare *P* and *Q* only by weak *r*-majorization. Note that $\sum_{x\in S_F} p(x)=0.909108>0.867464=\sum_{x\in S_F} q(x)$.

Moreover we obtain that:

$$U_P(x)=\int_0^x \frac{p}{f}^{(\mu_F)\uparrow} dt \geq \int_0^x \frac{q}{f}^{(\mu_F)\uparrow} = U_Q(x) \text{ for } x\in[0,1]$$

as shown in figure 3: thus $P \lesssim_F^w Q$.

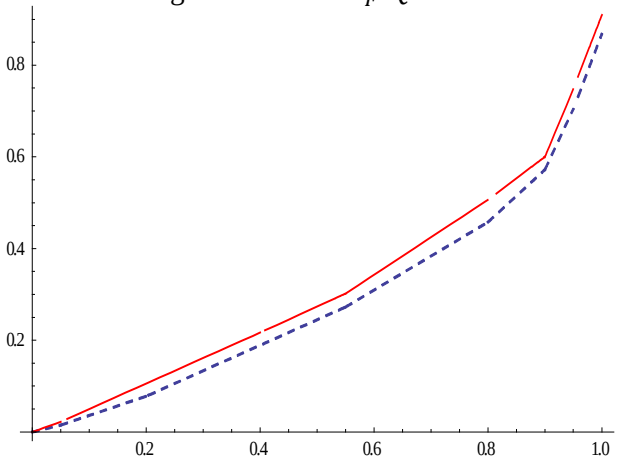

Figure 3: U_P (dashed) and U_Q.

Finally it is possible to verify that only KL (form A) and Chi^2 (form B) preserve the weak majorization pre-order (note that KL form B is not defined when $S_F\neq S$ as $\phi(t)=-\ln(t)$ is not finite in *t*=0).

$KL_A(F,Q)=0.349993>0.275265=KL_A(F,P)$;

$\chi_A(F,Q)=0.555154<0.560979=\chi_A(F,P)$;

$\chi_B(Q,F)=0.61325>0.451497=\chi_B(P,F)$.

We conclude that the choice of χ_A or KL_B in a situation when $S_F\neq S$ can seriously lead to choose the wrong distribution. This is confirmed by a simulation study.

Simulation 1

500 replications of random samples of dimension *n*=10 were generated from a Binomial distribution $Bi(2,0.5)$: for each replication we checked that $S_F=S=\{x=0,1,.2\}$. The MD method, applied to the KL and Chi^2 divergence measures (form A or B) lead in any case to acceptable results in terms of mean squared error (MSE). Actually, note that the minimization of KL (form A) exactly leads to the *maximum likelihood* estimate (MLE) whose optimal properties are well known. From table 1 we observe that MD estimates corresponding to the considered divergence measures (especially Chi^2 form A) are (on average) close to the MLE estimates in terms of MSE, hence it is appropriate to use any of those methods when $S_F=S$.

Simulation 2

500 replications of random samples of dimension *n*=10 were generated from a Binomial distribution $Bi(10,0.5)$: for each replication we surely have that $S_F\subseteq S=\{x=0,1,..,10\}$. As only KL form A and Chi^2 form B preserve the weak *r*-majorization pre-order, other MD methods are not appropriate for this situation, as confirmed by simulation results in terms of MSE. Actually, KL form B is not defined anytime $S_F\neq S$ (for theorem 3) and the MSE of Chi^2 form A is (on average) approximately 24 times the MSE of KL form A (MLE) and Chi^2 form B, which on the other hand are really close.

Average M.S.E.	*KL (A)*	*KL(B)*	*Chi^2 (A)*	*Chi^2 (B)*
Sim.1	0.012	0.02	0.013	0.021
Sim.2	0.00260	----	0.063	0.00265

Table 1: average MSE over 500 replications. As for simulation 1 the four methods provide good/acceptable estimates. As for simulation 2 only KL form A and Chi^2 form B seem to work appropriately.

4 Conclusion

In order to compare theoretical distributions with respect to a reference distribution (represented in our case by the empirical distribution), majorization theory can provide useful guidelines. In particular we use relative majorization [10] to analyze the distance between the theoretical and empirical distributions. As in many practical situations strong *r*-majorization cannot be fulfilled, we propose weak *r*-majorization (from above) as the most appropriate pre-order to compare distributions. Thus we search for a class of statistical functionals which preserve such pre-order. Within the class of divergence measures [1], classified in form A and B, the results can be summarized as follows. Assuming that the

distribution is discrete, divergence measures of the form A are consistent with respect to weak *r*-majorization from above, thus can be appropriately used in estimation, under the condition that ϕ is decreasing and convex. As for divergences of the form B, ϕ has to be convex and defined in 0. The theoretical results of the paper provide practical guidelines for the choice of divergence measures in estimation problem, as confirmed by examples in section 3.4. Although this paper refers to the context of statistical estimation, dissimilarity or distance measures can be applied in many different studies.

Acknowledgments
This paper has been elaborated in the framework of the project Opportunity for young researchers, reg. no. CZ.1.07/2.3.00/30.0016, supported by Operational Programme Education for Competitiveness and co-financed by the European Social Fund and the state budget of the Czech Republic.

References:

[1] Ali, S., M., Silvey, S.,D., A general class of coefficients of divergence of one distribution from another, *Journal of the Royal Statistical Society*, 28 (1), 1966, pp. 131-142.

[2] Beran, R., Minimum Hellinger distance estimates for parametric models, *The Annals of Statistics*, 5(3), 1977, pp. 445-463.

[3] Bertoli-Barsotti, L., *Maggiorazione e Schur-convessità: generalizzazioni e applicazioni statistiche*, Università degli studi di Torino, 2001.

[4] Berkson, J., Minimum Chi-square, not maximum likelihood!, *The annals of Statistics* 8 (3), 1980, pp. 457-487.

[5] Borovkov, A.,A., *Mathematical statistics*, Gordon and Breach science publishers, 1998.

[6] Broniatowskia, B., Leorato, S., An estimation method for the Neyman chi-square divergence with application to test of hypotheses, *Journal of multivariate analysis* 97, 2006, pp. 1409-1436.

[7] Broniatowski, M., Minimum divergence estimators, maximum likelihood and exponential families, arXiv:1108.0772, 2011.

[8] Cressie, N., Read, T.,R.,C., Multinomial goodness-of-fit tests, *Journal of the Royal Statistical society, series B*, 46 (3), 1984, pp. 440,464.

[9] Fishburn, P., C., Stochastic dominance and moments of distributions, *Math, Oper. Res.* 5, 1980, pp. 94-100.

[10] Joe, H., Majorization and Divergence, *Journal of mathematical analysis and applications*, 148 (2), 1990, pp. 287-305.

[11] Hardy, G.,H., Littlewood, J.,E., Pòlya, G., Some simple inequalities satisfied by convex functions, *Messenger of Math.* 58, 1929, pp. 145-152.

[12] Hardy, G.,H., Littlewood, J.,E., Pòlya, G., *Inequalities*, 1934, Cambridge.

[13] Kanji, H.,On the use of Minimum Chi-square estimation, *The Statistician*, 32, 1983, pp. 379-394.

[14] Karamata, J., Sur une inègalitè relative aux functions convexes, *Publ. Math. Univ. Belgrade*, 1, 1932, pp. 145-148.

[15] Karlin, S., Novikoff, A., Generalized convex functions, *Pacific J. Math.*, 13, 1963, pp. 173-180.

[16] Kullback, S., Leibler, R.,A., On information and sufficiency, *The Annals of Mathematical Statistics*,22 (1), 1951, pp. 79–86.

[17] Lando, T. Bertoli-Barsotti, L., Divergence measures and weak majorization in estimations problems, *Advanced in Applied and pure mathematics* (ed.J. Balicki), 2014, pp.152-157.

[18] Marshall, A.,W., Olkin, I., Arnold, B.C. *Inequalities: theory of majorizations and their applications,2nd Edition*, 2011, Springer, New York.

[19] Nielsen, F., Nock, R., On the Chi square and higher-order Chi distances for approximating f-divergences, *IEEE Signal Processing Letters* 21(1), 2013.

[20] Ortobelli, S., Petronio, F., Lando, T., Portfolio problems based on returns Consistent with investor's preferences, *Advanced in Applied and pure mathematics* (ed.J. Balicki), 2014, pp.283-290.

[21] Ortobelli, S.,Petronio, F., Tichy, T., Multivariate stochastic dominance among different financial markets, *Advanced in Appllied and pure mathematics* (ed.J. Balicki), 2014, pag.283-290.

[22] Ryff, J., V., On the representation of doubly stochatic operators, *Pacif. J. Math*, 13, 1963, pp. 1379-1386.

[23] Simpson, D., Minimum Hellinger estimation for the analysis of count data, *Journal of the American Statistical Association*, 82, No 399, 1987 pp. 802-807.

[24] Tong, Y., V., Some recent developments on majorization inequalities in Probability and Statistics, *Linear Algebra and its Applications*, 199, 1994, pp. 69-90.

[25] Van Evren, T., Harremoes, P., Renyi Divergence and Majorization, *Proceedings of the 2010 IEEE International Symposium on Information Theory*, 2010, pp. 1335-1339.

[26] Wolfowitz, J., The minimum distance method, *The Annals of Mathematical Statistics* 28 (1), 1957, pp. 75–88.