DISTANCES BASED ON THE PERIMETER OF THE RISK SET OF A TESTING PROBLEM

Ferdinand Österreicher
Institute of Mathematics, University of Salzburg, Austria
Department of Statistics and Actuarial Science,
University of Stellenbosch, Republic of South Africa (Visiting)

Abstract

At the core of this talk is a simple geometric object, namely the risk set of a statistical testing problem on the one hand and f-divergences, which were introduced by *Csiszár* (1963) on the other. f-divergences are measures for the 'hardness' of a testing problem depending on a convex real valued function f on the interval $[0, \infty)$. The choice of this parameter f can be adjusted so as to match the needs for specific applications.

After presenting the definition, mentioning the basic properties of a risk set and giving the integral geometric representation of f-divergences the talk will focus on the perimeter of the risk set, which has proved useful for the construction of least favourable distributions in robust statistics. The f-divergences based on the perimeter of the risk set and investigated in *Österreicher & Vajda* (2003) turn out to be metric divergences corresponding to a class of entropies introduced by *Arimoto* (1971).

Without essential loss of insight we restrict ourselves to discrete probability distributions and note that the extension to the general case relies strongly on the *Lebesgue-Radon-Nikodym* Theorem.

1 RISK SETS

Let $\Omega = \{x_1, x_2, ...\}$ be a set with at least two elements, $\mathfrak{P}(\Omega)$ the set of all subsets of Ω and $\mathcal{P}$ the set of all probability distributions $P = (p(x) : x \in \Omega)$ on Ω.

A pair $(P, Q) \in \mathcal{P}^2$ of probability distributions is called a *(simple versus simple) testing problem.* A subset $A \subset \Omega$ is called a *(simple) test.* It is associated with the following decision rule: one decides in favour of the hypothesis Q if $x \in A$ is observed and in favour of P if $x \in A^c = \Omega \backslash A$ is observed.

Then $P(A)$ and $Q(A^c)$ is the probability of type I error (probability of a decision in favour of Q although P is true), and the probability of type II error (probability of a decision in favour of P although Q is true) respectively.

Two probability distributions P and Q are called *orthogonal* ($P \perp Q$) if there exists a test $A \subset \Omega$ such that $P(A) = Q(A^c) = 0$. (In this extreme case only one observation is needed to decide between P and Q and the probabilities of committing both errors vanish.)

A testing problem $(P, Q) \in \mathcal{P}^2$ is called *least informative* if $P = Q$ and is called *most informative* if $P \perp Q$.

Let $0 \leq \pi < 1$ and let $(\pi, 1-\pi)$ be a prior distribution on the set $\{P, Q\} \subset \mathcal{P}$ associated with the testing problem (P, Q) . Then the quantity

$$\pi P(A) + (1-\pi)\, Q(A^c)$$

is called *Bayes risk of the test* A *with respect to the prior distribution* $(\pi, 1-\pi)$. Since the Bayes risk enables us to order the pairs $(P(A), Q(A^c))$, $A \in \mathfrak{P}(\Omega)$ of error probabilities, it is straightforward to ask for tests which provide the *minimal Bayes risk*. In fact, as can be easily checked, it holds

$$\begin{aligned} \pi P(A) + (1-\pi)\, Q(A^c) \quad &= \quad \sum_{x \in \Omega} \min\left(\pi p(x), (1-\pi)\, q(x)\right) + \\ &\quad + \sum_{x \in \Omega} \left(\pi p(x) - (1-\pi)\, q(x)\right) 1_{A \cap \{\pi p > (1-\pi) q\}} \\ &\quad + \sum_{x \in \Omega} \left((1-\pi)\, q(x) - \pi p(x)\right) 1_{A^c \cap \{(1-\pi) q > \pi p\}} \end{aligned}$$

where the two latter terms are nonnegative and vanish iff $\{(1-\pi)\, q > \pi p\} \subseteq A \subseteq \{(1-\pi)\, q \geq \pi p\}$.

In order to summarize let $t = \frac{\pi}{1-\pi}$, $A_t = \{q > tp\}$, $A_t^+ = \{q \geq tp\}$ and let $b_t(Q, P) = \sum_{x \in \Omega} \min\left(q(x), tp(x)\right)$ be the $(1+t)$-*multiple of the minimal Bayes risk with respect to the prior distribution* $(\frac{t}{1+t}, \frac{1}{1+t})$. Then

$$Q(A^c) + tP(A) \geq b_t(Q, P) \quad \forall\ A \in \mathfrak{P}(\Omega)$$

with equality iff $A_t \subseteq A \subseteq A_t^+$.

Definition 1: Let $(P, Q) \in \mathcal{P}^2$ be a testing problem. Then the set

$$R(P, Q) = co\left\{\left(P\left(A\right), Q\left(A^c\right)\right) : A \in \mathfrak{P}(\Omega),\ P\left(A\right) + Q\left(A^c\right) \leq 1\right\}$$

is called the *risk set of the testing problem* (P, Q) , whereby 'co' stands for 'the convex hull of'.

The geometric object of the risk set $R(P, Q)$ provides a qualitative measure for the deviation of P and Q . In fact, the family of risk sets define a uniform structure on the set $\mathcal{P}$. Cf. *Linhart & Österreicher* (1985).

Properties of Risk Sets

(R1) $R(P,Q)$ is a convex subset of the triangle $\Delta = \{(\alpha,\beta) \in [0,1]^2 : \alpha+\beta \le 1\}$ containing the diagonal $D = \{(\alpha,\beta) \in [0,1]^2 : \alpha+\beta = 1\}$. More specifically it holds

$$D \subseteq R(P,Q) \subseteq \Delta$$

with equality iff $P = Q$ and $P \perp Q$ respectively.

(R2) Let $t \ge 0$ and $b_t(Q,P)$ be the $(1+t)$-multiple of the minimal Bayes risk with respect to the prior distribution $\left(\frac{t}{1+t}, \frac{1}{1+t}\right)$. Then the risk set $R(P,Q)$ of a testing problem is determined by its family of supporting lines from below, namely

$$\beta = b_t(Q,P) - t \cdot \alpha \ , \quad t \ge 0 \ .$$

Consequence of (R2): Let (P,Q) and $(\tilde{P},\tilde{Q})$ be two testing problems. Then

$$R(P,Q) \supseteq R(\tilde{P},\tilde{Q}) \quad \Leftrightarrow \quad b_t(Q,P) \le b_t(\tilde{Q},\tilde{P}) \ \ \forall\ t \ge 0 \ .$$

Simple Example (Testing a fair tetrahedron versus a biased one):

$$\begin{array}{rcccccc} \Omega & = & \{ & 1, & 2, & 3, & 4 \ \} \\ P & = & (& \frac{1}{4}, & \frac{1}{4}, & \frac{1}{4}, & \frac{1}{4} \ ) \\ Q & = & (& \frac{5}{8}, & \frac{1}{4}, & \frac{1}{8}, & 0 \ ) \end{array}$$

Although the number of simple tests for a set Ω with m elements is $|\mathfrak{P}(\Omega)| = 2^m$ we need only $m+1$ pairs $(P(A), Q(A^c))$, $A \in \mathfrak{P}(\Omega)$ in order to determine the risk set $R(P,Q)$ economically. It is advisable to proceed as follows:

Order the set Ω so that the likelihood ratios are decreasing, i.e.

$$\frac{q(x_1)}{p(x_1)} \ge \frac{q(x_2)}{p(x_2)} \ge \ \ldots \ \ge \frac{q(x_m)}{p(x_m)} \ ,$$

take the tests

$$A_i = \begin{cases} \varnothing & \text{for} \quad i = 0 \\ \{1,\ldots,i\} & \text{for} \quad i \in \{1,\ldots,m\} \end{cases} \ ,$$

assign the set $S = \{\ (P(A_i), Q(A_i^c))\} : i \in \{0,1,\ldots,m\}\ \}$ of the pairs of error probabilities and form the convex hull $co(S)$ of this set. Then $co(S) = R(P,Q)$.

For our example the tests A_i and the corresponding pairs $(P(A_i), Q(A_i^c))$ of error probabilities are given in the following table.

A_i	$(P(A_i), Q(A_i^c))$
$\varnothing$	$(0,1)$
$\{1\}$	$(\frac{1}{4}, \frac{3}{8})$
$\{1,2\}$	$(\frac{1}{2}, \frac{1}{8})$
$\{1,2,3\}$	$(\frac{3}{4}, 0)$
Ω	$(1,0)$

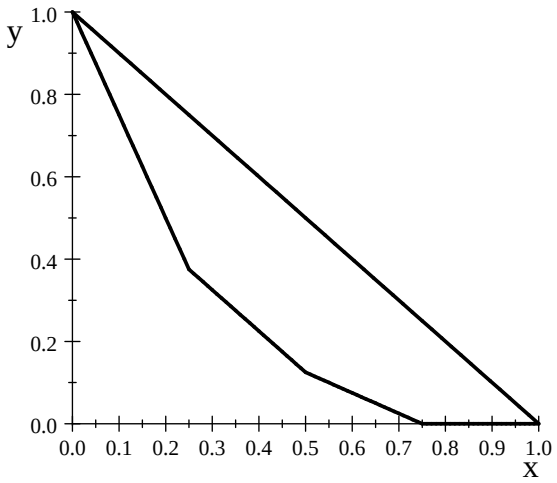

Fig. 1: Risk set of above testing problem

Remark 1: For special case $P = (\frac{1}{m}, ..., \frac{1}{m})$ and $Q = (q_1, ..., q_m)$, such that $q_1 > q_2 > \ ... \ > q_m$, the lower boundary of the set

$$co\,\{(P(A_i), Q(A_i)) : i \in \{0, ..., m\}\ \}\ ,$$

is the so-called *Lorenz curve.* It was already used by *Lorenz* (1905) in order to measure the inequality of the distribution of wealth within a given population. The translation of the following quotation from *Lorenz'* paper into our context describes exactly the purpose of the risk set.

"We wish to be able to say at which point a community is placed between the two extremes, equality on the one hand, and the ownership of all wealth by one individual on the other."

2 f-DIVERGENCES

2.1 GEOMETRIC APPROACH

In order to define a quantity for the 'hardness' of a testing problem (P, Q) we proceed, after the qualitative step which assigns the 'hardness' of a testing problem (P, Q) to the 'bulkiness' of the corresponding risk set $R(P, Q)$, by a first quantitative step.

To this end let $b_t(Q, P)$ be the $(1 + t)$-multiple of the minimal Bayes risk with respect to $(\frac{t}{1+t}, \frac{1}{1+t})$ of the testing problem (P, Q) and let $b_t(P, P) = \min(1, t)$ be the corresponding quantity for the least informative testing problem (P, P) . Then the differences

$$\min(1, t) - b_t(Q, P) \ , \ t \geq 0$$

compare the 'bulkiness' of the risk set $R(P, Q)$ with that of the risk set $R(P, P) = D$ of the least informative testing problem. The parameters $t \geq 0$ are the absolute values of the slopes of the supporting lines of the risk set from below.

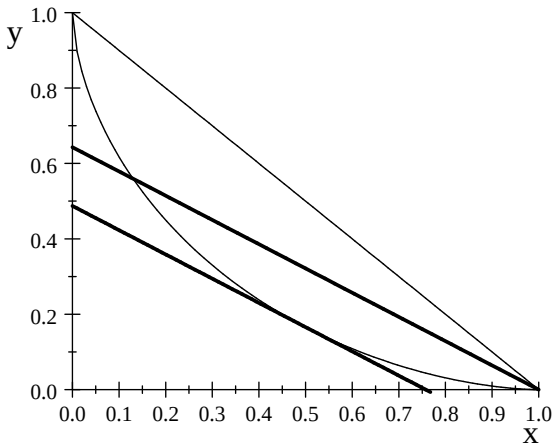

Fig. 2: Differences $\min(1, t) - b_t(Q, P)$

In a second quantitative step weights for the parameters $t \geq 0$ are assigned in terms of a suitable monotone function $F : [0, \infty) \mapsto [-\infty, \infty)$ so that the integral

$$\int_0^\infty [\min(1, t) - b_t(Q, P)] \, dF(t)$$

provides an essential extension of the above family of measures of the 'bulkiness' of the risk set. Due to the richness of the class of parameters F these weighted measures can be adjusted so as to match a given type of application.

The Perimeter of the Risk Set

In this subsection we are going to describe an interesting special case related to the well-known fact from integral geometry that the perimeter of a finite convex subset of $\mathbb{R}^2$ is the integral of its breadths.

Since $\max(1,t)-b_t(Q,P)$ is the vertical part of the the breadth of $R(P,Q)$ in direction of the vector $(\frac{t}{1+t},\frac{1}{1+t})$ of the prior distribution,

$$(\max(1,t)-b_t(Q,P))\cos(\varphi(t)) \quad \text{with} \quad \varphi(t)=\arctan(t)\in[0,\frac{\pi}{2})$$

is its breadth. Since the breadth of the risk set with respect to $\varphi\in[\frac{\pi}{2},\pi)$ is $(\cos(\pi-\varphi)+\sin(\pi-\varphi))$ and $\int_{\frac{\pi}{2}}^{\pi}(\cos(\pi-\varphi)+\sin(\pi-\varphi))\,d\varphi=2$ the perimeter $Per(R(P,Q))$ of the risk set is

$$\begin{aligned} Per(R(P,Q)) &= \int_0^{\pi/2}\left[\max(1,\arctan(\varphi))-b_{\arctan(\varphi)}(Q,P)\right]\cos(\varphi)d\varphi+2= \\ &= \int_0^{\infty}\left[\max(1,t)-b_t(Q,P)\right](1+t^2)^{-3/2}dt+2\ , \end{aligned}$$

whereby, by virtue of $\cos(\varphi(t))\frac{d\varphi(t)}{dt}=(1+t^2)^{-3/2}$, the density $1_{[0,\pi/2)}(\varphi)$ of uniform weight is transformed to the density $(1+t^2)^{-3/2}$ in the parametrization by $t\in[0,\infty)$. Since the perimeter of the risk set $R(P,P)=D$ of the least informative testing problem is obviously

$$Per(R(P,P))=\int_0^{\infty}\left[\max(1,t)-\min(1,t)\right](1+t^2)^{-3/2}+2=2\sqrt{2}$$

the difference

$$Per(R(P,Q))-Per(R(P,P))=\int_0^{\infty}\left[\min(1,t)-b_t(Q,P)\right](1+t^2)^{-3/2}dt$$

is the special case of our family of measures given by the density $(1+t^2)^{-3/2}$.

The above approach to define a family of measures of the 'hardness' of a testing problem, which stresses modelling, relies on the following representation theorem for so-called f-divergences $I_f(Q,P)$ given by *Feldman & Österreicher* (1981). In this setting the weight function F introduced above is the right-hand side derivative D_+f of a continuous convex function f on the interval $[0,\infty)$.

Representation Theorem:

$$I_f(Q,P)=\int_0^{\infty}\left[\min(1,t)-b_t(Q,P)\right]dD_+f(t)\ .$$

In the following section we will present the original definition of f-divergences by *Csiszár* (1963), a number of examples and the basic properties.

2.2 DEFINITION AND BASIC PROPERTIES

Let $\mathcal{F}_0$ be the set of convex functions $f : [0,\infty) \mapsto (-\infty,\infty]$ continuous at 0 (i.e. $f(0) = \lim_{u\downarrow 0} f(u)$) satisfying $f(1) = 0$ and (without loss of generality) $f(u) \geq 0 \ \forall\ u \in [0,\infty)$ and let D_+f denote the *right-hand side derivative* of f . Further, let $f^* \in \mathcal{F}_0$, defined by

$$f^*(u) = uf\left(\frac{1}{u}\right) \ , \ u \in (0,\infty) \ ,$$

the $*$-*conjugate* (convex) function of f and let a function $f \in \mathcal{F}$ satisfying $f^* \equiv f$ be called $*$-*self conjugate*. Then

$$\begin{array}{lcllcll} x \cdot f(0) & = & x \cdot f\left(\frac{0}{x}\right) & = & 0 \cdot f^*\left(\frac{x}{0}\right) & \text{for} & x \in (0,\infty) \\ y \cdot f^*(0) & = & y \cdot f^*\left(\frac{0}{y}\right) & = & 0 \cdot f(\frac{y}{0}) & \text{for} & y \in (0,\infty) \\ 0 \cdot f\left(\frac{0}{0}\right) & = & 0 \cdot f^*\left(\frac{0}{0}\right) & = & 0 \ . & & \end{array}$$

Definition 2 (*Cziszár* (1963), *Ali & Silvey* (1966)): Let $P, Q \in \mathcal{P}$. Then

$$I_f(Q,P) = \sum_{x\in\Omega} p(x) f\left(\frac{q(x)}{p(x)}\right)$$

is called the f-*Divergence* of the probability distributions Q and P .

Examples: Total Variation Distance ($f(u) = |u-1| = f^*(u)$)

$$I_f(Q,P) = V(Q,P) = \sum_{x\in\Omega} |q(x) - p(x)|$$

Squared Hellinger Distance ($f(u) = (\sqrt{u}-1)^2 = f^*(u)$)

$$I_f(Q,P) = H^2(Q,P) = \sum_{x\in\Omega} \left(\sqrt{q(x)} - \sqrt{p(x)}\right)^2$$

χ^2-Divergence ($f(u) = (u-1)^2$, $f^*(u) = \frac{(1-u)^2}{u}$)

$$I_f(Q,P) = \sum_{x\in\Omega} \frac{(q(x)-p(x))^2}{p(x)} = I_f^*(P,Q)$$

Kullback-Leibler Divergence ($f(u) = u\ln(u)$, $f^*(u) = -\ln(u)$)

$$I_f(Q,P) = \sum_{x\in\Omega} q(x) \ln\left(\frac{q(x)}{p(x)}\right) = I_f^*(P,Q)$$

Squared Perimeter Distance ($f(u) = \sqrt{1+u^2} - (1+u)/\sqrt{2} = f^*(u)$)

$$I_f(Q,P) = \sum_{x\in\Omega} \sqrt{p^2(x)+q^2(x)} - \sqrt{2}$$

Remark 2: Note that

$$\begin{aligned} I_f(Q,P) &= f(0)\cdot P(\{x : q(x)=0\}) + f^*(0)\cdot Q(\{x : p(x)=0\}) + \\ &+ \sum_{x:\ q(x)\cdot p(x)>0} p(x) f\left(\frac{q(x)}{p(x)}\right) \end{aligned}$$

and that $P(\{x : q(x)=0\})$ is the amount of singularity of the distribution P with respect to Q and $Q(\{x : p(x)=0\})$ is the amount of singularity of the distribution Q with respect to P . Therefore $f(0)=\infty$ and $f^*(0)=\infty$ imply $I_f(Q,P)=\infty$ unless $\{x\in\Omega :\ q(x)\cdot p(x)>0\} = \Omega$, i.e. all probabilities are positive.

Range of Values Theorem (*Vajda* (1972))**:** Let $f\in\mathcal{F}_0$. Then

$$0 \le I_f(Q,P) \le f(0)+f^*(0) \qquad \forall\ Q,P\in\mathcal{P}\ .$$

In the first inequality, equality holds if / iff $Q=P$. The latter provided that

(i) f is strictly convex at 1 .

In the second, equality holds if / iff $Q\perp P$. The latter provided that

(iii) $f(0)+f^*(0)<\infty$.

Characterization Theorem (*Csiszár*, 1974)**:** Given a mapping $I:\mathcal{P}^2\mapsto(-\infty,\infty]$ then the following two statements are equivalent

($*$) I is an f-divergence

i.e. there exists an $f\in\mathcal{F}_0$ such that $I(Q,P)=I_f(Q,P)\ \ \forall\ (P,Q)\in\mathcal{P}^2$

($**$) I satisfies the following three properties.

(a) $I(Q,P)$ is invariant under permutation of Ω ,

(b) Let $\mathcal{A}=(A_i,\ i\ge 1)$ be a partition of Ω and let

$$P_{\mathcal{A}} = (P(A_i),\ i\ge 1) \quad\text{and}\quad Q_{\mathcal{A}} = (Q(A_i),\ i\ge 1)$$

be the restrictions of the probability distributions P and Q to $\mathcal{A}$. Then

$$I(Q,P) \ge I(Q_{\mathcal{A}}, P_{\mathcal{A}})$$

with equality holding if $Q(A_i)\times p(x) = P(A_i)\times q(x)\ \ \forall\ x\in A_i$, $i\ge 1$ and

(c) Let $\alpha\in[0,1]$ and P_1,P_2 and Q_1,Q_2 probability distributions on Ω . Then

$$I(\alpha P_1+(1-\alpha)P_2, \alpha Q_1+(1-\alpha)Q_2) \le \alpha I(P_1,Q_1)+(1-\alpha)I(P_2,Q_2)\ .$$

2.3 METRIC f-DIVERGENCES

Let us now concentrate on those (further) properties of the convex function f which allows for metric divergences.

As we know already $I_f(Q,P)$ fulfils the basic property (M1) of a metric divergence, namely

$$I_f(Q,P) \geq 0 \quad \forall\ P,Q \in \mathcal{P} \qquad \text{with equality iff} \qquad Q = P\ , \tag{M1}$$

provided (i) f is strictly convex at 1 .

In addition $I_f(Q,P)$ is symmetric, i.e. satisfies

$$I_f(Q,P) = I_f(P,Q) \quad \forall\ P,Q \in \mathcal{P} \tag{M2}$$

iff (ii) f is $*$-self conjugate, i.e. satisfies $f \equiv f^*$.

It turns out that, in addition to the rather natural conditions (i) and (ii), the condition (iii) $f(0) + f^*(0) < \infty$, which is used to characterize $Q \perp P$, is crucial for metric divergences. However, since it cannot be expected in general that an f-divergence fulfils the triangle inequality we have to look for suitable powers to do so.

From the following two theorems given in *Kafka, Österreicher & Vincze* (1991) Theorem 4 offers a class (iii,α), $\alpha \in (0,1]$ of conditions which are sufficient for guaranteeing the power $[I_f(Q,P)]^\alpha$ to be a distance on $\mathcal{P}$. Theorem 5 determines, in dependence of the behaviour of f in the neighbourhoods of 1 and of $g(u) = f(0)(1+u) - f(u)$ in the neighbourhood of 0 , the maximal α providing a distance.

Theorem 4: Let $\alpha \in (0,1]$ and let $f \in \mathcal{F}_0$ fulfil, in addition to (ii), the condition

(iii,α) the function $h(u) = \frac{(1-u^\alpha)^{\frac{1}{\alpha}}}{f(u)}$, $u \in [0,1)$, is non-increasing.

Then

$$\rho_\alpha(Q,P) = [I_f(Q,P)]^\alpha$$

satisfies the triangle inequality

$$\rho_\alpha(Q,P) \leq \rho_\alpha(Q,R) + \rho_\alpha(R,P) \quad \forall\ P,Q,R \in \mathcal{P}\ , \tag{M3,α}$$

which effects, together with (M1) and (M2), that ρ_α is a metric.

Remark 3: The conditions (ii) and (iii,α) imply both (i) and (iii).

Theorem 5: Let (i) and (ii) hold true and let $\alpha_0 \in (0,1]$ be the maximal α for which (iii,α) is satisfied. Then the following statement concerning α_0 holds. If for some k_0, k_1, c_0, $c_1 \in (0,\infty)$

$$\begin{aligned} f(0)\cdot(1+u) - f(u) &\sim c_0 \cdot u^{k_0} \\ f(u) &\sim c_1 \cdot |u-1|^{k_1} \end{aligned}$$

then $k_0 \leq 1,\ k_1 \geq 1$ and $\alpha_0 \leq \min(k_0, 1/k_1) \leq 1$.

Finally we present a version of the refinement of the Range of Values Theorem which matches the assumptions (i), (ii) and (iii) which are necessary to allow for metric divergences.

Refinement of the Range of Values Theorem (*Feldman & Österreicher* (1989)**:** Let $f \in \mathcal{F}_0$ satisfy the conditions (i),(ii) and (iii), $x \in [0,1]$ and let the function $c_f : [0,1] \mapsto [0,\infty)$ be defined by

$$c_f(x) = (1+x)\, f\left(\frac{1-x}{1+x}\right) .$$

Then

$$c_f\left(V(Q,P)/2\right) \leq I_f\left(Q,P\right) \leq c_f(1) \cdot V(Q,P)/2,$$

where c_f satisfies $c_f(0) = 0$ and $c_f(1) = 2f(0) < \infty$ and is convex, strictly increasing and continuous on $[0,1]$.

Remark 4: Note that this theorem implies that any metric defined in terms of an f-divergence is equivalent to the total variation distance.

3 CONSTRUCTION OF LEAST FAVOURABLE DISTRIBUTIONS

Huber & Strassen (1973) proved the existence of least favourable pairs of distributions for composite versus composite testing problems under the assumption that both hypotheses are majorized by two-alternating capacities and characterized them in terms of f-divergences with strict convex functions f . The author restated the definition of least favourable pairs in terms of risk sets and demonstrated (1982) that their perimeter can be used to construct least favourable pairs. For further references in this context see e.g. *Österreicher* (1983).

For an application of the perimeter of the risk set for goodness of fit tests see *Reschenhofer & Bomze* (1991).

Definition 3: Let

$$R(P,\mathcal{Q}) = \cap_{Q' \in \mathcal{Q}} R(P,Q')$$

be the risk set of a simple versus composite testing problem, which is a pair $(P,\mathcal{Q})$ of an element P and a nontrivial subset $\mathcal{Q}$ of $\mathcal{P}$.

We will illustrate the construction of a least favourable distribution $Q^* \in \mathcal{Q}$ for the simple case

$$\begin{aligned} \mathcal{Q} &= U(Q,\varepsilon) = \{\ Q' \in \mathcal{P} : V(Q,Q')/2 \leq \varepsilon\ \} \\ &= \{\ Q' \in \mathcal{P} : Q'(A) \leq Q(A) + \varepsilon \ \ \forall\ A \in \mathfrak{P}(\Omega)\ \} \end{aligned}$$

of a total variation neighbourhood.

Theorem 7: Let $P, Q \in \mathcal{P}$ and let $\mathcal{Q} = U(Q, \varepsilon)$, $\varepsilon \in (0,1)$ be a total variation neighbourhood of Q which does not contain P . Let furthermore $R(P,Q)+(0,\varepsilon)$ be the risk set of the simple versus simple testing problem (P,Q) having been shifted upwards by the amount ε and let finally $\underline{t} < 1 < \bar{t}$ be the absolute values of the slopes of the supporting lines onto $R(P,Q)+(0,\varepsilon)$ through the points $(1,0)$ and $(1,0)$ respectively.

Then the least favourable distribution $Q^* \in \mathcal{Q}$ for $(P, U(Q,\varepsilon))$ is given by the censored version

$$q^*(x) = \max\left(\underline{t} \cdot p(x), \min(q(x), \bar{t} \cdot p(x)) \right)$$

of the density q .

Simple Example (Continuation): In order to illustrate Theorem 7 let us continue our simple example from Section 1 by replacing the distribution Q by the total variation neighbourhood

$$\mathcal{Q} = U(Q, \frac{1}{8}) = \left\{ Q' \in \mathcal{P} : Q'(A) \leq Q(A) + \frac{1}{8} \ \ \forall\ A \in \mathfrak{P}(\Omega) \right\} .$$

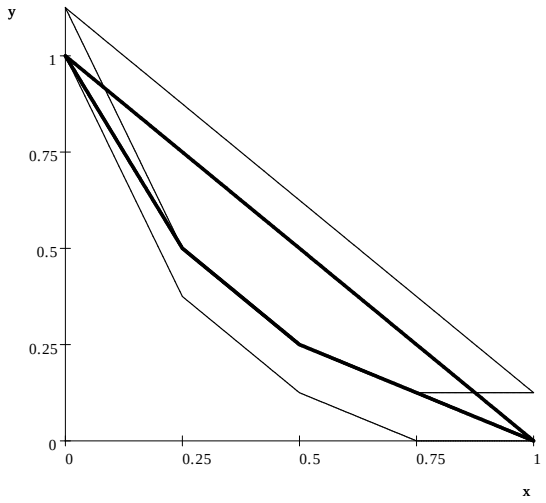

Fig. 3: Modification of the risk set

When comparing the distribution Q in the center of the variation neighborhood $\mathcal{Q} = U(Q, \frac{1}{8})$ with the least favourable distribution $Q^* \in \mathcal{Q}$

$$\begin{array}{rcl} Q & = & (\ \frac{5}{8}, \ \frac{1}{4}, \ \frac{1}{8}, \ 0 \ ) \\ Q^* & = & (\ \frac{4}{8}, \ \frac{1}{4}, \ \frac{1}{8}, \ \frac{1}{8} \ ) \end{array}$$

notice that the probability $\frac{1}{8}$ is shifted from the most probable element to the least probable.

Remark 5: For the special case $\Omega = \{1, ..., n\}$, $P = \left(\frac{1}{n}, ..., \frac{1}{n}\right)$ and $Q = (q_1, ..., q_n)$ the above theorem has the following econometric interpretation.

If the distribution Q of income (with total amount 1) of a population of n individuals has to be redistributed so that the inequality in income is minimized under the constraint that the portion of income of no group of the population is cut or raised more than ε , one has to proceed as follows: If a person's income exceeds a certain amount $\bar{t}/n$, her or his income has to be cut to this bound. The total amount ε of income collected that way must be allotted to those persons, whose income is smaller than a certain lower bound $\underline{t}/n$ so that every person is guaranteed the minimal income $\underline{t}/n$.

The principle of income transfer was first clearly described by Dalton (1920) as follows:

"If there are only two income receivers and a transfer of income takes place from the richer to the poorer, inequality is diminished. There is, indeed, an obvious limiting condition. The transfer must be so large as to more than reverse the relative position of the two income receivers, and it will produce its maximum result, that is to say, create equality, when it is equal to the half of difference between the two incomes. And, we may safely go further and say that, however great the number of income receivers and whatever the amount of their incomes, any transfer of the two of them or, in general, any series of such transfers, subject to the above condition, will diminish inequality. It is possible that, in comparing two distributions, in which both the total of the income of the number of the income receivers are the same, we may see that one might be able to be evolved from the other by means of a series of transfers of this kind. In such a case we would say that the inequality of one was less than that of another."

4 DIVERGENCES OF PERIMETER-TYPE

If both the arc length of the lower boundary of the risk set and the diagonal D are measured in terms of the l_p-norm in $\mathbb{R}^2$ then the ordinary case ($p = 2$) can be extended to the perimeter-type family

$$I_{f_p}(Q, P) = \begin{cases} \sum_{x \in \Omega} \left[q^p(x) + p^p(x)\right]^{1/p} - 2^{1/p} & \text{for} \quad p \in (1, \infty) \\ \frac{1}{2} \sum_{x \in \Omega} |q(x) - p(x)| & \text{for} \quad p = \infty \end{cases}$$

(cf. *Österreicher* (1996)). In taking the $(1 - 1/p)$-th part of the corresponding convex function $(1 + u^p)^{1/p} - 2^{1/p-1}(1 + u)$ we make a second step of gener-

alization yielding the family of f-divergences defined by the convex functions

$$f_p(u) = \begin{cases} \frac{1}{1-1/p}\left[(1+u^p)^{1/p} - 2^{1/p-1}(1+u)\right] & \text{if} \quad p \in (0,\infty)\backslash\{1\} \\ (1+u)\ln(2) + u\ln(u) - (1+u)\ln(1+u) & \text{if} \quad p = 1 \\ |u-1|/2 & \text{if} \quad p = \infty\,, \end{cases}$$

where both cases $p = 1$ and $p = \infty$ are limiting cases. As a matter of fact, this family relates due to

$$f_p(u) = (1+u)\left[h_{1/p}(1/2) - h_{1/p}(u/(1+u))\right],\ u \in [0,\infty),$$

to the class of entropies investigated by *Arimoto* (1971)

$$h_\alpha(t) = \begin{cases} \frac{1}{1-\alpha}\left[1 - \left(t^{1/\alpha} + (1-t)^{1/\alpha}\right)^\alpha\right] & \text{if} \quad \alpha \in (0,\infty)\backslash\{1\} \\ -\left[t\ln t + (1-t)\ln(1-t)\right] & \text{if} \quad \alpha = 1 \\ \min(t, 1-t) & \text{if} \quad \alpha = 0\,. \end{cases}$$

Note that our class of f-divergences includes, in addition to the case for $p > 1$ already discussed for the case $p = \frac{1}{2}$ ($f_{1/2}(u) = (\sqrt{u}-1)^2$), the squared *Hellinger* distance $H^2(Q,P)$ and for $p = 1$

$$\begin{aligned} I_{f_1}(Q,P) &= I(Q, \frac{P+Q}{2}) + I(P, \frac{P+Q}{2}) \\ &= 2H(\frac{P+Q}{2}) - \left[H(P) + H(Q)\right], \end{aligned}$$

where I and H is the classical *Kullback-Leibler* divergence (f-divergence for $f(u) = u\ln u$), respectively *Shannon's* entropy.

Theorem 8 (*Österreicher & Vajda* (2003)): This class of f-divergence provides the distances

$$\left[I_{f_p}(Q,P)\right]^{\min(p,\frac{1}{2})} \quad \text{for} \quad p \in (0,\infty) \qquad \text{and} \qquad V(Q,P)/2 \quad \text{for} \quad p = \infty\,.$$

For further results, including those in connection with possible applications, we refer to the paper mentioned above.

References

Arimoto, S. (1971) : Information-theoretical considerations on estimation problems. Information and Control, **19**, 181-194.

Csiszár, I.: Information measures: A critical survey. In: *Trans. 7^{th} Prague Conf. on Information Theory*, Academia Prague 1974, Vol. A, 73-86.

Dalton, H. (1920): The measurement of the inequality of incomes. *Economic. J.*, **30**, 348-361.

Feldman, D. and Österreicher, F. (1981): Divergenzen von Wahrscheinlichkeitsverteilungen – integralgeometrisch betrachtet. *Acta Math. Acad. Sci. Hungar.*, **37**/4, 329–337.

Feldman, D. and Österreicher, F. (1989): A note on f-divergences. *Studia Sci. Math. Hungar.*, **24**, 191–200.

Huber, P.J. and Strassen, V. (1973): Minimax tests and Neyman-Pearson lemma for capacities. *Ann. Statist.*, **1**, 251-263.

Kafka, P., Österreicher, F. and Vincze, I. (1991): On powers of f-divergences defining a distance. *Studia Sci. Math. Hungar.*, **26**, 415–422.

Liese, F. and Vajda, I.: Convex Statistical Distances. Teubner–Texte zur Mathematik, Band **95**, Leipzig 1987

Linhart, J. and Österreicher, F. (1985): Uniformity and distance - a vivid example form statistics. *Int. J. Edu. Sci. Technol.,* **16**/5, 645-649.

Lorenz, M.O. (1905): Methods of measuring concentration of wealth. *J. Amer. Statist. Assoc.*, **9**, 209-219.

Österreicher, F. and Thaler, M. (1978): The fundamental Neyman-Pearson lemma and the Radon-Nikodym theorem from a common statistical point of view. *Int. J. Edu. Sci. Technol.,* **9**, 163-176.

Österreicher, F. (1983): Least favourable distributions. *Entry from Kotz-Johnson: Encyclopedia of Statistical Sciences*, Vol. 3, 588-592, John Wiley & Sons, New York

Österreicher, F. (1992): The risk set of a testing problem – A vivid statistical tool. In: *Trans. of the 11^{th} Prague Conference on Information Theory*, Academia, Prague, Vol. A, 175–188.

Österreicher, F. and Vajda, I. (1993): Statistical information and discrimination. *IEEE Trans. Inform. Theory*, **39**, 3, 1036–1039.

Österreicher, F. (1996): On a class of perimeter-type distances of probability distributions. *Kybernetika*, **32**, 389–393.

Österreicher, F. and Vajda, I. (2003): A new class of metric divergences on probability spaces and its applicability in statistics. *Ann. Inst. Statist. Math.*, Vol. 55, No. 3, 639-653.

Reschenhofer, E. and Bomze, I. M. (1991) Lengths tests for goodness of fit. *Biometrika* **78**, 207–216.

Vajda, I. (1972): On f-divergence and and singularity of probability measures. *Period. Math. Hungar.*, **2**, 223-234.