AUSTRIAN JOURNAL OF STATISTICS
Volume 42 (2013), Number 1, 3–19

Distances Based on the Perimeter of the Risk Set of a Testing Problem

Ferdinand Österreicher[1]
Institute of Mathematics, University of Salzburg, Austria

Abstract: At the core of this paper 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 hand. 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.

One of these adjustments of the parameter f is exemplified in Section 3 of this paper. There it is illustrated that the appropriate choice of f for the construction of least favourable distributions in robust statistics is the convex function $f(u) = \sqrt{1+u^2} - (1+u)/\sqrt{2}$ yielding the perimeter of the risk set of a testing problem.

After presenting the definition, mentioning the basic properties of a risk set and giving the integral geometric representation of f-divergences the paper will focus on the perimeter of the risk set.

All members of the class of f-divergences of perimeter-type introduced and investigated in Österreicher and Vajda (2003) and Vajda (2009) 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.

Zusammenfassung: Den Kern dieses Artikels bilden einerseits ein einfaches geometrisches Objekt, nämlich die Risikomenge eines statistischen Testproblems, und andererseits die von Csiszár (1963) eingeführten f-Divergenzen. Letztere sind Größen, welche die *Schwierigkeit* eines Testproblems messen und die durch eine konvexe reellwertige Funktion f auf dem Intervall $[0, \infty)$ parametrisiert sind. Die Wahl des Parameters f kann den Bedürfnissen spezifischer Anwendungen angepasst werden.

Eine von diesen Anpassungen des Parameters f wird in Abschnitt 3 dieses Artikels beschrieben. In diesem wird nämlich illustriert, dass es für die Konstruktion von ungünstigsten Verteilungen in der robusten Statistik zweckmäßig ist, als Parameter die konvexe Funktion $f(u) = \sqrt{1+u^2} - (1+u)/\sqrt{2}$ zu wählen, welche den Umfang der Risikomenge des Testproblems liefert.

[1] Dedicated to the Memory of Igor Vajda (1942-2010)

Nachdem Definition und grundlegende Eigenschaften der Risikomenge eines Testproblems gegeben und die integralgeometrische Darstellung von f-Divergenzen präsentiert werden, konzentriert sich der vorliegende Artikel auf den Umfang der Risikomenge.

Alle Elemente der Klasse von f-Divergenzen vom Umfangstyp, welche in den Arbeiten von Österreicher and Vajda (2003) und Vajda (2009) eingeführt und untersucht werden, stellen sich als metrische Divergenzen heraus, die einer von Arimoto (1971) eingeführten Familie von Entropien entsprechen.

Ohne Verlust von Einsicht beschränken wir uns hier auf diskrete Wahrscheinlichkeitsverteilungen und merken an, dass die Fortsetzung auf den allgemeinen Fall auf dem Satz von Lebesque-Radon-Nikodym beruht.

Keywords: Testing Problem, Risk Set, Dissimilarity Measure (f-divergence), Integral-geometric Representation, Least Favourable Distribution, Metric Divergence.

1 Introduction

What is basic for this paper is a (simple versus simple) testing problem (P, Q), which is a pair of probability distributions P and Q defined on a set $\Omega = \{x_1, x_2, \dots\}$ of at least two elements.

In Section 2 the central entity of this paper, namely the risk set of a testing problem, and its properties are presented:

Let $A \subseteq \Omega$ be a (nonrandomized) test. Then the convex hull of all pairs $(P(A), Q(A^c))$, $A \subseteq \Omega$, of the probabilities $P(A)$ and $Q(A^c)$ of type I and type II error satisfying $P(A) + Q(A^c) \leq 1$ is the risk set $R(P, Q)$ of the testing problem (P, Q). Expressed colloquially, its essence may be summarized as follows: The 'bulkier' its risk set the 'easier' the testing problem. Section 3 is devoted to f-divergences. Subsection 3.3 contains their precise definition and their basic properties.

Motivation 1: The most widely used measure of the deviation of two probability distributions in statistics is Pearson's χ^2-divergence

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

Another well-known measure of deviation, originally designed for applications in cryptanalysis and later used both in information theory and statistics is the I- or Kullback-Leibler divergence

$$I(Q \parallel P) = \sum_{x \in \Omega} q(x) \log \frac{q(x)}{p(x)} .$$

Another measure of deviation is the total variation distance

$$||Q - P||/2 = \frac{1}{2} \sum_{x \in \Omega} |q(x) - p(x)| .$$

These and many other measures of deviation of two probability distributions are special cases of f-divergences

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

defined in terms of a convex function $f : [0,\infty) \mapsto \mathbb{R}$ continuous at 0 and introduced by Csiszár (1963). So χ^2-, I-divergence and total variation distance are the f-divergences with the convex function $f(u) = (u-1)^2$, $f(u) = u\log u$ and $f(u) = |u-1|/2$, respectively.

Subsections 3.1 and 3.2 deliver the integral-geometric approach to f-divergences which is based on the risk set of a testing problem:

The 'Representation Theorem', the main result of this subsection, states that every f-divergence is a certain way of measuring the 'bulkiness' of the risk set. The most natural measure of its 'bulkiness' is perhaps its perimeter: It turns out that the latter is the f-divergence given by the convex function

$$f(u) = \sqrt{1+u^2} - (1+u)/\sqrt{2}\,. \tag{1}$$

More generally, the 'parameter' f of an f-divergence is nothing but a certain function for the weights of the breadths of the corresponding risk set measured for all different directions.

By the way, the area of the risk set — the well-known Gini coefficient — is not an f-divergence. Subsection 3.4 is entitled 'Metric f-Divergences'.

Motivation 2: Pearson's χ^2-divergence $\chi^2(Q,P)$ is obviously not apt to define a metric divergence. However, the square root of its symmetrized version

$$\chi^2(Q,(P+Q)/2) = \frac{1}{2}\sum_{x\in\Omega} \frac{(q(x)-p(x))^2}{p(x)+q(x)}$$

is a metric. This divergence, which is defined in terms of the convex function $f(u) = \frac{(u-1)^2}{2(1+u)}$ and which goes back to Sanghvi (1953) is studied in detail by Puri and Vincze (1988). Thus, from the mathematical point of view it is very natural to ask which properties of a convex function f are sufficient for an f-divergence to be a metric divergence. Theorems 4 and 5 answer this question.

Remark 1: Metric divergences do not only fulfil the symmetry property

$$I_f(Q,P) \equiv I_f(P,Q)\,,$$

i.e. that f is $*$-self conjugate, but also the property

$$I_f(Q,P) \le 2f(0) < \infty \qquad \text{with equality if and only if } \; P \perp Q\,,$$

whereby $f(0) < \infty$ is crucial for

- establishing bounds for the error probabilities of a sequence (P^n, Q^n), $n \in \mathbb{N}$, of testing problems (i.e. in case of n iid observations with distribution P and Q, respectively) and
- the characterization of an entirely separated sequence P_n, Q_n, $n \in \mathbb{N}$, of probability distributions (cf. e.g. Chapter 11. Various statistical applications of f-divergence in Vajda, 1989).

Section 4 is devoted to robust testing: We are given a simple versus composite testing problem $(P, \mathcal{Q})$, where

$$\mathcal{Q} = \{Q' : ||Q' - Q||/2 \leq \varepsilon\}$$

is the set of all probability distribution with total variation distance $\leq \varepsilon$ from a given probability distribution Q and let $P \notin \mathcal{Q}$. A least favourable distribution is a distribution $Q^* \in \mathcal{Q}$, which is 'closest' to P. In early papers on robust testing least favourable distributions were characterized by f-divergences. In this section, we show how to construct a least favourable distribution Q^*. This is done in geometric terms: We construct the intersection of risk sets

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

and select the element $Q^* \in \mathcal{Q}$ which satisfies $R(P, Q^*) = R(P, \mathcal{Q})$. Therefore, in order to construct least favourable distributions the appropriate choice of the convex function f of the f-divergence is the one which gives rise to the perimeter of the risk set of the testing problem, i.e. the one given by (1). f-divergences are obviously used in several areas of statistics, furthermore in proving limit theorems in probability theory, in analyzing the limiting behavior of Markov chains, in information theory and quantum physics.

Section 5 is devoted to the class of f-divergences of perimeter type, introduced and studied in Österreicher and Vajda (2003) and Vajda (2009). It is based on the class of entropies due to Arimoto (1971) and contains, next to the f-divergence given by (1), the total variation distance and a symmetrized version of the I-divergence also the squared Hellinger distance (with $f(u) = (\sqrt{u} - 1)^2$) and the squared Puri-Vincze distance. All f-divergences of this class are metric divergences.

2 Risk Sets

Let $\Omega = \{x_1, x_2, \dots\}$ 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

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

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 $b_t(Q,P) = \sum_{x\in\Omega} \min(q(x), tp(x))$ 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

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

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 and Österreicher, 1985).

2.1 Properties of Risk Sets

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

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

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

(R2) Let $t \geq 0$ and $b_t(Q, P)$ 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 the risk set $R(P, Q)$ of a testing problem is determined by its family of supporting lines from below, namely

$$\beta = b_t(Q, P) - t\alpha \,, \qquad t \geq 0 \,.$$

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

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

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

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

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.

$$\frac{q(x_1)}{p(x_1)} \geq \frac{q(x_2)}{p(x_2)} \geq \cdots \geq \frac{q(x_m)}{p(x_m)} \,,$$

take the tests

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

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

Remark 2: For the special case $P = (\frac{1}{m}, \ldots, \frac{1}{m})$ and $Q = (q_1, \ldots, q_m)$, such that $q_1 > q_2 > \cdots > q_m$, the lower boundary of the set

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

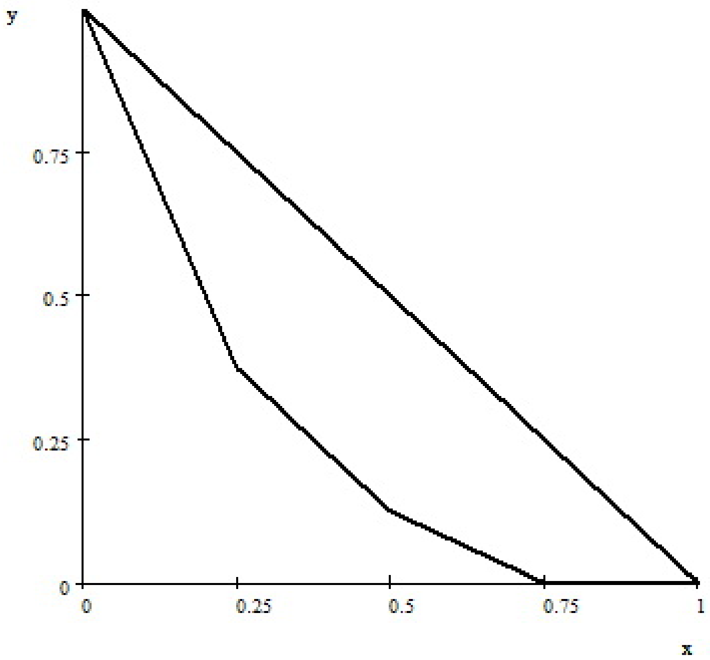

Figure 1: Risk set of above testing problem.

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."

3 f-Divergences

3.1 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

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

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.

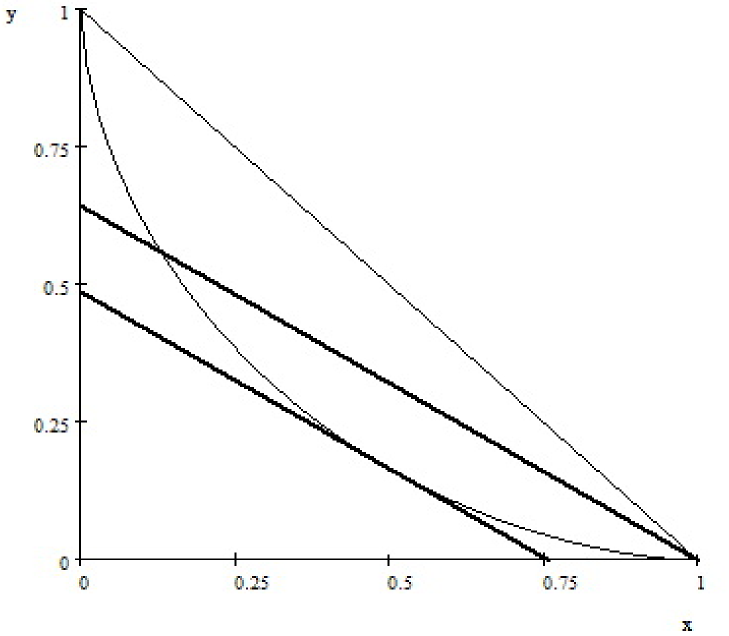

Figure 2: Differences $\min(1,t)$–$b_t(Q,P)$.

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.

3.2 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 breadth of $R(P,Q)$ in direction of the vector $(\frac{t}{1+t}, \frac{1}{1+t})$ of the prior distribution,

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

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

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

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

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

the difference

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

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 and 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:

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

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

3.3 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

$$f^*(u) = uf(1/u)\,, \qquad u \in (0,\infty)\,,$$

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

$$\begin{aligned} x \cdot f(0) &= x \cdot f(0/x) = 0 \cdot f^*(x/0) && \text{for } x \in (0,\infty) \\ y \cdot f^*(0) &= y \cdot f^*(0/y) = 0 \cdot f(y/0) && \text{for } y \in (0,\infty) \\ 0 \cdot f(0/0) &= 0 \cdot f^*(0/0) = 0\,. \end{aligned}$$

Definition 2 (Csiszár, 1963; Ali and 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 ($f(u) = |u-1|/2 = f^*(u)$)

$$I_f(Q,P) = ||Q-P||/2 = \frac{1}{2}\sum_{x\in\Omega}|q(x)-p(x)|$$

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

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

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

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

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

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

Squared Puri-Vincze Distance ($f(u) = (u-1)^2/2(1+u) = f^*(u)$)

$$I_f(Q,P) = \frac{1}{2}\sum_{x\in\Omega}\frac{(q(x)-p(x))^2}{p(x)+q(x)}$$

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 3: Note that

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

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 $\{x\in\Omega : q(x)p(x)>0\}=\Omega$, i.e., all probabilities are positive.

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

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

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

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

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

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

with equality holding if $Q(A_i) \times p(x) = P(A_i) \times q(x) \, \forall x \in A_i, \ i \geq 1$.
(c) Let $\alpha \in [0,1]$ and P_1, P_2 and Q_1, Q_2 probability distributions on Ω. Then

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

3.4 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

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

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, and 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 $h(u) = (1-u^\alpha)^{1/\alpha}/f(u)$, $u \in [0,1)$, 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 4: 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 and 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

$$c_f(||Q-P||/2) \leq I_f(Q,P) \leq c_f(1)\cdot ||Q-P||/2\,,$$

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

4 Construction of Least Favourable Distributions

Huber and 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 and 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

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

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

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

of the density q.

Simple Example (Continuation): In order to illustrate Theorem 7 let us continue our simple example from Section 2 by replacing the distribution Q by the total variation neighbourhood

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

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

$$\begin{aligned} Q &= (5/8, 1/4, 1/8, 0) \\ Q^* &= (4/8, 1/4, 1/8, 1/8) \end{aligned}$$

notice that the probability $1/8$ is shifted from the most probable element to the least probable.

Remark 6: For the special case $\Omega = \{1, \ldots, n\}$, $P = (1/n, \ldots, 1/n)$ and $Q = (q_1, \ldots, q_n)$ 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

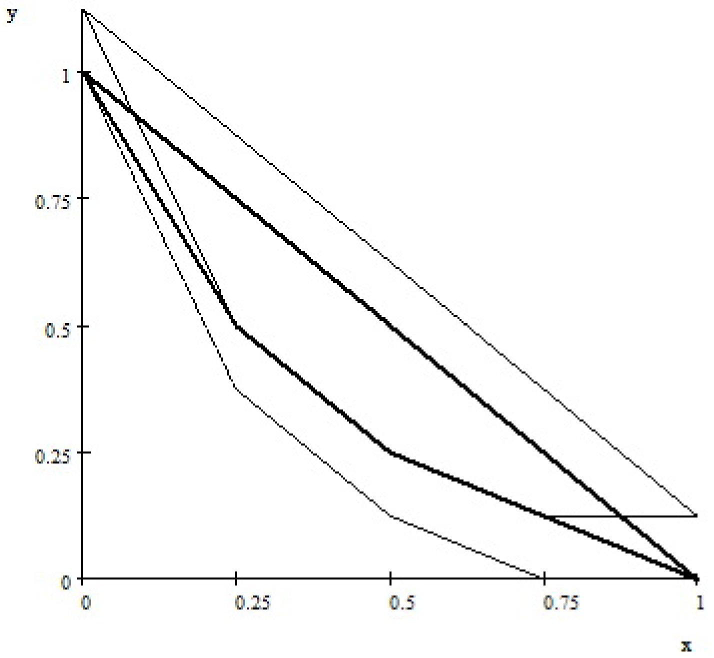

Figure 3: Modification of the risk set.

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."

5 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

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

(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 generalization yielding the family of f-divergences defined by the convex functions

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

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

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

to the class of entropies investigated by Arimoto (1971)

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

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

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

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

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

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

For further results, including those in connection with possible applications for minimum f-estimation, we refer to the paper mentioned above.

Vajda extended this family of divergences — in rechanging the parameter from p to $\alpha = 1/p$ — to

$$\varphi_\alpha(u) = \frac{\operatorname{sgn}(\alpha)}{1-\alpha}[(1+u^{1/\alpha})^\alpha - 2^{\alpha-1}(1+u)]\,, \qquad \alpha \in \mathbb{R}\backslash\{0,1\}$$

with the limiting cases $\varphi_0 = f_\infty$ and $\varphi_1 = f_1$ and Theorem 8 to

Theorem 9 (Vajda, 2009): This class of f-divergence provides the distances

$$[I_{\varphi_\alpha}(Q,P)]^{1/\max(\alpha,2)} \quad \text{for } \alpha \in \mathbb{R}\backslash\{0\} \text{ and } ||Q-P||/2 \text{ for } \alpha = 0\,.$$

This family includes for $\alpha = -1$ also the function

$$4 \cdot \varphi_{-1}(u) = \frac{(u-1)^2}{2(1+u)}$$

yielding the well-known divergence

$$\chi^2(Q,(P+Q)/2) = \frac{1}{2}\sum_{x\in\Omega} \frac{q(x)-p(x)^2}{p(x)+q(x)}$$

introduced by Sanghvi (1953). This divergence again, was extended to the family of metric divergences defined by the convex functions

$$\Phi_\alpha(u) = \frac{|u-1|^\alpha}{2(1+u)^{\alpha-1}} \qquad \text{for } \alpha \in [1,\infty)$$

by Puri and Vincze (1988).

Acknowledgement

The author would like to thank the referee for her/his valuable suggestions, especially the one to include a(n introductory) summary.

References

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

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

Csiszár, I. (1963). Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, *8*, 85–107.

Csiszár, I. (1974). Information measures: A critical survey. In J. Kozesnik (Ed.), *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes and of the 1974 European European Meeting of Statisticians* (Vol. A, p. 73-86). Academia Prague.

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

Feldman, D., and Österreicher, F. (1981). Divergenzen von Wahrscheinlichkeitsverteilungen – integralgeometrisch betrachtet. *Acta Mathematica Hungarica*, *37*, 329-337.

Feldman, D., and Österreicher, F. (1989). A note on f-divergences. *Studia Scientiarum Mathematicarum Hungarica*, *24*, 191-200.

Huber, P. J., and Strassen, V. (1973). Minimax tests and Neyman-Pearson lemma for capacities. *The Annals of Statistics*, *1*, 251-263.

Kafka, P., Österreicher, F., and Vincze, I. (1991). On powers of f-divergences defining a distance. *Studia Scientiarum Mathematicarum Hungarica*, *26*, 415-422.

Linhart, J., and Österreicher, F. (1985). Uniformity and distance – a vivid example from statistics. *International Journal of Mathematical Education in Science and Technology*, *16*, 645-649.

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

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

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. *Annals of the Institute of Statistical Mathematics*, *55*, 639-653.

Puri, M. L., and Vincze, I. (1988). Information and mathematical statistics. In P. Mandl and M. Huskova (Eds.), *Proceedings of the 4th Conference on Asymptotic Statistics.* Prague: Charles University.

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

Sanghvi, L. D. (1953). Comparison of genetics and morphological methods for a study of biological differences. *American Journal of Physical Anthropology*, *11*, 385-404.

Vajda, I. (1972). On f-divergence and and singularity of probability measures. *Periodica Mathematica Hungarica*, *2*, 223-234.

Vajda, I. (1989). *Theory of Statistical Inference and Information.* Dordrecht-Boston-London: Kluwer Academic Publishers.

Vajda, I. (2009). On metric divergences of probability distributions. *Kybernetika*, *45*, 885-900.

Author's address:

Ferdinand Österreicher
Department of Mathematics
University of Salzburg
Hellbrunner Straße 34
5020 Salzburg
Austria
E-mail: `Ferdinand.Oesterreicher@sbg.ac.at`