# ON THE MINIMUM DISCRIMINATION INFORMATION THEOREM

L. Rüschendorf

Abstract: A basic result in the information theoretic approach to statistics developped essentially by Kullback [7] is the minimum discrimination information theorem, which allows in many cases to determine the projection of a probability measure on a set of probability measures with given linear constraints w.r.t. the Kullback-Leibler distance. A partial converse of this result is due to Csiszár [3]. In the present paper we prove analogical results for a large class of divergence-type distances, thus opening the way to extend the information theoretic approach to a larger class of distances. It turns out, that the Kullback-Leibler distance has some properties, which allow specific simple calculations.

## 1. Introduction

Let $(X, \mathcal{A})$ be a measurable space and let for $P, R \in M^1(X)$ - the set of probability measures on $(X, \mathcal{A})$ - $I(P|R)$ denote the Kullback-Leibler distance, i.e.

AMS subject classification: 60-00-E05, 62-B10.
Key words and phrases: Minimum discrimination information, $\varphi$-divergences, measures with linear constraints.

$$(1) \qquad I(P|R) = \begin{cases} \int \ln \frac{dP}{dR}\, dP, & \text{if } P \ll R \\ \infty, & \text{else} \end{cases}.$$

For $Q \in M^1(X)$ and a vector subspace $F < L^1(Q)$ with $1 \in F$ let

$$(2) \qquad M = \{P \in M^1(X);\ F \subset L^1(P) \text{ and } \int f dP = \int f dQ,\ f \in F\}.$$

The minimum discrimination information theorem (MDIT), which is due to Kullback [7], Kullback and Khairat [8] states: If $R \in M^1(X)$, $P^* \in M$ and $f \in F$ satisfy $\frac{dP^*}{dR} = \exp f$, then

$$(3) \qquad I(P^*|R) = \inf\{I(P|R);\ P \in M\},$$

i.e. $P^*$ is the I-projection of R on M.

A partial converse of (3) was found by Csiszár [3]: If $P^*$ is the I-projection of R on M with $I(P^*|R) < \infty$, then there exists a $g \in L^1(F, P^*)$ - the closure of F in $L^1(P^*)$ - such that

$$(4) \qquad \frac{dP^*}{dR} = \exp g \ [P^*].$$

(Since each $P \in M$ with $I(P|R) < \infty$ satisfies $P \ll P^*$ (cf. relation (14)) and, furthermore, we can in the definition of M assume w.l.g. that F is a vector space, (4) is equivalent to Csiszár's formulation).

A consequence of (3), (4) ist, that (4) gives a necessary and sufficient condition for an I-projection if F is closed in $L^1(P)$ for all $P \in M$ with $I(P|R) < \infty$, assuming the existence of $P \in M$ with $R \ll P$ and $I(P|R) < \infty$.

Consider now the functional

$$(5) \qquad I_\varphi(P|R) = \begin{cases} \int \varphi\left(\frac{dP}{dR}\right) dR, \text{ if the integral exists} \\ \infty \text{ , else} \end{cases}.$$

where $\varphi$: $(0,\infty) \to R$ is a convex function and $\varphi(0) = \lim_{x \downarrow 0} \varphi(x)$, $\varphi(\infty) = \lim_{x \to \infty} \varphi(x)$. (We define $\frac{dP}{dR} = \frac{dP}{d\lambda}/\frac{dR}{d\lambda}$, where $P \ll \lambda$, $R \ll \lambda$, and use the convention $0 \cdot \infty = 0$. Note, that $R\{\frac{dP}{dR} = \infty\} = 0$.). We will show in the following, that results similar to (3) and (4) can be obtained also for many functionals $I_\varphi$.

For convex functions $\varphi$ as above Csiszár [2] introduced the $\varphi$-divergence measures $J_\varphi(P|R) = \int \varphi\left(\frac{dP}{d\lambda}/\frac{dR}{d\lambda}\right) \frac{dR}{d\lambda} d\lambda$, using the convention $0\varphi\left(\frac{0}{0}\right) = 0$, $0\varphi\left(\frac{a}{0}\right) = a \lim_{u\to\infty} \frac{\varphi(u)}{u}$ for $0 < a < \infty$. It has been shown by Csiszár [2] and in several further papers, that $J_\varphi$ shares many of the useful statistical properties of the Kullback-Leibler distance. The reason for the introduction of the modification $I_\varphi$ of $J_\varphi$ is of technical nature. It turns out, that $I_\varphi$ is more adapted to the considered minimization problem but, simultaneously, looses in the general case some useful statistical properties of the divergence measure $J_\varphi$. But, as will become clear from the following discussion, the minimization results for $I_\varphi$ imply corresponding results for $J_\varphi$ in many cases.

$I_\varphi$ and $J_\varphi$ are related by

$$(6) \qquad J_\varphi(P|R) = I_\varphi(P|R) + \lim_{u\to\infty} \frac{\varphi(u)}{u} P\{\frac{dR}{d\lambda} = 0\}.$$

Therefore, for $P \ll R$ we have $I_\varphi(P|R) = J_\varphi(P|R)$. If $\lim_{u\to\infty} \frac{\varphi(u)}{u} = \infty$ then the $J_\varphi$-projection $P^*$ of $R$ on $M$ satisfies $P^* \ll R$ and, therefore, is identical with the $I_\varphi$-projection of $R$ on

$$(7) \qquad M(R) = \{P \in M;\ P \ll R\} ;$$

examples for this situation are $\varphi(x) = x \log x$ or $\varphi(x) =$

$|x-1|^{\alpha}$, $\alpha > 1$, leading to the Kullback-Leibler distance, resp. $\chi^{\alpha}$ - distance. Also, as is immediate from (6), the case $\lim_{u\to\infty} \frac{\varphi(u)}{u} = 0$ implies equality of $I_\varphi$- and $J_\varphi$-projections. An example is $\varphi(x) = x^{-\alpha}-1, \alpha > 0$. Note that in this case $\varphi(0) = \infty$, which implies, that a projection $P^*$ on M satisfies $R \ll P^*$. So a difference can only occur for $0 < \lim_{u\to\infty} \frac{\varphi(u)}{u} < \infty$. If e.g. $\varphi(u) = |u-1|$, then $J_\varphi(P|R)$ is the total variation distance of P and R, while $I_\varphi(P|R)$ is the total variation distance between $P_R$ - the continuous part of P w.r.t R - and R.

Since $I_\varphi$ is a functional of $\frac{dP}{dR}$ it will be more adapted for the problem of determining $I_\varphi$-projections to consider linear constraints of the following form: For $F < L^1(Q_R)$ define

$$(8) \qquad \tilde{M} = \{P \in M^1(X);\ F \subset L^1(P_R) \text{ and } \int f dP_R = \int f dQ_R,\ f \in F\}.$$

If $Q \ll R$, then $\tilde{M} \cap M \supset M(R)$; if $1 \in F$ and $Q \not\ll R$, then $\tilde{M} \cap M = \phi$. Let $F_1 = \langle 1, F\rangle$, the vectorspace generated by 1 and F and let $M_\varphi$ resp. $\tilde{M}_\varphi$ denote the elements P of M resp. $\tilde{M}$ with $I_\varphi(P|R) < \infty$. For the first reading one may like to concentrate on the most important case of projection on M(R), where $Q \ll R$ and $1 \in F$. In this case the notations will simplify considerably.

## 2. Optimization of linear functions and closedness properties

In this section we derive some results for the optimization of linear functions, which are used in the subsequent sections for the determination of $I_\varphi$-projections.

An interesting characterization of the extreme points of the convex set M is due to Douglas [4]: If $P^* \in M$, then $P^* \in ex(M)$ - the set of extreme points - iff

$$(9) \qquad F_1 \text{ is dense in } L^1(P^*).$$

Since $M(R)$ and $M(R,s) = \{P \in M;\ P$ has a bounded density w.r.t. $R\}$ are extremal subsets of $M$, we have $ex\,(M(R)) \subset ex(M)$, $ex\,(M(R,s)) \subset ex(M)$, so that the same characterization of extreme points is true in $M(R)$ and $M(R,s)$. The idea that linear functions on $M$ take their infima in extreme points leads to.

Proposition 1. If $P^* \in M$ and $g \in L^1(P^*)$ satisfies $\int g dP^* = \inf\{\int g dP;\ P \in M(P^*,s)\}$, then

(10) $g \in L^1(F_1,P^*)$ or, equivalently, $\int g dP = \int g dP^*$ for all $P \in M(P^*,s)$.

Proof. The proof follows the lines of Csiszár [3], Theorem 3.1, and Douglas [4]. Let $F_1^{\perp} = \{h \in L^{\infty}(P^*); \int h f dP^* = 0$ for all $f \in F_1\}$, then

(11) $\int g h dP^* = 0$ for all $h \in F_1^{\perp}$,

since, otherwise, there would exist a $h \in F_1^{\perp}$, such that $\int g h dP^* < 0$. Then $\tilde{P} = (1 + \frac{h}{\|h\|_{\infty}})\ P^*$ - the measure which has density $1 + \frac{h}{\|h\|_{\infty}}$ w.r.t. $P^*$, $\|h\|_{\infty}$ denoting the norm in $L^{\infty}(P^*)$ - is an element of $M$. Now $\int g d\tilde{P} < \int g dP^*$ leads to a contradiction to our assuption. By Hahn - Banach's theorem (11) implies, that $g \in L^1(F_1,P^*)$. The equivalence of this condition to the second condition of (10) is again immediate from the Hahn-Banach theorem.

□

Remark 1. a) If $(X,\mathcal{A})$ is a topological space with Borel $\sigma$-algebra $\mathcal{A}$, if $M$ is a closed subset of the set of tight measures supplied with weak topology (i.e. the topology of the convergence of integrals of bounded continuous functions) and if $g$ is a bounded universally measureable function, then by an extension of the Choquet-Bishop-de Leeuw theorem due to v. Weizsäcker [17] there exists an extremal solution

$\tilde{P}$ of the inf problem and, therefore, by Douglas theorem $g \in L^1(F_1,\tilde{P})$. Proposition 1 states this property for any solution.

b) If $P^* \in \tilde{M}$, $g \in L^1(P_R)$ satisfies $\int g dP_R^* = \inf\{\int g dP_R;\ P \in \tilde{M},\ P$ has a bounded density w.r.t. $P^*\}$, then similarly to Proposition 1, $\tilde{g} \in L^1(\tilde{F}_1,P^*)$, where $\tilde{g} = g \cdot 1_{\{\frac{dP^*}{dR} < \infty\}}$, $\tilde{F} = \{f \cdot 1_{\{\frac{dP^*}{dR}<\infty\}};\ f \in F\}$. For $P^* \ll R$ this is equivalent to $g \in L^1(F_1,P^*)$.

Proposition 1 can be sharpened under additional assumptions on F.

Proposition 2. If $F_1$ is a vectorlattice in $L^\alpha(P)$, $1 < \alpha \leq \infty$, if $P^* \in M$, $g \in L^\alpha(P^*)$ and $\int g dP^* = \inf\{\int g dP;\ P \in M(P^*,s)\}$ then

(12) $g \in L^\alpha(F_1,P^*)$ - the closure of $F_1$ in $L^\alpha(P^*)$ - for $\alpha < \infty$ resp. w.r.t. weak $*$-topology for $\alpha = \infty$.

Proof. We only consider the case $\alpha = \infty$; the case $1 < \alpha < \infty$ is similarly proved. By Proposition 1, $g \in L^1(F_1,P^*)$ and, therefore, there exists a sequence $(f_n) \subset F_1$ with $\lim_{n\to\infty} f_n = g\,[P^*]$. If $|g| \leq K\ [P^*]$, then $(f_n \wedge K) \vee (-K) \in F_1$ ($\wedge$ denoting the infimum, $\vee$ the supremum), and, therefore, w.l.g. $|f_n| \leq K$. So for all $h \in F_1^\perp = \{h' \in L^1(P^*);\ \int f h' dP^* = 0$ for all $f \in F_1\}$ we have by the theorem of majorized convergence $0 = \lim \int f_n h dP^* = \int g h dP^*$, implying by Hahn-Banach's theorem that $g \in L(F_1,P^*)$. □

We now consider the following special case. Let $A_i \subset A$ be sub $\sigma$-algebras, $1 \leq i \leq k$, and let $F = \{\sum_{i=1}^k f_i;\ f_i \in B(X,A_i)$, $1 \leq i \leq k\}$, where $B(X,A_i)$ are the bounded $A_i$-measurable functions. Then M is the set of probability measures with marginals $Q_i = Q/A_i$, $1 \leq i \leq k$. (Equivalently, we can take $F = \{\sum_{i=1}^k f_i;\ f_i \in L^\alpha(Q,A_i)\ 1 \leq i \leq k\}$, $1 \leq \alpha \leq \infty$. We shall use the following two assumptions:

A1) $(X, \mathcal{A}_i, Q_i)$, $1 \leq i \leq k$, are compactly approximable, i.e. there exist compact set-systems $\mathcal{E}_i \subset \mathcal{A}_i$ with $Q_i(A_i) = \sup \{Q_i(E_i),\ E_i \subset A_i,\ E_i \in \mathcal{E}_i\}$, $1 \leq i \leq k$.

A2) $(X, \mathcal{A})$ is a topological space with Borel $\sigma$-algebra $\mathcal{A}$ and $R = R(\bigcup_{i=1}^{k} \mathcal{A}_i)$ - the algebra generated by $\bigcup_{i=1}^{k} \mathcal{A}_i$ - contains a countable basis of the topology.

Let H denote the uniform closure of the set of finite linear combinations of characteristic functions of sets in R.

Theorem 3. If $g \in H$ and A1) holds or if g is bounded upper - or lower - semicontinuous and A1), A2) hold, then

$$
\begin{aligned}
m &= \inf \{\int g\,dP;\ P \in M\} \\
&= \sup \Big\{\sum_{i=1}^{k} \int f_i\,dQ_i;\ f_i \in B(X, \mathcal{A}_i),\ 1 \leq i \leq k,\ \sum_{i=1}^{k} f_i \leq g\Big\} \\
&= M
\end{aligned} \tag{13}
$$

Proof. The proof follows the lines of the proof of Theorem 5 in [13], observing the following facts:

1) Unter A1) each finitely additive set function P on $(X, \mathcal{A})$ with $P/\mathcal{A}_i = Q_i$, $1 \leq i \leq k$, is $\sigma$-additive on $R(\bigcup_{i=1}^{k} \mathcal{A}_i)$.

2) Under the conditions A1), A2) each regular bounded additive set function P on $(X, \mathcal{A})$ with $P/\mathcal{A}_i = Q_i$, $1 \leq i \leq k$, is $\sigma$-additive (the regular bounded additive set functions are the dual of the space generated by bounded semicontinuous functions).

□

For the existence of solutions of the dual problem in (13) we need some further conditions allowing to bound a sequence $f_i^{(n)}$ with $\sum_{i=1}^{k} f_i^{(n)} \leq g$ and $\sum_{i=1}^{k} \int f_i^{(n)}\,dQ_i \to m$. Let $P^*$ be an optimal solution of the inf-problem in Theorem 3, then we

clearly have $\int |g - \sum_{i=1}^{k} f_i^{(n)}| dP^* \to 0$ and, therefore, (for a subsequence) $\sum_{i=1}^{k} f_i^{(n)} \to g[P^*]$. The following lemma gives a sufficient condition to imply that $g \in F$.

Lemma 4. Let $1 \leq \alpha \leq \infty$, $f_i^{(n)} \in L^\alpha(P,A_i)$, $1 \leq i \leq k$, $n \in \mathbb{N}$, and $g \in L^\alpha(P,A)$ such that $\sum_{i=1}^{k} f_i^{(n)} \to g$ [P].
If $\int |f_i^{(n)}|^\alpha dP \leq K$ for $1 < \alpha < \infty$, $|f_i^{(n)}| \leq K$ for $\alpha = \infty$, and $\{f_i^{(n)}\}$ uniformly P-integrable and bounded for $\alpha = 1$, then there exist $f_i \in L^\alpha(P,A_i)$, $1 \leq i \leq k$, with $g = \sum_{i=1}^{k} f_i$ [P].

Proof. 1) $\alpha = 1$: Since $\{f_i^{(n)}; n \in \mathbb{N}\}$, $1 \leq i \leq k$, are uniformly integrable and bounded, $\{f_i^{(n)}, n \in \mathbb{N}\}$ is weakly sequentially compact in $L^1(P,A)$. Therefore, we can find subsequences and $f_i \in L^1(P,A)$ such that $f_i^{(n)} \to f_i$ w.r.t. weak topology, $1 \leq i \leq k$. Since the weak and the strong closure of convex sets are identical there exists a sequence $g_n \in \operatorname{con}\{f_i^{(n)}; n \in \mathbb{N}\}$ converging strongly to $f_i$, so we may assume that $f_i \in L^1(P,A_i)$, $1 \leq i \leq k$. Furthermore, $\int (\sum_{i=1}^{k} f_i^{(n)} - \sum_{i=1}^{k} f_i) h dP \to 0$ for all $h \in L^\infty(P,A)$. Since by assumption $\sum_{i=1}^{k} f_i^{(n)}$ converges to g in $L^1(P,A)$, we have $g = \sum_{i=1}^{k} f_i$ [P].

2) $1 < \alpha < \infty$: The proof is similar to the proof of 1) observing, that bounded subsets of $L^\alpha(P,A)$ are weakly sequentially compact.

3) $\alpha = \infty$: $\{f_i^{(n)}; n \in \mathbb{N}\}$ are weakly compact subsets of $L^1(P,A)$ and, therefore, by 1) $g = \sum_{i=1}^{n} f_i$ [P] with $f_i \in L^1(P,A_i)$. Since $g_n$ of the proof of 1) are bounded by K, clearly also $|f_i| \leq K$. □

Remark 2. a) $F = \{\sum_{i=1}^{k} f_i; f_i \in L^1(Q_i,A_i), 1 \leq i \leq k\}$ is generally not closed in $L^1(Q,A)$ as is shown by an example due to Lindenstrauß [10] in the case $X = [0,1]^2$, $A = B^2 \cap [0,1]^2$ and $A_i$ are the $\sigma$-algebras generated by the projections $\pi_i$, $i = 1,2$. So $f_i(x_1,x_2) = f_i(x_i)$ in this case. Therefore, the proof of Corollaries 3.1, 3.2 of Csiszár [3] ( a generaliza-

tion of a result of Sinkhorn [15] and Hobby, Pyke [6]) is not correct in the general case. (A referee has pointed out to the author, that a corrected proof of these results will be published).

b) For $(X,A) = \bigotimes_{i=1}^{n} (X_i,B_i)$, a product of polish spaces, and for $J_1,\dots,J_k \subset \{1,\dots,n\}$ let $A_i$ be the $\sigma$-algebras induced by $\bigotimes_{j\in J_i} B_j$ in $X$, $1 \le i \le k$, $Q_i = Q_{J_i}$ the marginals of $Q$ on $A_i$ and $F = \{\sum_{i=1}^{k} f_i;\ f_i \in B(X,A_i),\ 1 \le i \le k\}$. Then assumptions A1), A2) of Theorem 3 are fulfilled (with the compact systems $B_i = \{K_i \times \prod_{j\in J_i^c} X_j;\ K_i \text{ compact in } \prod_{j\in J_i} X_j\}$).

If $J_1,\dots,J_k$ are pairwise disjoint, then it has been proved in [5] that in (13) we can restrict to $f_i \in B_K(X,A_i)$, $1 \le i \le k$, - the elements of $B(X,A_i)$, which are bounded by a suitable constant $K$ - and, therefore, by Lemma 4 there exist solutions $f_1^*,\dots,f_k^*$ of the dual problem of Theorem 3 and are characterized by $g = \sum_{i=1}^{k} f_i^*$ [$P^*$]. If $J_1,\dots,J_k$ are not pairwise disjoint but if there exists a subset $J \subset J_{i_0}$ such that $J_i \setminus J$ are pairwise disjoint, then the same arguments yield the possibility to restrict to $B_K(X,A_i)$ (arguing for the $x_J$ sections of $f_i^{(n)}$ separately). If e.g. $n = 5$, $k = 4$, $J_1 = \{1,2,3\}$, $J_2 = \{2,3,4\}$, $J_3 = \{2,5\}$ and $J_4 = \{2,6\}$, we can use $J = \{2,3\}$ and, therefore, have the existence of solutions also in this case.

## 3. Minimum discrimination w.r.t. $\varphi$-type divergences

We now consider $\varphi$-type divergence measures as defined in (5). We consider at first the case of differentiable $\varphi$, then the case of $\chi^\alpha$-divergences, $1 \le \alpha < \infty$, and finally remark on certain similar distances.

### 3.1 The differentiable case

Let $\varphi\colon (0,\infty) \to R^1$ be a continuous, strictly convex and differentiable function and call an element $P^*$ of a convex

subset $M'$ of $M$ a $I_\varphi$-projection of $R$ on $M'$, if $I_\varphi(P^*|R) = \inf\{I_\varphi(P|R);\ P \in M'\}$ (note that $I_\varphi$-projections $P^*$ are uniquely determined, if $I_\varphi(P^*|R) < \infty$ and $M' \subset M(R)$. For certain existence results on the corresponding $J_\varphi$-projections cf. Liese [9]. Remember the definition of $\tilde{F}_1$ in Remark 1. b).

Theorem 5. Let $M' \subset M$ be convex, let $P^* \in M'$ satisfy $I_\varphi(P^*|R) < \infty$ and $\varphi'(\frac{dP^*}{dR}) \in L^1(P^*_R)$.

a) $P^*$ is the $I_\varphi$-projection of $R$ on $M'$ iff

$$\int \varphi'(\frac{dP^*}{dR})\ (\frac{dP^*}{dR} - \frac{dP}{dR})\ dR \leq 0 \text{ for all } P \in M' \cap M_\varphi.$$

b) If $P^*$ is the $I_\varphi$-projection on $M$, then

$$\varphi'(\frac{dP^*}{dR})\ 1_{\{\frac{dP^*}{dR} < \infty\}} \in L^1(F_1, P^*).$$

c) If $P^*$ is the $I_\varphi$-projection on $\tilde{M}$, then

$$\varphi'(\frac{dP^*}{dR})\ 1_{\{\frac{dP^*}{dR} < \infty\}} \in L^1(\tilde{F}_1, P^*).$$

d) If $P^* \in \tilde{M}$ and $\varphi'(\frac{dP^*}{dR}) \in F$, then $P^*$ is the $I_\varphi$-projection on $\tilde{M}$.

Proof. a) For $P \in M_\varphi \cap M'$ and $\alpha \in [0,1]$ define $h_\alpha = \frac{1}{\alpha - 1}\ (\varphi(\alpha\frac{dP}{dR} + (1-\alpha)\frac{dP^*}{dR}) - \varphi(\frac{dP^*}{dR}))$. For $\alpha \uparrow 1$, $h_\alpha$ converges monotonically nondecreasing to $\varphi'(\frac{dP^*}{dR})\ (\frac{dP}{dR} - \frac{dP^*}{dR})$ and, therefore, by the monotone convergence theorem (using $h_\alpha \geq \varphi(\frac{dP}{dR}) - \varphi(\frac{dP^*}{dR})$), $\frac{1}{\alpha-1} \int (\varphi(\alpha\frac{dP}{dR} + (1-\alpha)\frac{dP^*}{dR}) - \varphi(\frac{dP^*}{dR}))\ dR$ converges monotonically nondecreasing to $\int \varphi'(\frac{dP^*}{dR})\ (\frac{dP}{dR} - \frac{dP^*}{dR})\ dR$.

If $P^*$ is the $I_\varphi$-projection on $M'$, then the left hand side is $\leq 0$ for each $\alpha$, which implies, that also the limit on the right hand side is $\leq 0$. If, conversely, the right hand side is

$\leq 0$, then by the nondecreasing property of $h_\alpha$ we have that
$\int h_0 dR = -(I_\varphi(P|R) - I_\varphi(P^*|R)) \leq \int (\lim_{\alpha \uparrow 1} h_\alpha)\, dR \leq 0$;
i.e. $I_\varphi(P|R) \geq I_\varphi(P^*|R)$.

b) Since $\int \varphi'(\frac{dP^*}{dR}) \frac{dP^*}{dR}\, dR = \int \varphi'(\frac{dP^*}{dR})\, 1_{\{\frac{dP^*}{dR} < \infty\}}\, dP^*$ and for
$P \ll P^*$ : $\{\frac{dP}{dR} < \infty\} = \{\frac{dP^*}{dR} < \infty\} \cup \{\frac{dP}{dR} = 0\}$, we have

$$\int \varphi'(\frac{dP^*}{dR}) \frac{dP}{dR}\, dR = \int \varphi'(\frac{dP^*}{dR})\, 1_{\{\frac{dP}{dR} < \infty\}}\, dP = \int \varphi'(\frac{dP^*}{dR})\, 1_{\{\frac{dP^*}{dR} < \infty\}}\, dP.$$

Therefore, b) follows from a) and from Proposition 1).

c) follows similarly from a modified version of a) and from Remark 1, b).

d) If $P^* \in M$ and $\varphi'(\frac{dP^*}{dR}) \in F$, then

$$\int \varphi'(\frac{dP^*}{dR}) \frac{dP^*}{dR}\, dR = \int \varphi'(\frac{dP^*}{dR})\, dP^*_R = \int \varphi'(\frac{dP^*}{dR})\, dP_R = \int \varphi'(\frac{dP^*}{dR}) \frac{dP}{dR}\, dR.$$

Therefore, d) follows from a) (in the modified version for subsets of $\tilde{M}$.

□

Remark 3. a) If $\varphi(x) = x \log x$ and $M' = M(R)$, then $I_\varphi(P|R)$ is the Kullback-Leibler distance and condition a) says: $P^* \in M_\varphi$ is the I-projection of R on M iff

$$\int \log \frac{dP^*}{dR}\, dP^* = \inf \{\int \log \frac{dP^*}{dR}\, dP;\ P \in M_\varphi\} \text{ or,}$$
(14) equivalently
$$I(P|R) \geq I(P|P^*) + I(P^*|R), \text{ for all } P \in M_\varphi.$$

This geometric property of the Kullback-Leibler distance was proved by Cziszár [3], Th. 2.2. Note that the MDIT (3) and its converse (4), therefore, are consequences of Theorem 5, d) resp. Theorem 5, a) and Proposition 1 (with $g = \log \frac{dP^*}{dR}$).

b) A special case of part d) was proved by Perez [11], Lemma 2.1.

c) Sufficient conditions for the $I_\varphi$-projection on M (and not only on M(R)) are not easy to give. If $\varphi'(\frac{dP^*}{dR}) 1_{\{\frac{dP^*}{dR}<\infty, \frac{dP}{dR}<\infty\}} \in F$ for all $P \in M_\varphi$, then $P^*$ is a $I_\varphi$-projection on M.

But this condition seems to be difficult to verify.

3.2 $\chi^\alpha$-divergences

We next consider an important class of not necessarily differentiable functions, namely $\varphi(u) = |1-u|^\alpha$, $1 \leq \alpha < \infty$, and denote $I_\alpha(P|R) = I_\varphi(P|R) = \int |\frac{dP}{dR} - 1|^\alpha dR$. The $I_\alpha$-distances have been introduced in literature under the name $\chi^\alpha$-divergences by Vajda [16], who investigated several interesting properties of $I_\alpha$.

Consider at first the case $\alpha = 1$; then $I_1$ is the variation distance between $P_R$ and R. Define

$$\operatorname{sgn}(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ -1, & \text{if } x < 0 \end{cases}.$$

Theorem 6. Let $M' \subset \tilde{M}$ and $P^* \in M'$ satisfy $P^*\{\frac{dP^*}{dR} = 1\} = 0$, then

a) $P^*$ is a $I_1$-projection of R on M' iff

$$\int \operatorname{sgn}\,(1 - \frac{dP^*}{dR})\ (\frac{dP^*}{dR} - \frac{dP}{dR})\ dR \geq 0 \text{ for all } P \in M'.$$

b) If $P^* \in M(R)$, then $P^*$ is a $I_1$-projection on M(R) iff

$$P^*\{\frac{dP^*}{dR} \leq 1\} \geq P\ \{\frac{dP^*}{dR} \leq 1\} \text{ for all } P \in M(R).$$

c) A necessary (sufficient) condition for a $I_1$-projection on

$$\tilde{M} \text{ is } \operatorname{sgn}\left(1 - \frac{dP^*}{dR}\right) 1_{\{\frac{dP^*}{dR} < \infty\}} \in L^1(\tilde{F}_1, P)\ (\operatorname{sgn}\left(1 - \frac{dP^*}{dR}\right) \in F)$$

d) A necessary (sufficient) condition for a $I_1$-projection on $M(R)$ is $1_{\{\frac{dP^*}{dR} \leq 1\}} \in L^1(F_1, P^*)$ $(\in F)$.

Proof. a) Let $E = \{\frac{dP}{dR}\ ;\ P \in M'\}$; then $E$ is a convex subset of $L^1(R)$. By Theorem 1.1 of Singer [14] (pg. 360, using the equivalent conditions 1.2, 1.7, 1.8) $P^* \in M'$ is a $I_1$-projection of $R$ on $M'$ iff there exists a $h \in L^\infty(R)$ with

1) $\quad \|h\|_\infty = 1,$

2) $\quad \int \left(1 - \frac{dP^*}{dR}\right) h dR = \int \left|1 - \frac{dP^*}{dR}\right| dR$ and

3) $\quad \int \left(\frac{dP^*}{dR} - \frac{dP}{dR}\right) h dR \geq 0$ for all $P \in M'$.

Conditions 1), 2) are equivalent to $|h| \leq 1$ $[R]$and

$$h(x) = \begin{cases} 1 & < 1 \\ \text{if } \frac{dP^*}{dR} & \quad [R] \text{ or,} \\ -1 & > 1 \end{cases}$$

equivalently, using our assumption on $P^*$ to $h(x) = \operatorname{sgn}\left(1 - \frac{dP^*}{dR}\right)$ $[R]$.

b) Using the relation $\operatorname{sgn}\left(1 - \frac{dP^*}{dR}\right) = 2 \cdot 1_{\{\frac{dP^*}{dR} \leq 1\}} - 1$ the proof of b) is similar to that of a).

c), d). The proof of c), d) is analogical to that of Theorem 5.

□

Remark 4. The proof of the sufficiency parts of Theorem 6 did not use the assumption $P^*\{\frac{dP^*}{dR} = 1\} = 0$. Furthermore, as in Theorem 5, $\operatorname{sgn}\left(1 - \frac{dP^*}{dR}\right) 1_{\{\frac{dP^*}{dR} < \infty\}} \in L^1(F_1, P^*)$ is a necessary condition for a $I_1$-projection on $M$.

A sufficient condition can also be given for the variation distance $J_1(P|R) = \int |\frac{dP}{d\lambda} - \frac{dR}{d\lambda}| \, d\lambda$, where $P \ll \lambda$, $R \ll \lambda$.

Theorem 7. Let $P^* \in M$ with $1_{\{\frac{dP^*}{dR} \geq 1\}} \in F$, then $P^*$ is a $J_1$-projection of $R$ on $M$.

Proof. For $P \in M$, $J_1(P|R) = 2(P(A) - R(A))$, where $A = \{\frac{dP}{dR} \geq 1\}$. If $1_{\{\frac{dP}{dR} \geq 1\}} \in F$, then

$$J_1(P^*|R) = 2(P^*\{\tfrac{dP}{dR} \geq 1\} - R\{\tfrac{dP^*}{dR} \geq 1\})$$

$$= 2(P\{\tfrac{dP^*}{dR} \geq 1\} - R\{\tfrac{dP^*}{dR} \geq 1\})$$

$$\leq 2 \sup_{A \in} (P(A) - R(A)) = J_1(P|R),$$

for all $P \in M$.

For the case $\alpha > 1$ let $M_\alpha = \{P \in M;\ I_\alpha(P|R) < \infty\}$, $\tilde{M}_\alpha = \{P \in \tilde{M};\ I_\alpha(P(R) < \infty\}$. ¤

Theorem 8. If $1 < \alpha < \infty$ and $P^* \in M'$ where $I_\alpha(P^*|R) < \infty$ and $M'$ is convex, then

a) $P^*$ is the $I_\alpha$-projection of $R$ on $M'$, iff

$$\int \operatorname{sgn}(1 - \tfrac{dP^*}{dR}) \, |\tfrac{dP^*}{dR} - 1|^{\alpha-1} (\tfrac{dP}{dR} - \tfrac{dP^*}{dR}) \, dR \geq 0, \text{ for all } P \in M' \cap M_\alpha.$$

b) A necessary (sufficient) condition for a $I_\alpha$-projection on $\tilde{M}_\alpha$ is

$$1_{\{\frac{dP^*}{dR} < \infty\}} \operatorname{sgn}(1 - \tfrac{dP^*}{dR}) \, |\tfrac{dP^*}{dR} - 1|^{\alpha-1} \in L^1(\tilde{F}_1, P^*)$$

$$(\operatorname{sgn}(1 - \tfrac{dP^*}{dR}) \, |\tfrac{dP^*}{dR} - 1|^{\alpha-1} \in F).$$

c) A necessary (sufficient) condition for a $I_\alpha$-projection on $M(R)$ is for $P^* \in M(R)$, $|\frac{dP^*}{dR} - 1|^{\alpha-1} \operatorname{sgn}(1 - \frac{dP^*}{dR}) \in L^1(F_1, P^*)$ $(\in F)$.

Proof. The proof of Theorem 8 follows from that of Theorem 5. Alternatively, we can also use Theorem 1.1 of Singer [14], pg. 360 implying, that $P^* \in M'$ is the $I_\alpha$-projection of R on M', iff there exists a $h \in L^\beta(R)$, $\frac{1}{\alpha} + \frac{1}{\beta} = 1$, such that

1) $\quad ||h||_\beta = 1,$

2) $\quad \int (1 - \frac{dP^*}{dR})\, h dR = ||1 - \frac{dP^*}{dR}||_\alpha$ and

3) $\quad \int (\frac{dP^*}{dR} - \frac{dP}{dR})\, h dR \geq 0$ for all $P \in M' \cap M_\alpha$ (we use $E = \{\frac{dP}{dR};$ $P \in M', I_\alpha(P|R) < \infty\} \subset L^\alpha(R)$ in this case).

Observing, that the only element $h \in L^\beta(R)$ with 1) and 2) is given by

$$h = |1 - \frac{dP^*}{dR}|^{\alpha-1} \operatorname{sgn}(1 - \frac{dP^*}{dR}) / (||1 - \frac{dP^*}{dR}||_\alpha)^{\alpha-1}$$

we obtain a).

□

Remark 5. Similar results as in Theorem's 5, 6, 7, 8 can also be given for certain related distances.

Let e.g. $D(P|P') = \int \varphi(\frac{dP}{dR} - \frac{dP'}{dR})\, dR$, where $\varphi: R^1 \to R^1$ is convex, differentiable or $\varphi(u) = |u|^\alpha$, $1 \leq \alpha < \infty$, and the integral is assumed to exist. Necessary (sufficient) conditions for a D-projection of $P'_*$ on $\tilde{M}$ are in the case of a convex, differentiable $\varphi, \varphi'$ $(\frac{dP^*}{dR} - \frac{dP'}{dR})\, 1_{\{\frac{dP^*}{dR} < \infty\}} \in L^1(\tilde{F}_1, P^*)$ $(\varphi'(\frac{dP^*}{dR} - \frac{dP'}{dR}) \in F)$, while for $\varphi(u) = |u|^\alpha$ the corresponding conditions are $1_{\{\frac{dP^*}{dR} < \infty\}} \operatorname{sgn}(1 - \frac{dP^*}{dR}) |\frac{dP}{dR} - \frac{dP'}{dR}|^{\alpha-1} \in L^1(\tilde{F}_1, P^*)$

$$(\operatorname{sgn}\ (1 - \frac{dP^*}{dR})\ |\frac{dP^*}{dR} - \frac{dP'}{dR}|^{\alpha-1} \in F).$$

4. Remarks and examples

1) We did not use essentially the assumption, that R is a probability measure. So e.g. with R = $\lambda$-the Lebesgue-measure on $(R^1, B^1)$ - our results allow to determine distributions in M with maximum value of the entropy $H(P) = -\int \log \frac{dP}{d\lambda}\, dP$. The results of section 3 also hold true for projections on $\tilde{M}^s$ - the set of all signed measures P on $(X, A)$ with $\int f \frac{dP}{dR}\, dR = \int f \frac{dQ}{dR}\, dR$, $f \in F$. For general divergence-type distances it is easier to find projections in $\tilde{M}^s$ than in $\tilde{M}$. The Kullback-Leibler distance is an exception, since the necessary condition $\ln \frac{dP^*}{dR} = g\ [P^*]$, $g \in L^1(F_1, P^*)$ leads automatically to a nonnegative density $\exp(g)$. The same is true for differentiable $\varphi$ with $(\varphi')^{-1}(f) \geq 0$ for all $f \in F$.
Note that our results give lower bounds for the considered distances, even if there does not exist a nonnegative density satisfying the conditions of our Theorem's.

2) Let $F = \langle f_1, \ldots, f_n \rangle \subset L^1(Q)$, where $f_1, \ldots, f_n$ are not linearly dependent w.r.t. Q, then F is closed w.r.t. $L^1(P)$ for all $P \in M$. M is the set of all probability measures with given moments $\int f_j dP = a_j = \int f_j dQ$, $1 \leq j \leq n$.

a) If $X = \mathbb{R}^n$, $R = N(a, \Sigma)$ - the normal distribution with mean $a \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n \times n}$ - and $F = \langle 1, x_j, x_i x_j;\ 1 \leq i, j \leq n \rangle$ and if Q has mean b and covariance $\Psi$, then the I-projection of R on M is $P^* = N(b, \psi)$, since $\ln \frac{dP^*}{dR} \in F$. Similarly, $N(b, \psi)$ is the distribution in M (with mean b and covariance $\psi$) with maximum value of the entropy $H(P) = \int \ln \frac{dP}{d\lambda^n}\, dP$, $P \in M$, where $\lambda^n$ is the n-dimensional Lebesgue-measure.

b) Let R be the restriction of $\lambda^n$ on $[0, \infty)^n$ and $F = \langle 1, x_i;$

$1 \leq i \leq n > \subset L^1(Q, B^n [0,\infty)^n)$. Let Q have first moment vector $b > 0$, then $P^* = \bigotimes_{i=1}^n P^*_{b_i}$ - where $P_{b_i}$ is the exponential distribution with mean $b_i$-is the I-projection on M, i.e. $P^*$ has under all distributions on $\mathbb{R}^n_+$ with mean $b > 0$ the maximum value of the entropy.

c) If R is the restriction of $\lambda^1$ on $[-M',M']$, $M' \in \mathbb{R}^1$, and $F = < 1,x >$, then the I-projection on M is given by $\frac{dP^*}{dR}(x) = (M'-a)e^{(M'-a)x}(e^{(M'-a)M'} - e^{-(M'-a)M'})$, $a = \int x dQ$ ($|a| \leq M'$). If we use a $\chi^2$-distance $I_2(P|R) = \int (\frac{dP}{dR} - 1)^2 dR$, $P \in M_2$, we obtain from Theorem 8 the projection $\frac{dP^*_a}{dR}(x) = \frac{1}{2M'} + \frac{3a}{2M'^3} x$, on M(R) (assuming $|a| \leq \frac{M'}{6}$ in order to obtain a nonnegative density) and have the relation $I_2(P^*_a|R) = I_2(P^*_o|R) + \int_{-M'}^{M'} (\frac{3a}{2M'^3} x)^2 dx \geq I_2(P^*_o|R)$, which is intuitively obvious.

3) Let $A_o \subset A$ be a sub $\sigma$-algebra of $A$ and let $Q/A_o \ll R/A_o$ and $F = L^1(Q, A_o)$. Then M is the set of all extensions of $Q/A_o$ to the larger $\sigma$-algebra $A$. In this case F is a closed subset of $L^1(P, A)$ for each $P \in M$. Clearly, M is closed w.r.t variation distance. Therefore, a $I_1$-projection and the $J_\varphi$-projections exist, if $\lim \frac{\varphi(u)}{u} = \infty$ (cf. Liese [9]). Let

$$(15) \qquad P^*(A) = \int_A \frac{dQ/A_o}{dR/A_o} dR, \quad A \in A, \text{ then } P^* \in M(R) \text{ and}$$

by Theorems 5, 6, 7, 8, $P^*$ is the I-projection on M (since $\ln \frac{dP^*}{dR} \in F$), the $I_\alpha$-projection on M(R) for $1 \leq \alpha \leq \infty$ with $P^* \in M_\alpha$ (since $\operatorname{sgn}(1 - \frac{dP^*}{dR}) |\frac{dP^*}{dR} - 1|^{\alpha-1} \in L^1(Q, A_o) = F$). It is also the projection on M w.r.t variation distance $J_1$ (since $1_{\{\frac{dP^*}{dR} \geq 1\}} \in F$) and, finally, it is the projection on M(R) for all differentiable divergences $J_\varphi$ with $M \neq \emptyset$ (since $\varphi'(\frac{dP^*}{dR}) \in F$). So $P^*$ is 'universally' the best approximation w.r.t all considered distances. Some related results and the

statistical consequences were developped by Plachky, Rüschendorf [12].

4) Let $(X,A) = \bigotimes_{i=1}^{n} (X_i,A_i)$, let $F = \{\sum_{i=1}^{n} f_i(x_i); f_i \in L^1(Q_i,A_i)\}$, where $Q_i$ are marginals of Q. Then M is the set of probability measures with given marginals $Q_1,\ldots,Q_n$. If e.g. $R = \bigotimes_{i=1}^{n} R_i$, $R_i \in M^1(X_i,A_i)$, then the I-projection on M is $P^* = \bigotimes_{i=1}^{n} Q_i$. A simple direct argument shows for $P \in M$ with $I(P|R) < \infty$ (cf. also (14)).

$$(16) \qquad I(P|\bigotimes_{i=1}^{n} R_i) \geq I(P|\bigotimes_{i=1}^{n} Q_i) + I(\bigotimes_{i=1}^{n} Q_i|\bigotimes_{i=1}^{n} R_i),$$

i.e. w.r.t. Kullback-Leibler distance for a given measure P the closest product measure is the product of its marginals, a property which is usually not true w.r.t. divergence type measures. If we use a $\chi^2$-distance (with $\varphi(u) = (1-u)^2$) and if we assume, that there exist $\alpha_i \geq 0$, $\sum_{i=1}^{n} \alpha_i = 1$ with $Q_i = (\alpha_i f_i + \sum_{j \neq i} \alpha_j) R_i$, then the $I_2$-projection is given by the mixture $P^* = \sum_{i=1}^{n} \alpha_i R_1 \otimes \ldots \otimes f_i R_i \otimes \ldots \otimes R_n$. Therefore, w.r.t. $\chi^2$-distance the measure closest to $\bigotimes_{i=1}^{n} R_i$ with marginals $Q_i$ is not $\bigotimes_{i=1}^{n} Q_i$.

If $n = 2$, $X_1 = \{1,\ldots,m\}$, $X_2 = \{1,\ldots,\ell\}$ and $R = (\pi_{ij})$ and if $Q = (q_{ij})$ with $q_{i.} = p_i$, $q_{.j} = q_j$, then we obtain the following conditions for projections $p^*_{ij}$ on M.

a) For I-projections: $1_{\{\pi_{ij} > 0\}} \frac{p^*_{ij}}{\pi_{ij}} = a_i b_j$ is necessary and sufficient (only an iterative solution of this equation is generally known).

b) $I_2$-projection on M(R) : $1_{\{\pi_{ij} > 0\}} \frac{p^*_{ij}}{\pi_{ij}} = a_i + b_j$ is sufficient, where $a_i, b_j$ are solutions of $p_i = a_i \pi_{i.} + \sum_j b_j \pi_{ij}$, $q_j = \sum_i a_i \pi_{ij} + b_j \pi_{.j}$, $1 \leq i \leq m$, $1 \leq j \leq \ell$.

c) For $J_1$-projection a sufficient condition is that $\{(i,j) :$

$p^*_{ij} \geq \pi_{ij}\} \in \{I \times X_2,\ X_1 \times J;\ I \subset X_1,\ J \subset X_2\}$.

5) For $A_1,\ldots,A_n \in \mathcal{A}$, $n \leq \infty$, let M be the set of all distributions P with $P(A_i) = q_i = Q(A_i)$, $i \leq n$. If $\frac{dP^*}{dR} = \sum_{i=1}^{n} a_i 1_{A_i} + a_0$ and $P^* \in M$, then $P^*$ is the $\chi^2$-projection on M(R) and the projection w.r.t. total variation distance on M. The condition $P^*(A_i) = q_i$, $1 \leq i \leq n$, can be solved explicitely in several cases - e.g. for disjoint $A_i$, $1 \leq i \leq n$, with $\bigcup_{i=1}^{n} A_i = X$, $a_i = \frac{q_i}{R(A_i)}$ - while for several further cases one can give iterative solutions.

Acknowledgement: The author wishes to thank the referees for several useful remarks, which lead especially to a more careful consideration of the relation between I- and J-projections. Also the relevance of the papers of Liese, Perez and Vajda was pointed out by them.

References

[1] Chow, Y. S., Teicher, H.: Probability Theory. New York - Heidelberg - Berlin, Springer (1980).

[2] Csiszár, I.: Information type measures of difference of probability distributions and indirect observations. Studia Sci. Math. Hungar. 2, 299 - 318 (1967).

[3] Csiszár, I.: I-divergence geometry of probability distributions and minimization problems. Ann. Prob. 3, 146 - 158 (1975).

[4] Douglas, R. G.: On extremal measures and subspace density. Michigan Math. J. 11, 243 - 246 (1964).

[5] Gaffke, N., Rüschendorf, L.: On a class of extremal problems in statistics. Math. Operationsforsch., Ser. Optimization 12, 123 - 135 (1981).

[6] Hobby, C., Pyke, R.: Doubly stochastic operators obtained from positive operators. Pacific J. Math. 15, 153 - 157 (1965).

[7] Kullback, S.: Information Theory and Statistics. Wiley, New York (1959).

[8] Kullback, S., Khairat, M. A.: A note on minimum discrimination information. Ann. Math. Statist. 37, 279 - 280 (1966).

[9] Liese, F.: On the existence of f-projections. In: Topics in Information Theory (ed. I. Csiszár and P. Elias). Colloqia Math. Soc. J. Bolyai 16, North Holland, 431 - 446 (1977).

[10] Lindenstrauß, J.: A remark on extreme doubly stochastic measures. Amer. Math. Monthly 72, 379 - 382 (1965).

[11] Perez, A.: Information-theoretic risk estimates in statistical decisions. Kybernetika 3, 1 - 22 (1967).

[12] Plachky, D., Rüschendorf, L.: Conservation of the UMP-resp. maximin-property of statistical tests under extensions of probability measures. To appear in: Statistics and Decisions.

[13] Rüschendorf, L.: Sharpness of Fréchet-bounds. Z. Wahrscheinlichkeitstheorie verw. Gebiete 57, 293 - 302 (1981).

[14] Singer, I.: Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces. Springer (1970).

[15] Sinkhorn, R.: Diagonal equivalence to matrices with prescribed row and column sums. Amer. Math. Monthly 74, 402 - 405 (1967).

[16] Vajda, I.: $\chi^{\alpha}$-divergence and generalized Fisher's infor-

mation. Trans. 6th Prague Conf. on Information Theory. Statistical Decision Functions, Random Processes, Prague 873 - 886 (1973).

[17] v. Weizsäcker, H.: Der Satz von Choquet-Bishop-de Leeuw für konvexe nicht kompakte Mengen straffer Maße über beliebigen Grundräumen. Math. Z. 142, 161 - 165 (1975).

L. Rüschendorf
Institut für Math. Stochastik
Hebelstraße 27
7800 Freiburg