Ann. Inst. Statist. Math.
30 (1978), Part A, 105-113

f-DISSIMILARITY: A GENERALIZATION OF THE AFFINITY OF SEVERAL DISTRIBUTIONS

L. GYÖRFI AND T. NEMETZ

(Received Dec. 4, 1975; revised Aug. 17, 1977)

1. Introduction

In many areas of statistics (e.g. discriminant analysis, hypotheses testing, nonparametric statistic, pattern classification etc.) there is a need for some appropriate distances of probability distributions. There are two basic properties which an appropriate distance should satisfy:

(i) it should be non-increasing on transformed observation

(ii) be unchanged if the transformed observation (statistic) is sufficient one.

For the case of two distributions such distances are widely used and investigated. The case of more than two distributions, however, is mainly treated by using pairwise distances. Such a concept may result in overlooking characterizations and possible solutions for many problems. It seems reasonable to replace pairwise distances with measures of separation among (dissimilarity, discrimination of) more than two distributions. In this respect the pionir work was done by K. Matusita [8] who introduced the notion of affinity of several distributions. Although the affinity of distributions represents the likeness of distributions it serves as well as a measures for discriminating among distributions [10]. The negative of Matusita's affinity satisfies both (i) and (ii) (see [9]). Motivated by Matusita's works several authors proposed "affinity" and/or separation measures. Extracting the common feature of these measures, in [5] we had proposed a wide class of dissimilarity measures of several distributions:

DEFINITION 1. Let $f(s_1, \cdots, s_n)$ be a continuous, convex, homogeneous function defined on the set

$$S_n \stackrel{\Delta}{=} \{(s_1, \cdots, s_n);\ 0 \leqq s_i \leqq \infty,\ i=1, \cdots, n\}\ . \tag{1}$$

Let $P_1, \cdots, P_n$ be probability measures on the measurable space $(X, \mathcal{X})$ with Radon-Nikodym derivatives $p_1(x), \cdots, p_n(x)$ with respect to a dominating σ-finite measure μ. The f-dissimilarity of $P_1, \cdots, P_n$ is defined by

$$(2)\qquad D_f(P_1,\cdots,P_n)=\int_X f(p_1(x),\cdots,p_n(x))\mu(dx)\ .$$

We have shown in [6] that this class includes the following separation measures: the probability of correct decision of the Bayesian decision rule, the class of f-divergences ([1], [4]), the negative of Matusita's affinity ρ_n, the negative of Toussaint's affinity ρ_n^* ([12]) Ito's generalized Chernoff-bound C_n ([7]), Toussaint's dispersion R_r^n ([11]), the asymptotic probability of correct classification of the nearest neighbor decision rule ([3]).

In this paper we are concerned with the basic properties of the f-dissimilarities. In particular, we prove that the f-dissimilarity satisfies both (i) and (ii) for different kinds of indirect observations. A characterization similar to that of Theorem 3 [9] is given by finite partitions.

2. f-dissimilarity and indirect observation

For the sake of mathematical rigour we recall the following

DEFINITION 2. The function $f(s)$ $s\in S_n$ is called homogeneous if $f(ts)=tf(s)$ for all reals $t\geqq 0$ and $s\in S_n$. The function $f(s)$ is called convex on S_n if for any $s_1, s_2\in S_n$ and real τ, $0<\tau<1$

$$(3)\qquad f(\tau s_1+(1-\tau)s_2)\leqq \tau f(s_1)+(1-\tau)f(s_2)\ .$$

A homogeneous convex function is said to be strictly convex if equality holds in (3) iff (=if and only if) s_1 and s_2 are linearly dependent.

In our derivation the following lemma plays a fundamental role:

LEMMA 1. *For a vector $\tilde{s}\in S_n$, let $S^*(\tilde{s})$ be the subspace of all vectors $s\in S_n$ such that their ith coordinate is zero whenever the ith coordinate of $\tilde{s}$ is zero. Let, in addition, $f(s)$ be a continuous, homogeneous convex function on S_n. Then for any $\tilde{s}\in S_n$, $\tilde{s}\neq 0$ there exists a vector $w=w(\tilde{s})$ such that*

$$(4)\qquad f(s)\geqq (w(\tilde{s}), \tilde{s}-s)+f(\tilde{s})\ ,\qquad s\in S^*(\tilde{s})$$

where $(w, \tilde{s}-s)$ denotes the inner product.

If $f(s)$ is strictly convex, the equality holds in (4) iff s and $\tilde{s}$ are linearly dependent.

PROOF. This lemma is a simple consequence of Theorem 2.2.6 in [2]. This theorem ensures the existence of the vector $w(\tilde{s})$ for all interior points $\tilde{s}$ of a closed convex set in the n-dimensional space provided that f is a convex continuous function on it. In order to apply this theorem we have only to note that the restriction of $f(s)$ into the subspace $S^*(\tilde{s})$ is a homogeneous, continuous convex function on $S^*(\tilde{s})$. Obviously $S^*(\tilde{s})$ is a closed convex set and $\tilde{s}$ is an interior point thereof. Therefore, the first assertion of the lemma is true. The second statement will be proved in an indirect way. Suppose there are independent vectors s_1 and $s_2 \in S^*(s_1)$ such that

$$f(s_2)=(w(s_1),\ s_1-s_2)+f(s_1)\ . \tag{5}$$

Letting $s \overset{\Delta}{=} (s_1+s_2)/2$ we have $s \in S^*(s_1)$. Since s and s_1 are linearly independent and f is strictly convex

$$f(s)<\frac{f(s_1)+f(s_2)}{2}\ . \tag{6}$$

Substituting the right-hand side of (5) into (6) and using $(s_1-s_2)/2= s_1-s$ we get

$$f(s)<(w(s_1),\ s_1-s)+f(s_1)\ . \tag{7}$$

That is, under the condition that $f(s)$ is strictly convex, the strict inequality (4) holds when s and $\tilde{s}$ are linearly independent.

The converse, namely that under the same condition the equality in (4) holds when s and $\tilde{s}$ are linearly dependent is easily shown. (Actually, the proof runs as follows. Suppose that s_1 and s_2 are linearly dependent, say $s_2=\alpha\cdot s_1$, and the strict inequality (4) holds for s_1 and s_2. Then we have

$$\alpha f(s_1)=f(s_2)>(w(s_1),\ s_1-s_2)+f(s_1)\ ,$$

hence

$$0>(w(s_1),\ s_1-s_2)+(1-\alpha)f(s_1)=(1-\alpha)[(w(s_1),\ s_1)+f(s_1)]=(1-\alpha)f(0)=0\ ,$$

which is a contradiction.)

In the sequel this lemma will be used to prove that the f-dissimilarity has the properties (i) and (ii).

THEOREM 1. *Let $(X, \tilde{\mathcal{X}})$ be a subspace of $(X, \mathcal{X})$, and $\tilde{P}_1, \cdots, \tilde{P}_n$ be the restrictions to $(X, \tilde{\mathcal{X}})$ of the probability measures $P_1, \cdots, P_n$ defined on $(X, \mathcal{X})$. Then for the f-dissimilarity the following inequality holds:*

$$D_f(P_1, \cdots, P_n) \geqq D_f(\tilde{P}_1, \cdots, \tilde{P}_n)\ . \tag{8}$$

If f is strictly convex on S_n then equality holds in (8) iff $\tilde{\mathcal{X}}$ is a sufficient σ-algebra of $\mathcal{X}$. (For the definition of the sufficient σ-algebra we *refer to Loève* [13] *Section* 24.4, *pp.* 344–347.)

PROOF. Let us choose a probability measure μ as a dominating measure (e.g. $\mu=(P_1+\cdots+P_n)/n$) and for notational brevity let

$$\boldsymbol{p}=\{p_1(x),\cdots, p_n(x)\}$$

and

$$\tilde{\boldsymbol{p}}=\{E_\mu(p_1(x)\,|\,\tilde{\mathscr{X}}),\cdots, E_\mu(p_n(x)\,|\,\tilde{\mathscr{X}})\}$$

where $E_\mu(\cdot\,|\,\tilde{\mathscr{X}})$ denotes conditional expectation. In the notations of Lemma 1 obviously $\boldsymbol{p}\in S^*(\tilde{\boldsymbol{p}})$ with μ-probability 1. Therefore, the inequality

$$f(\boldsymbol{p})\geqq(\boldsymbol{w}(\tilde{\boldsymbol{p}}), \tilde{\boldsymbol{p}}-\boldsymbol{p})+f(\boldsymbol{p})$$

holds with μ-probability 1. For strictly convex f equality holds iff $\tilde{\boldsymbol{p}}$ and $\boldsymbol{p}$ are linearly dependent. Taking conditional expectation of both sides we have

$$\begin{aligned}E_\mu(f(\boldsymbol{p})\,|\,\tilde{\mathscr{X}})&\geqq E_\mu(w(\tilde{\boldsymbol{p}}), \tilde{\boldsymbol{p}}-\boldsymbol{p})\,|\,\tilde{\mathscr{X}})+E_\mu(f(\tilde{\boldsymbol{p}})\,|\,\tilde{\mathscr{X}})\\ &=(\boldsymbol{w}(\tilde{\boldsymbol{p}}), \tilde{\boldsymbol{p}}-E_\mu(\boldsymbol{p}\,|\,\tilde{\mathscr{X}}))+f(\tilde{\boldsymbol{p}})=f(\tilde{\boldsymbol{p}})\end{aligned}$$

with μ-probability 1. For strictly convex f equality holds iff $\boldsymbol{p}$ and $\tilde{\boldsymbol{p}}$ are linearly dependent with μ-probability 1. Taking expectation we have

$$D_f(P_1,\cdots, P_n)\geqq D_f(\tilde{P}_1,\cdots, \tilde{P}_n)\ ,$$

with equality iff $\tilde{\mathscr{X}}$ is sufficient, provided that f is strictly convex. Choosing $\tilde{\mathscr{X}}$ to be the trivial σ-algebra $\tilde{\mathscr{X}}=\{\phi, X\}$ we have

COROLLARY 1.

$$D_f(P_1,\cdots, P_n)\geqq f(1,\cdots, 1)$$

with equality iff $P_1\equiv\cdots\equiv P_n$ *provided that* f *is strictly convex.*

Remark 1. If $\varphi(t)$ is a strictly monotone increasing function on $[f(1, 1,\cdots, 1), \infty)$ then $\varphi(D_f)$ also satisfies (i) and (ii).

THEOREM 2. *Let T be a measurable transformation of $(X, \mathscr{X})$ into the measurable space $(Y, \mathscr{Y})$ and let $P_1^T,\cdots, P_n^T$ denote the measures generated by T on $(Y, \mathscr{Y})$. Then*

$$D_f(P_1,\cdots, P_n)\geqq D_f(P_1^T,\cdots, P_n^T) \tag{9}$$

with equality iff T is a sufficient transformation provided that f is strictly convex.

PROOF. Let $\tilde{\mathcal{X}}$ be the σ-algebra generated by the sets $T^{-1}(B)$ $B \in \mathcal{Y}$, and let $\tilde{P}_i$ be the restriction of P_i to $\tilde{\mathcal{X}}$. Choosing the dominating measure μ_T on $(Y, \mathcal{Y})$ as the measure generated by T and μ we have $\tilde{p}_i(x)=p_i^T(Tx)$, $i=1, 2, \cdots, n$ which means

$$D_f(\tilde{P}_1, \cdots, \tilde{P}_n(x))=D_f(P_1^T, \cdots, P_n^T) \tag{10}$$

The assertion of Theorem 2 follows from that of Theorem 1.

3. f-dissimilarity and randomization

In this section we show that the f-dissimilarity does not change when considering randomization independent of i, $i \in \{1, 2, \cdots, n\}$. If, in addition, a transformation is applied after the randomization then, in general, the f-dissimilarity decreases. For strictly convex f it does not change iff the transformation is sufficient (in Halmos-Savage sense). The kind of indirect observations we discuss in this section is sometimes referred to as observation channel, see e.g. [4].

THEOREM 3. *Let $P_1, \cdots, P_n$ be probability measures on $(X, \mathcal{X})$, and for every $x \in X$ let $R(C, x)$, $C \in \mathcal{Z}$ be given probability measures on the measurable space $(Z, \mathcal{Z})$ such that*

(a) *there is a measure ν on $(Z, \mathcal{Z})$ which dominates $R(\cdot|x)$ for every $x \in X$*

(b) *$R(C|x)$ is $\mathcal{X}$-measurable for every fixed $c \in \mathcal{Z}$.*

Let $(Y, \mathcal{Y})$ be the Cartesian product of $(X, \mathcal{X})$ and $(Z, \mathcal{Z})$. Define P_i^* as the extension of

$$P_i^*(A * C)=\int_A R(C|x)p_i(x)\mu(dx) , \qquad A \in \mathcal{X},\ C \in \mathcal{Z} \tag{11}$$

to $(Y, \mathcal{Y})$. Then

$$D_f(P_1^*, \cdots, P_n^*)=D_f(P_1, \cdots, P_n) . \tag{12}$$

PROOF. Let $p_i^*(x, z)$ be the Radon-Nikodym derivative of P_i^* with respect to the Cartesian product $\mu * \nu$. Then, obviously

$$p_i(x, z)=p_i(x)r(z/x) , \tag{13}$$

where $r(z/x)$ is the Radon-Nikodym derivative of $R(\cdot/x)$ with respect to ν. Using (13), the homogeneity of $f(\cdot)$ and Fubini's theorem, we obtain the following chain of equalities which proves the theorem

$$\begin{aligned} D_f(P_1^*, \cdots, P_n^*) &= \int_Y f(p_1^*(x, z), \cdots, p_n^*(x, z))d(\mu * \nu) , \\ &= \int_Y f(p_1(x), \cdots, p_n(x))r(z/x)d(\mu * \nu) \end{aligned}$$

$$=\int_X \left\{f(p_1(x),\cdots,p_n(x))\int_Z r(z/x)d\nu\right\}d\mu$$
$$=\int_X f(p_1(x),\cdots,p_n(x))d\mu$$
$$=D_f(p_1,\cdots,p_n)\ .$$

The following corollary follows from Theorem 3 by choosing $R(C/x)$ independently of x.

COROLLARY 2. *Let $P_1,\cdots,P_n$ resp. R be probability measures on $(X,\mathcal{X})$ resp. $(Z,\mathcal{Z})$. Define P_i^* as the product measure P_i*R on the product space $(X*Z,\mathcal{X}*\mathcal{Z})$. Then*

$$D_f(P_1,\cdots,P_n)=D_f(P_1^*,\cdots,P_n^*)\ .$$

THEOREM 4. *Suppose that the conditions of Theorem 3 are fulfilled. Define the probability measures $\bar{P}_i$ on $(Z,\mathcal{Z})$ as*

$$\bar{P}_i(C)\stackrel{\Delta}{=}\int_X R(C/x)p_i(x)\mu(dx)\ ,\qquad C\in\mathcal{Z}\ .$$

Then

$$D_f(P_1,\cdots,P_n)\geqq D_f(\bar{P}_1,\cdots,\bar{P}_n)\ . \tag{14}$$

If f is strictly convex then equality holds iff the randomized transformed observation is a sufficient one i.e.

$$p_i(x)=\bar{p}_i(z)g(x,z)\ ,\qquad i=1,2,\cdots,n\ \text{a.e. w.r.t. }\mu*\nu$$

for some function $g(x,z)$.

PROOF. Clearly $\bar{P}_i(C)=P_i^*(X*C)$, where P_i^* was given by (11) $(i=1,\cdots,n)$. Thus $\bar{P}_i$ can be considered as the restriction of P_i^* to the sub σ-algebra $\bar{\mathcal{Y}}=X*\mathcal{Z}$. Therefore Theorem 1 applies to this case, and (14) follows from (12). For strictly convex f equality holds iff $\bar{\mathcal{Y}}$ is sufficient σ-algebra of $\mathcal{Y}$ i.e. iff

$$p_i^*(x,z)=\bar{p}_i(z)g^*(x,z)\qquad i=1,2,\cdots,n,\ \text{a.e. w.r.t. }\mu*\nu$$

for some function $g^*(x,z)$. The condition of equality in (14) follows by (13).

4. Characterization of the f-dissimilarity

The main result of this section is the analogue of Theorem 3 in [9]. It is shown that the f-dissimilarity can be approximated by considering the f-dissimilarity of measures generated by finite partitions.

THEOREM 5. *Let e_k be the unit vector whose kth coordinate is 1 and let*

$$M_f=\sum_{k=1}^{n} f(e_k) .$$

Then

$$D_f(P_1,\cdots, P_n)\leqq M_f .$$

If f is strictly convex then the equality holds iff $P_1, P_2,\cdots, P_n$ are pairwise orthogonal.

PROOF. Because of the homogeneity of f we have

$$\int f(p_1(x),\cdots, p_n(x))d\mu=\int \left(\sum_{i=1}^{n} p_i(x)\right)f\left(\sum_{k=1}^{n}\left(p_k(x)e_k\Big/\sum_{j=1}^{n} p_j(x)\right)\right)d\mu . \tag{15}$$

The convexity of f implies that

$$f\left(\sum_{k=1}^{n}\left(p_k(x)\Big/\sum_{j=1}^{n} p_i(x)\right)e_k\right)\leqq\sum_{k=1}^{n}\left(p_k(x)\Big/\sum_{i=1}^{n} p_i(x)\right)f(e_k) . \tag{16}$$

Since $e_1, e_2,\cdots, e_n$ are linearly independent, equality holds in (16) for strictly convex f iff one of the weight $p_k(x)\Big/\sum_{i=1}^{n} p_i(x)$ is 1 and all the other are 0. Substituting (16) into (15) we have

$$D_f(P_1,\cdots, P_n)\leqq\sum_{k=1}^{n} f(e_k)\int p_k(x)d\mu=M_f .$$

Remark 2. For any constants a, b $(a\geqq 0)$ $\tilde{D}=aD_f+b$ is an f-dissimilarity as well. Indeed, this can be seen by considering the function $\tilde{f}(s_1,\cdots, s_n)\overset{\Delta}{=}af(s_1,\cdots, s_n)+b((s_1+s_2+\cdots+s_n)/n)$. This means that one may consider "normalized dissimilarities," that is, generating functions $\tilde{f}$ yielding dissimilarities between 0 and 1.

THEOREM 6. *It holds that*

$$D_f(P_1,\cdots, P_n)=\sup_{\mathcal{A}}\sum_{i=1}^{n} f(P_1(A_i),\cdots, P_n(A_i))$$

where the supremum is taken over all finite measurable partitions $\mathcal{A}=\{A_1,\cdots, A_m\}$ of X.

Note that the sum on the right-hand side is the f-dissimilarity of the restrictions of $P_1,\cdots, P_n$ to the algebra generated by $A_1,\cdots, A_m$.

PROOF. In this proof it will be convenient to choose $\mu=P_1+P_2+\cdots+P_n$ as a dominating measure. In this case $p_i(x)\leqq 1$, $i=1, 2,\cdots, n$. Since $f(s_1,\cdots, s_n)$ is continuous on the compact set $S_n^*=\{(s_1,\cdots, s_n):$

$0 \leqq s_n \leqq 1,\ i=1, \cdots, n\}$, it is uniformly continuous on S_n^*. This means that for any $\varepsilon>0$ there exists a partition $C_1, \cdots, C_N$ of S_n^* into n-dimensional rectangles such that the total variation of f on any of $C_1, \cdots, C_n$ is less than ε. Let $B_j=\{x: (p_1(x), \cdots, p_n(x)) \in C_j,\ j=1, \cdots N\}$. Since C_j is an n-dimensional rectangle we have

$$\frac{1}{\mu(B_j)} \int_{B_j} p_1(x)\mu(dx), \cdots, \frac{1}{\mu(B_j)} \int_{B_j} p_n(x)\mu(dx) \in C_j$$

provided $\mu(B_j)>0$, $j=1, 2, \cdots, N$. Since the contribution to the dissimilarity, of sets of μ-measure 0 is zero we may disregard such sets. Thus we will suppose that $\mu(B_i)>0$. Then

$$f(p_1(x), \cdots, p_n(x)) \leqq f\left(\frac{P_1(B_j)}{\mu(B_j)}, \cdots, \frac{P_n(B_j)}{\mu(B_j)}\right)+\varepsilon\ , \qquad x \in B_j\ .$$

Integrating both sides on B_j and summing over $j=1, \cdots, N$ we get

$$\begin{aligned} D_f(P_1, \cdots, P_n) &\leqq \sum_{j=1}^{N} f\left(\frac{P_1(B_j)}{\mu(B_j)}, \cdots, \frac{P_n(B_j)}{\mu(B_j)}\right)\mu(B_j)+\varepsilon\mu(X) \\ &= \sum_{j=1}^{N} f(P_1(B_j), \cdots, P_n(B_j))+n\varepsilon\ . \end{aligned} \tag{17}$$

Theorem 1, in turn, implies that

$$D_f(P_1, \cdots, P_n) \geqq \sum_{j=1}^{m} (P_1(A_j), \cdots, P_n(A_j)) \tag{18}$$

for all finite partitions $\{A_1, \cdots, A_m\}$ of X. (17) and (18) prove the assertion of the theorem.

Acknowledgement

We would like to acknowledge the editor's and the anonymous reviewers' many helpful suggestions which have contributed to the paper.

TECHNICAL UNIVERSITY OF BUDAPEST
MATHEMATICAL INSTITUTE OF HUNGARIAN ACADEMY OF SCIENCE

REFERENCES

[1] Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another, *J. R. Statist. Soc.*, B, **28**, 131-142.

[2] Blackwell, D. and Girschick, M. A. (1954). *Theory of Games and Statistical Decisions*, John Wiley and Sons Inc.

[3] Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification, *IEEE Trans. Inf. Theory*, IT-**13**, 21-27.

[4] Csiszár, I. (1967). Information-type measures of divergence of probability distributions and indirect observation, *Studia Sci. Math. Hung.*, **2**, 299-318.

[5] Györfi, L. and Nemetz, T. (1975). On the dissimilarity of probability measures, *Technical Report*, Math. Inst. of Hungarian Acad. of Sci.

[6] Györfi, L. and Nemetz, T. (1977). f-dissimilarity: a general class of separation measures of several probability distributions, *Topics in Information Theory*, Ed. I. Csiszar, P. Elias, North-Holland, 309-321.

[7] Ito, T. (1974). Approximate bounds of misrecognition and their computational evaluation, *Proc. of the Second International Joint Conference on Pattern Recognition*, Copenhagen.

[8] Matusita, K. (1969). The notion of affinity of several distribution, *Ann. Inst. Statist. Math.*, **19**, 181-192.

[9] Matusita, K. (1971). Some properties of affinity and application, *Ann. Inst. Statist. Math.*, **23**, 137-155.

[10] Matusita, K. (1973). Discrimination and the affinity of distributions, in *Discriminant Analysis and Applications*, Academic Press, New York, 213-223.

[11] Toussaint, G. T. (1972). Feature evaluation criteria and contextual decoding algorithms in statistical pattern recognition, Ph. D. thesis, Dept. of Elect. Engineering, University of British Columbia.

[12] Toussaint, G. T. (1974). Some properties of Matusita's measure of affinity of several distributions, *Ann. Inst. Statist. Math.*, **26**, 389-394.

[13] Loève, M. (1960). *Probability Theory*, D. van Nostrand, Princeton.