Z. Wahrscheinlichkeitstheorie verw. Geb. 14, 149–160 (1969)

Information Radius*

ROBIN SIBSON

Summary. This paper is an account of a new method of constructing measures of divergence between probability measures; the new divergence measures so constructed are called *information radius* measures. They are information-theoretic in character, and are based on the work of Rényi [8] and Csiszár [2, 3]. The divergence measure K_1 can be used for the measurement of dissimilarity in numerical taxonomy, and its application to this field is discussed in Jardine and Sibson [5]; it was this application which originally motivated the study of information radius. Other forms of information radius are related to the variation distance, and the *normal information radius* discussed in § 3 is related to Mahalanobis' D^2 Statistic. This paper is in part intended to lay the mathematical foundations for [5], but because information radius appears to be of some general interest, the investigation of its properties is here carried further than is needed for the applications discussed in [5].

1. Information Gain

The concept of information gain of order 1 is by now a familiar one in probability and statistics; details may be found in Kullback [6]. A general definition runs as follows. Let X be a set, $\mathscr{E}$ a σ-algebra of measurable subsets of X. Let μ, ν be probability measures on $(X, \mathscr{E})$ with $\nu \gg \mu$.

Definition 1.1 (1). The *information gain of order* 1, $I_1(\mu|\nu)$, is given by

$$I_1(\mu|\nu) = \int_X \log_2 \frac{d\mu}{d\nu}\, d\mu.$$

$d\mu/d\nu$ is a particular choice of the Radon-Nikodym derivative; it is clear that different choices of $d\mu/d\nu$ do not affect the value of $I_1(\mu|\nu)$ assuming that this exists, and Lemma 1.1 of [2] shows that $I_1(\mu|\nu)$ does indeed exist for all such pairs μ, ν, although it may take the value $+\infty$.

Now suppose that λ is a further probability measure on $(X, \mathscr{E})$ with $\lambda \gg \nu$. $I_1(\mu|\nu)$ may then be written in the more familiar form

$$I_1(\mu|\nu) = \int_X p \cdot \log_2 (p/q)\, d\lambda,$$

where $p = d\mu/d\lambda$, $q = d\nu/d\lambda$.

Rényi [8] has given an axiomatic characterisation for the information gain of order 1, and a more general characterisation for the information gain of order α; this is defined as follows.

Definition 1.1 (α). The *information gain of order* α, $I_\alpha(\mu|\nu)$, for $0 < \alpha < \infty$, $\alpha \neq 1$, is given by

$$I_\alpha(\mu|\nu) = \frac{1}{\alpha - 1} \log_2 \int_X \left(\frac{d\mu}{d\nu}\right)^{\alpha - 1} d\mu = \frac{1}{\alpha - 1} \log_2 \int_X p^\alpha q^{1-\alpha}\, d\lambda,$$

in the same notation as before.

* Part of this work was carried out during the author's tenure of an S.R.C. Research Studentship.

Again, Lemma 1.1 of [2] shows that this is well-defined, although it may take the value $+\infty$. For $0<\alpha<\infty$, $I_\alpha(\mu|\nu)$ has the property that it is nonnegative, and is zero if and only if $\mu=\nu$.

We now introduce a further case, that of information gain of order ∞. This was not considered by Rényi, but it is clearly a generalisation in the same spirit as those which he suggests. The main interest of the order ∞ case is that it shows how to attach an information-theoretic interpretation to the variation distance, as will be seen in § 2. First, some definitions and lemmas.

Definition 1.2. Let f be a measurable real function on $(X, \mathscr{E})$, and let λ be a probability measure on $(X, \mathscr{E})$. Define

$$\sup_\lambda f = \sup\{h\colon \lambda[f^{-1}(h,\infty)]>0\}.$$

Lemma 1.3. *If* $f=g\,[\lambda]$, *then* $\sup_\lambda f=\sup_\lambda g$. ⌟

Lemma 1.4. *If* μ, ν *are probability measures on* $(X, \mathscr{E})$, *and if* $\nu \gg \mu$, *then both*

$$\sup_\nu \frac{d\mu}{d\nu} \quad \textit{and} \quad \sup_\mu \frac{d\mu}{d\nu}$$

are well-defined and nonnegative, and they are equal.

Proof. $d\mu/d\nu \geqq 0$ almost everywhere with respect to ν, and hence with respect to μ. So both

$$\sup_\nu \frac{d\mu}{d\nu} \quad \text{and} \quad \sup_\mu \frac{d\mu}{d\nu}$$

are well-defined and nonnegative, although they may take the value $+\infty$; Lemma 1.3 shows that their values are unaffected by changing $d\mu/d\nu$ on a set of ν-measure zero. Now $\mu(A)>0$ implies $\nu(A)>0$ for all $A\in\mathscr{E}$, so immediately we have

$$\sup_\mu \frac{d\mu}{d\nu} \leqq \sup_\nu \frac{d\mu}{d\nu}.$$

Suppose

$$\sup_\mu \frac{d\mu}{d\nu} < \sup_\nu \frac{d\mu}{d\nu};$$

then there is some set $A\in\mathscr{E}$ such that $\mu(A)=0$ and $\nu(A)>0$ and

$$\frac{d\mu}{d\nu} \geqq k > \sup_\mu \frac{d\mu}{d\nu}$$

on A. But

$$0=\mu(A)=\int_A \frac{d\mu}{d\nu}\,d\nu \geqq k\cdot\nu(A).$$

Now $\nu(A)>0$, and since, by the above,

$$\sup_\mu \frac{d\mu}{d\nu} \geqq 0,$$

it follows that $k>0$, and this yields a contradiction. Hence

$$\sup_\mu \frac{d\mu}{d\nu} = \sup_\nu \frac{d\mu}{d\nu}. \quad ⌟$$

Definition 1.1 (∞). The *information gain of order* ∞, $I_\infty(\mu|\nu)$, is given by

$$I_\infty(\mu|\nu)=\log_2 \sup_\mu \frac{d\mu}{d\nu}.$$

Lemma 1.5. $I_\infty(\mu|\nu)\geqq 0$, *with equality if and only if* $\mu=\nu$.

Proof. $I_\infty(\mu|\nu)\leqq 0$ if and only if $\sup_\mu \frac{d\mu}{d\nu}\leqq 1$. By Lemma 1.4 this is so if and only if $\sup_\nu \frac{d\mu}{d\nu}\leqq 1$, that is, if and only if $\frac{d\mu}{d\nu}\leqq 1$ except on a set of ν-measure zero. Let S_h be defined by

$$S_h=\left\{x\in X:\ \frac{d\mu}{d\nu}(x)<(1-h)\right\},$$

where a particular choice has been made for $d\mu/d\nu$. $S_0=\bigcup_{n=1}^{\infty} S_{1/n}$, so $\nu(S_0)>0$ implies that $\nu(S_{1/n})>0$ for some n, since the union is countable. Now suppose that $d\mu/d\nu\leqq 1$ almost everywhere with respect to ν. If $\nu(S_0)>0$, then choose N so that $\nu(S_{1/N})>0$. Then

$$\begin{aligned}1=\mu(X)=\int_X \frac{d\mu}{d\nu}\,d\nu &\leqq \nu(X\setminus S_{1/N})+\nu(S_{1/N})\cdot\left(1-\frac{1}{N}\right)\\ &\leqq 1-\frac{1}{N}\cdot\nu(S_{1/N})\\ &<1\end{aligned}$$

and this is a contradiction. It follows that the assumption that $\nu(S_0)>0$ is false, and hence that $d\mu/d\nu=1$ almost everywhere with respect to ν, and hence that $\mu=\nu$. Thus $I_\infty(\mu|\nu)\leqq 0$ implies $\mu=\nu$, and this in turn implies $I_\infty(\mu|\nu)=0$; this result follows immediately from this. ⌟

Definition 1.6. Let μ_1, μ_2 be probability measures on $(X, \mathscr{E})$, and let λ be a measure on $(X, \mathscr{E})$ such that $\lambda \gg \mu_1, \mu_2$. The *variation distance* $\rho(\mu_1, \mu_2)$ is defined by

$$\rho(\mu_1, \mu_2)=\int_X |p_1-p_2|\,d\lambda,$$

where $p_i=d\mu_i/d\lambda$.

It is well-known that this is independent of the choice of λ, and that it is a metric on the set of probability measures on $(X, \mathscr{E})$. The induced topology is called the *variation distance topology*.

2. Information Radius

Suppose that $\mu_1, \ldots, \mu_n$ are probability measures on $(X, \mathscr{E})$, and that $w_1, \ldots, w_n \geqq 0$, with $\sum w_i > 0$. The basic idea of information radius is to use an I_α to compare each of the $\mu_1, \ldots, \mu_n$ in turn with a probability measure ν such that $\nu \gg \sum w_i \mu_i$, and to find an appropriate generalised weighted mean value for these $I_\alpha(\mu_i|\nu)$ using the w_i as weights. The information radius is the infimum of this under variation of ν.

Rényi [8] points out that ordinary information measures (that is, information measures of order 1) are based on taking a weighted mean in the usual way by formation of the weighted sum

$$\Sigma_1\begin{pmatrix} x_1, \ldots, x_n \\ w_1, \ldots, w_n \end{pmatrix} = \sum w_i x_i / \sum w_i,$$

and that information measures of order α are similarly based on

$$\Sigma_\alpha\begin{pmatrix} x_1, \ldots, x_n \\ w_1, \ldots, w_n \end{pmatrix} = \frac{1}{\alpha - 1} \log_2 \{\{\sum w_i 2^{(\alpha-1)x_i}\} / \sum w_i\},$$

and that these are the only functions of a certain kind yielding information gain and entropy measures satisfying certain postulates. A further form not considered by Rényi is given by

$$\Sigma_\infty\begin{pmatrix} x_1, \ldots, x_n \\ w_1, \ldots, w_n \end{pmatrix} = \max\{x_i \colon w_i \neq 0\}.$$

It is easy to see that Σ is a continuous function of α, and this shows why the notation Σ_∞ is employed.

Now we define, for $\nu \gg \sum w_i \mu_i$,

$$K_\alpha\left(\begin{matrix} \mu_1, \ldots, \mu_n \\ w_1, \ldots, w_n \end{matrix}\middle| \nu\right) = \Sigma_\alpha\begin{pmatrix} I_\alpha(\mu_1|\nu), \ldots, I_\alpha(\mu_n|\nu) \\ w_1, \quad \ldots, \quad w_n \end{pmatrix}$$

and

$$K_\alpha\begin{pmatrix} \mu_1, \ldots, \mu_n \\ w_1, \ldots, w_n \end{pmatrix} = \inf\left\{K_\alpha\left(\begin{matrix} \mu_1, \ldots, \mu_n \\ w_1, \ldots, w_n \end{matrix}\middle| \nu\right) \colon \nu \gg \sum w_i \mu_i\right\}.$$

If $w_1 = \cdots = w_n$, it is convenient to have available the notation $K_\alpha(\mu_1, \ldots, \mu_n|\nu)$, $K_\alpha(\mu_1, \ldots, \mu_n)$, omitting the weights. This notation is used in [5].

Definition 2.1. K_α is called the *information radius of order* α.

The next theorem is fundamental in establishing the properties of the information radius.

Theorem 2.2.

$$K_\alpha\begin{pmatrix} \mu_1, \ldots, \mu_n \\ w_1, \ldots, w_n \end{pmatrix} = K_\alpha\left(\begin{matrix} \mu_1, \ldots, \mu_n \\ w_1, \ldots, w_n \end{matrix}\middle| \nu\right)$$

for precisely one value $\nu = \nu_0$ *of* ν, *and* ν_0 *is given by*

$$q_0 = \sum w_i p_i / \sum w_i \qquad \text{if } \alpha = 1,$$

$$q_0 = \left(\sum w_i p_i^\alpha\right)^{1/\alpha} / \int_X \left(\sum w_i p_i^\alpha\right)^{1/\alpha} d\lambda \qquad \text{if } \alpha \in (0,1) \cup (1, \infty),$$

$$q_0 = \max_{w_i \neq 0} p_i / \int_X \max_{w_i \neq 0} p_i \, d\lambda \qquad \text{if } \alpha = \infty,$$

where $\lambda \gg \sum w_i \mu_i$, *and* $p_i = d\mu_i/d\lambda$, $q_0 = d\nu_0/d\lambda$.

Proof. The result follows from the identity

$$K_\alpha\left(\begin{matrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{matrix}\middle|\, v\right)-K_\alpha\left(\begin{matrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{matrix}\middle|\, v_0\right)=I_\alpha(v_0|v)$$

and the fact that I_α is positive-definite for $\alpha\in(0,\infty]$. The identity is easy to verify, but the task is rather lengthy, and details are not given here. ⌟

Corollary 2.3. *K_α takes the following forms:*

$$K_1\left(\begin{matrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{matrix}\right)=\int_X\sum_i\left\{\frac{w_i p_i}{\sum_j w_j}\log_2\frac{p_i\sum_j w_j}{\sum_j w_j p_j}\right\}d\lambda;$$

$$K_\alpha\left(\begin{matrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{matrix}\right)=\frac{1}{\alpha-1}\log_2\left\{\left\{\int_X\left[\sum w_i p_i^\alpha\right]^{1/\alpha}d\lambda\right\}^\alpha/\sum w_i\right\}\quad \text{if } \alpha\in(0,1)\cup(1,\infty);$$

$$K_\infty\left(\begin{matrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{matrix}\right)=\log_2\int_X\max_{w_i\neq 0}p_i\,d\lambda.\quad ⌟$$

Lemma 2.4. *If $n=2$, $w_1, w_2>0$, then*

$$K_\infty\left(\begin{matrix}\mu_1 & \mu_2\\ w_1 & w_2\end{matrix}\right)=\log_2\{1+\tfrac{1}{2}\rho(\mu_1,\mu_2)\}.$$

Proof. This follows immediately from the last part of Corollary 2.3. ⌟

Thus K_∞ is very closely related to the variation distance, and may be regarded as providing an information-theoretic interpretation of it. If $w_1, w_2>0$, all the other K_α are related to ρ in a weaker sense made precise in Theorem 2.7 below.

Lemma 2.5.

$$K_\alpha\left(\begin{matrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{matrix}\right)$$

is zero if and only if all μ_i for which $w_i>0$ are equal. It is symmetric in the sense that if σ is a permutation of $1,\ldots,n$, then

$$K_\alpha\left(\begin{matrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{matrix}\right)=K_\alpha\left(\begin{matrix}\mu_{\sigma(1)},\ldots,\mu_{\sigma(n)}\\ w_{\sigma(1)},\ldots,w_{\sigma(n)}\end{matrix}\right).$$

Proof. This result follows immediately from Theorem 2.2 and Corollary 2.3. ⌟

Definition 2.6.

$$N_\varepsilon(\mu_1;\alpha,w_1,w_2)=\left\{\mu_2\colon K_\alpha\left(\begin{matrix}\mu_1 & \mu_2\\ w_1 & w_2\end{matrix}\right)<\varepsilon\right\},$$

where $w_1, w_2>0$, $\varepsilon>0$.

Theorem 2.7. *For fixed α, w_1, w_2 the $N_\varepsilon(\mu_1;\alpha,w_1,w_2)$ for varying μ_1, ε form a basis for the variation distance topology.*

Proof. In the case $\alpha=\infty$ this is a corollary of Lemma 2.4. If $\alpha<\infty$, then it is enough to observe that information radius is, or is a monotone invertible continuous function of, an f-divergence in the sense of Csiszár to which Theorems 1 and 2 of [3] apply. ⌟

Theorem 2.8.

(1) $$K_1\begin{pmatrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{pmatrix}\leqq\sum_i\left\{\frac{w_i}{\sum_j w_j}\log_2\frac{\sum_j w_j}{w_i}\right\};$$

(α) $$K_\alpha\begin{pmatrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{pmatrix}\leqq\frac{1}{\alpha-1}\log_2\{[\sum w_i^{1/\alpha}]^\alpha/\sum w_i\}\quad \textit{if } \alpha\in(0,1)\cup(1,\infty);$$

(∞) $$K_\infty\begin{pmatrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{pmatrix}\leqq\log_2 m,$$

where $w_i>0$ for just m values of i. In each case the bound is attained if and only if those μ_i for which $w_i>0$ are mutually singular.

Proof. (1) $\alpha=1$

$$K_1\begin{pmatrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{pmatrix}-\sum_i\left\{\frac{w_i}{\sum_j w_j}\log_2\frac{\sum_j w_j}{w_i}\right\}$$

$$=\int_X\sum_i\left\{\frac{w_i p_i}{\sum_j w_j}\log_2\frac{p_i\sum_j w_j}{\sum_j w_j p_j}\right\}d\lambda-\sum_i\left\{\frac{w_i}{\sum_j w_j}\log_2\frac{\sum_j w_j}{w_i}\right\}$$

$$=\int_X\sum_i\left\{\frac{w_i p_i}{\sum_j w_j}\log_2\frac{p_i\sum_j w_j}{\sum_j w_j p_j}-\frac{w_i p_i}{\sum_j w_j}\log_2\frac{\sum_j w_j}{w_i}\right\}d\lambda$$

$$=\int_X\sum_i\left\{\frac{w_i p_i}{\sum_j w_j}\log_2\frac{w_i p_i}{\sum_j w_j p_j}\right\}d\lambda$$

$$=M_1,\quad \text{say}.$$

Now $w_i p_i\leqq\sum_j w_j p_j$, so $M_1\leqq 0$, whence the result follows. Clearly $M_1=0$ if and only if $w_i p_i=\sum_j w_j p_j[\mu_i]$ whenever $w_i\neq 0$; that is, if and only if all the μ_i for which $w_i\neq 0$ are mutually singular.

(α) $\alpha\in(0,1)\cup(1,\infty)$

$$K_\alpha\begin{pmatrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{pmatrix}-\frac{1}{\alpha-1}\log_2\{[\sum w_i^{1/\alpha}]^\alpha/\sum w_i\}$$

$$=\frac{1}{\alpha-1}\log_2\{\{\int_X[\sum w_i p_i^\alpha]^{1/\alpha}d\lambda\}^\alpha/\sum w_i\}-\frac{1}{\alpha-1}\log_2\{[\sum w_i^{1/\alpha}]^\alpha/\sum w_i\}$$

$$=\frac{1}{\alpha-1}\log_2\{\int_X[\sum w_i p_i^\alpha]^{1/\alpha}d\lambda/\sum(w_i^{1/\alpha})\}^\alpha$$

$$=\frac{\alpha}{\alpha-1}\log_2\{\{\int_X[\sum w_i p_i^\alpha]^{1/\alpha}d\lambda\}\cdot\{\int_X[\sum(w_i p_i^\alpha)^{1/\alpha}]d\lambda\}^{-1}\}$$

$$=M_\alpha,\quad \text{say}.$$

Now if $x_1, \ldots, x_n > 0$, we have

$$\left(\sum x_i\right)^k \geqq \sum x_i^k \qquad \text{for all } k > 1,$$

and

$$\left(\sum x_i\right)^k \leqq \sum x_i^k \qquad \text{for all } k \in (0, 1),$$

with equality if and only if $x_1 = \cdots = x_n$. This is a standard inequality, and may be found, for example, in Hardy, Littlewood, and Pólya [7, p. 28]. It follows from this inequality that $M_\alpha \leqq 0$, with equality if and only if all the μ_i for which $w_i \neq 0$ are mutually singular.

(∞) $\alpha = \infty$

$$\begin{aligned} K_\infty\left(\begin{matrix}\mu_1, \ldots, \mu_n \\ w_1, \ldots, w_n\end{matrix}\right) &= \log_2 \int_X \max_{w_i \neq 0} p_i \, d\lambda \\ &\leqq \log_2 \int_X \sum_{w_i \neq 0} p_i \, d\lambda \\ &\leqq \log_2 m, \end{aligned}$$

where $w_i \neq 0$ for just m values of i. Again, equality holds if and only if all μ_i for which $w_i \neq 0$ are mutually singular. ⌟

Theorem 2.9.

$$K_\alpha\left(\begin{matrix}\mu_1, \ldots, \mu_n \\ w_1, \ldots, w_n\end{matrix}\right) \leqq \log_2 n.$$

Proof. (1) $\alpha = 1$. In this case the result required is simply the standard entropy inequality for information of order 1, applied to the $w_i / \sum w_j$ as though they were probabilities, given the result of Theorem 2.8(1). Details of the entropy inequality may be found in Feinstein [4, p. 15].

(α) $\alpha \in (0, 1) \cup (1, \infty)$. In this case the bound given by Theorem 2.8(α) can be written in the form

$$\frac{\alpha}{\alpha - 1} \log_2 \sum_i \left[w_i / \sum_j w_j\right]^{1/\alpha}.$$

Now $\sum_i \left[w_i / \sum_j w_j\right]^k$ takes its stationary value (maximum if $k < 1$, minimum if $k > 1$) for $w_1 = \cdots = w_n$, and the result follows from this.

(∞) $\alpha = \infty$. This case is a trivial corollary of Theorem 2.8(∞). ⌟

3. Relative Information Radius

Ordinary information radius is an infimum value obtained as the probability measure ν varies over all $\nu \gg \sum w_i \mu_i$. The definition may be relativised by requiring that ν satisfy further conditions; in order to obtain a meaningful result it may be necessary to place some restrictions on $\mu_1, \ldots, \mu_n$. If ν is restricted to lie in some subset $\mathscr{R}$ of the set of all probability measures on $(X, \mathscr{E})$, the resultant infimum value is called the *rel $\mathscr{R}$ information radius.* No attempt will be made here to explore the general properties of relative information radius; instead, two special cases of order 1 relative information radius will be considered in detail. We take $\alpha = 1$ throughout this section.

The first special case to be considered is that of *product information radius*, denoted by $K^\times$. Let $(X_1, \mathscr{E}_1), \ldots, (X_p, \mathscr{E}_p)$ be probability spaces, and let $(X, \mathscr{E})$ be their product. The product information radius is obtained by requiring that ν be of the form $\nu_1 \times \cdots \times \nu_p$, where for each j the factor ν_j is a probability measure on $(X_j, \mathscr{E}_j)$. The main interest of product information radius is that it satisfies an additivity theorem, as follows.

Theorem 3.1. *Let $\mu_1, \ldots, \mu_n$ be probability measures on $(X, \mathscr{E})$ with $\mu_i = \mu_{i1} \times \cdots \times \mu_{ip}$, μ_{ij} being a probability measure on $(X_j, \mathscr{E}_j)$. Then*

$$K^\times \begin{pmatrix} \mu_1, \ldots, \mu_n \\ w_1, \ldots, w_n \end{pmatrix} = \sum_j K_1 \begin{pmatrix} \mu_{1j}, \ldots, \mu_{nj} \\ w_1, \ldots, w_n \end{pmatrix}.$$

Proof. This result is an easy consequence of the additivity theorems for I_1; see, for example, Kullback [6, p. 12]. ⌟

From the identity given in Theorem 2.2 above, it follows that

$$K^\times \begin{pmatrix} \mu_1, \ldots, \mu_n \\ w_1, \ldots, w_n \end{pmatrix} = K_1 \begin{pmatrix} \mu_1, \ldots, \mu_n \\ w_1, \ldots, w_n \end{pmatrix} + I_1(\nu_0 | \nu_1 \times \cdots \times \nu_p),$$

where $\mu_1, \ldots, \mu_n$ are as in Theorem 3.1 and where $\nu_0 = \sum w_i \mu_i / \sum w_i$ and $\nu_j = \sum w_i \mu_{ij} / \sum w_i$. Thus $K^\times \geqq K_1$, the difference being a term $I_1(\nu_0 | \nu_1 \times \cdots \times \nu_p)$.

Discussion of the second special case and of its relation to other divergence measures will occupy the rest of this section. The second case is that of *normal information radius*, denoted by N. Let X be m-dimensional Euclidean space, $\mathscr{E}$ the σ-algebra of Borel sets. We shall be concerned only with probability measures on $(X, \mathscr{E})$ which are absolutely continuous with respect to Lebesgue measure, and which possess mean vectors and positive-definite covariance matrices, and have finite entropy relative to Lebesgue measure; such probability measures will be called *good*. The notation in Chapter 9 of Kullback [6] will be followed as far as possible, with the exception that our use of μ to denote a probability measure precludes its use to denote a mean, for which purpose β will be used. Σ will, as in Kullback, denote a covariance matrix; note that the entry Σ_{rr} is the variance along the r-axis, not the standard deviation. The transpose operator on a matrix or vector will be denoted by superscript T.

The multivariate normal probability measure with mean β and (nonsingular) covariance matrix Σ will be denoted by $\mathscr{N}(\beta, \Sigma)$. Its density function with respect to Lebesgue measure is given by

$$p(x) = (\det 2\pi\Sigma)^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(x-\beta)^\mathsf{T} \Sigma^{-1}(x-\beta)\}.$$

If μ is an arbitrary good probability measure, the multivariate normal probability measure with the same mean and covariance will be written $\mathscr{N}(\mu)$.

Lemma 3.2. *Let μ_1 be a good probability measure with mean β_1 and covariance Σ_1, and with entropy $\mathscr{H}(\mu_1)$. Then*

$$I_1(\mu_1 | \mathscr{N}(\beta, \Sigma)) = -\mathscr{H}(\mu_1) + (2 \log 2)^{-1} \{\log(\det 2\pi\Sigma) + (\beta_1 - \beta)^\mathsf{T} \Sigma^{-1}(\beta_1 - \beta) + \operatorname{tr}(\Sigma^{-1} \Sigma_1)\}.$$

Proof. Clearly we have

$$I_1(\mu_1|\mathcal{N}(\beta,\Sigma)) = -\mathcal{H}(\mu_1) + (\log 2)^{-1}\int(-\log p)\,d\mu_1$$

where p is the density function for $\mathcal{N}(\beta,\Sigma)$. But

$$-\log p(x) = \tfrac{1}{2}\log(\det 2\pi\Sigma) + \tfrac{1}{2}(x-\beta)^{\mathrm{T}}\Sigma^{-1}(x-\beta)$$

so we have

$$I_1(\mu_1|\mathcal{N}(\beta,\Sigma)) = -\mathcal{H}(\mu_1) + (2\log 2)^{-1}\{\log(\det 2\pi\Sigma) + \int(x-\beta)^{\mathrm{T}}\Sigma^{-1}(x-\beta)\,d\mu_1\}.$$

Now

$$\begin{aligned}\int(x-\beta)^{\mathrm{T}}\Sigma^{-1}(x-\beta)\,d\mu_1 &= \int(x-\beta_1)^{\mathrm{T}}\Sigma^{-1}(x-\beta_1)\,d\mu_1 + (\beta_1-\beta)^{\mathrm{T}}\Sigma^{-1}(\beta_1-\beta)\\ &= \int \mathrm{tr}\,[\Sigma^{-1}(x-\beta_1)(x-\beta_1)^{\mathrm{T}}]\,d\mu_1 + (\beta_1-\beta)^{\mathrm{T}}\Sigma^{-1}(\beta_1-\beta)\\ &\quad [\text{because } x^{\mathrm{T}}Ax = \mathrm{tr}(Axx^{\mathrm{T}})]\\ &= \mathrm{tr}(\Sigma^{-1}\Sigma_1) + (\beta_1-\beta)^{\mathrm{T}}\Sigma^{-1}(\beta_1-\beta),\end{aligned}$$

whence it follows that $I_1(\mu_1|\mathcal{N}(\beta,\Sigma))$ has the form stated in the theorem. ⌟

Corollary 3.3. *For fixed μ_1, $I_1(\mu_1|\mathcal{N}(\beta,\Sigma))$ is minimised by taking $\beta=\beta_1$, $\Sigma=\Sigma_1$, and the resultant minimum value is*

$$I_1(\mu_1|\mathcal{N}(\mu_1)) = N(\mu_1) = -\mathcal{H}(\mu_1) + \tfrac{1}{2}\log_2(\det 2\pi e\,\Sigma_1). \quad ⌟$$

Corollary 3.4. *Among all good probability measures with given mean and covariance, the multivariate normal probability measure has uniquely the largest entropy, namely $\frac{1}{2}\log_2(\det 2\pi e\,\Sigma)$.* ⌟

The normal information radius is obtained by restricting v to lie in the set of multivariate normal probability measures. A formal definition runs as follows.

Definition 3.5. Let $\mu_1,\ldots,\mu_n$ be good probability measures and suppose $w_1,\ldots,w_n \geqq 0$ with $\sum w_i > 0$. Define

$$N\begin{pmatrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{pmatrix} = \inf K_1\left(\begin{pmatrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{pmatrix}\middle|\mathcal{N}(\beta,\Sigma)\right);$$

N is called *normal information radius*. If $w_1=\cdots=w_n$, the notation $N(\mu_1,\ldots,\mu_n)$ will be used. It is clear that this conforms with the usage $N(\mu_1)=I_1(\mu_1|\mathcal{N}(\mu_1))$ in Corollary 3.3.

Lemma 3.6.

$$N\begin{pmatrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{pmatrix} = N\begin{pmatrix}\mathcal{N}(\mu_1),\ldots,\mathcal{N}(\mu_n)\\ w_1,\quad\ldots,\quad w_n\end{pmatrix} + \sum w_i\ N(\mu_i)/\sum w_i.$$

Proof.

$$N\begin{pmatrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{pmatrix}=\inf\{\sum w_i I_1(\mu_i|\mathcal{N}(\beta,\Sigma))/\sum w_i\}$$
$$=-\sum w_i\mathcal{H}(\mu_i)/\sum w_i+(2\log 2)^{-1}\inf\{\log(\det 2\pi\Sigma)$$
$$+\sum w_i(\beta_i-\beta)^{\mathrm{T}}\Sigma^{-1}(\beta_i-\beta)/\sum w_i+\sum w_i\,\mathrm{tr}(\Sigma^{-1}\Sigma_i)/\sum w_i\}$$
$$=-\sum w_i\mathcal{H}(\mu_i)/\sum w_i+[\sum w_i\log_2(\det 2\pi e\,\Sigma_i)]/2\sum w_i$$
$$-[\sum w_i\log_2(\det 2\pi e\,\Sigma_i)]/2\sum w_i+(2\log 2)^{-1}\inf\{\log(\det 2\pi\Sigma)$$
$$+\sum w_i(\beta_i-\beta)^{\mathrm{T}}\Sigma^{-1}(\beta_i-\beta)/\sum w_i+\sum w_i\,\mathrm{tr}(\Sigma^{-1}\Sigma_i)/\sum w_i\}$$
$$=\sum w_i N(\mu_i)/\sum w_i+N\begin{pmatrix}\mathcal{N}(\mu_1),\ldots,\mathcal{N}(\mu_n)\\ w_1,\quad\ldots,\quad w_n\end{pmatrix}.\quad\square$$

This lemma shows that general normal information radius can be expressed as the sum of two terms, one arising from the non-normality of the probability measures, and the other from the failure of their first and second moments to coincide.

Lemma 3.7. *There exists precisely one pair* $(\beta,\Sigma)=(\beta_0,\Sigma_0)$ *given by*

$$\beta_0=\sum w_i\beta_i/\sum w_i,$$
$$\Sigma_0=\sum w_i[(\beta_i-\beta_0)(\beta_i-\beta_0)^{\mathrm{T}}+\Sigma_i]/\sum w_i$$

for which

$$N\begin{pmatrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{pmatrix}=K_1\left(\begin{pmatrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{pmatrix}\middle|\mathcal{N}(\beta,\Sigma)\right).$$

Proof.

$$K_1\left(\begin{pmatrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{pmatrix}\middle|\mathcal{N}(\beta,\Sigma)\right)=\sum w_i I_1(\mu_i|\mathcal{N}(\beta,\Sigma))/\sum w_i$$
$$=-\sum w_i\mathcal{H}(\mu_i)/\sum w_i+(2\log 2)^{-1}\{\log(\det 2\pi\Sigma)$$
$$+\sum w_i(\beta_i-\beta)^{\mathrm{T}}\Sigma^{-1}(\beta_i-\beta)/\sum w_i+\sum w_i\,\mathrm{tr}(\Sigma^{-1}\Sigma_i)/\sum w_i\}.$$

For any choice of Σ this is easily seen to be minimised by $\beta=\beta_0$, where β_0 is as above; it may then be written

$$K_1\left(\begin{pmatrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{pmatrix}\middle|\mathcal{N}(\beta_0,\Sigma)\right)=-\sum w_i\mathcal{H}(\mu_i)/\sum w_i+(2\log 2)^{-1}\{\log(\det 2\pi\Sigma)$$
$$+\mathrm{tr}\{\Sigma^{-1}[\sum w_i(\{\beta_i-\beta_0\}\{\beta_i-\beta_0\}^{\mathrm{T}}+\Sigma_i)/\sum w_i]\}\},$$

which is minimised by taking $\Sigma=\Sigma_0$. $\square$

Corollary 3.8.

$$N\begin{pmatrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{pmatrix}=-\sum w_i\mathcal{H}(\mu_i)/\sum w_i+\tfrac{1}{2}\log_2(\det 2\pi e\,\Sigma_0).$$

In particular, if $\mu_i=\mathcal{N}(\beta_i,\Sigma_i)$, *then*

$$N\begin{pmatrix}\mu_1,\ldots,\mu_n\\ w_1,\ldots,w_n\end{pmatrix}=\tfrac{1}{2}\log_2(\det\Sigma_0)-\tfrac{1}{2}[\sum w_i\log_2(\det\Sigma_i)]/\sum w_i.\quad\square$$

Lemma 3.7 and Corollary 3.8 between them play a rôle in the discussion of normal information radius similar to the rôle of Theorem 2.2 and Corollary 2.3 in the discussion of ordinary information radius.

A special case of normal information radius which is of particular interest because of its relation to Mahalanobis' D^2 Statistic is that given by taking $n=2$, $w_1=w_2$, $\mu_i=\mathcal{N}(\beta_i, \Sigma_i)$ $(i=1, 2)$. In this case the expression given for N in Corollary 3.8 reduces to

$$N(\mu_1, \mu_2)=\tfrac{1}{2}\log_2\left\{\frac{\det\{\frac{1}{2}(\Sigma_1+\Sigma_2)+\frac{1}{4}(\beta_1-\beta_2)(\beta_1-\beta_2)^{\mathrm{T}}\}}{(\det\Sigma_1)^{\frac{1}{2}}(\det\Sigma_2)^{\frac{1}{2}}}\right\}.$$

Lemma 3.9.

$$N(\mu_1, \mu_2)=\tfrac{1}{2}\log_2\left\{\frac{\det\frac{1}{2}(\Sigma_1+\Sigma_2)}{(\det\Sigma_1)^{\frac{1}{2}}(\det\Sigma_2)^{\frac{1}{2}}}\right\}$$
$$+\tfrac{1}{2}\log_2\{1+\tfrac{1}{4}(\beta_1-\beta_2)^{\mathrm{T}}[\tfrac{1}{2}(\Sigma_1+\Sigma_2)]^{-1}(\beta_1-\beta_2)\}.$$

Proof.

$$\det\{\tfrac{1}{2}(\Sigma_1+\Sigma_2)+\tfrac{1}{4}(\beta_1-\beta_2)(\beta_1-\beta_2)^{\mathrm{T}}\}=\det\{\tfrac{1}{2}(\Sigma_1+\Sigma_2)\}\det(\mathrm{I}+a\,a^{\mathrm{T}}),$$

where $a=[\frac{1}{2}(\Sigma_1+\Sigma_2)]^{-\frac{1}{2}}[\frac{1}{2}(\beta_1-\beta_2)]$; $\frac{1}{2}(\Sigma_1+\Sigma_2)$ is positive-definite symmetric, so there is a unique positive-definite symmetric matrix $[\frac{1}{2}(\Sigma_1+\Sigma_2)]^{\frac{1}{2}}$ whose square is $\frac{1}{2}(\Sigma_1+\Sigma_2)$. But, for any vector x, $\det(\mathrm{I}+x\,x^{\mathrm{T}})=1+x^{\mathrm{T}}x$, so we have

$$\det(\mathrm{I}+a\,a^{\mathrm{T}})=1+\tfrac{1}{4}(\beta_1-\beta_2)^{\mathrm{T}}[\tfrac{1}{2}(\Sigma_1+\Sigma_2)]^{-1}(\beta_1-\beta_2)$$

and the result follows. ⌟

This lemma shows that $N(\mu_1, \mu_2)$ can be expressed as a sum of two terms, the first of which can be regarded as measuring the difference between the covariance matrices of μ_1 and μ_2, and the second as measuring the standardised difference between the means. If $\Sigma_1=\Sigma_2=\Sigma$, the first term vanishes, and the second becomes

$$\tfrac{1}{2}\log_2\{1+\tfrac{1}{4}(\beta_1-\beta_2)^{\mathrm{T}}\Sigma^{-1}(\beta_1-\beta_2)\}=\tfrac{1}{2}\log_2(1+\tfrac{1}{4}D^2),$$

where D^2 is Mahalanobis' D^2 Statistic.

Ali and Silvey [1] show that there is a wide class of divergence measures of a certain form which reduce to a monotone function of Mahalanobis' D^2 Statistic in the case of multivariate normal distributions of equal covariance. Normal information radius is not a divergence measure of the form considered by Ali and Silvey, neither can Csiszár's general treatment [2, 3] be applied to it; this is because it is not in the form of an expectation of a function of the likelihood ratio. Thus the result that in the case of equal covariance the information radius reduces to $\frac{1}{2}\log_2(1+\frac{1}{4}D^2)$ is more than merely the explicit form of a result deducible from general theory. Any divergence measure which reduces essentially to Mahalanobis' D^2 in the equal covariance case may be regarded as providing a generalisation of D^2, and different generalisations may be of interest in different situations. Normal information radius is appropriate to the study of the possibility of pooling normal populations to form larger populations which are to be treated as though they were normal.

It is a pleasure to acknowledge my indebtedness to Professor D. G. Kendall for his advice and helpful criticism in connection with this paper.

References

1. Ali S. M., Silvey, S. D.: A general class of coefficients of divergence of one distribution from another. J. roy. statist. Soc. Ser. B **27**, 131 – 142 (1966).
2. Csiszár, I.: Information-type measures of difference of probability distributions and indirect observations. Studia Sci. Math. Hungar. **2**, 299 – 318 (1967).
3. – On topological properties of f-divergences. Studia Sci. Math. Hungar. **2**, 329 – 339 (1967).
4. Feinstein, A.: Foundations of Information Theory. New York: McGraw Hill 1958.
5. Jardine, N., Sibson, R.: The measurement of dissimilarity. (To appear.)
6. Kullback, S.: Information theory and statistics. New York: Wiley 1959.
7. Hardy, G. H., Littlewood, J. E., Pólya, G.: Inequalities. Cambridge University Press 1934.
8. Rényi, A.: On measures of entropy and information. Proc. 4th Berkeley Sympos. math. Statist. Probab. 547 – 561 (1961).

R. Sibson
King's College Research Centre
King's College
Cambridge, England

(Received December 6, 1968)