DIFFERENTIAL METRICS IN PROBABILITY SPACES

C. R. Rao*

*Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, PA

1. INTRODUCTION

In an early paper (Rao, 1945), the author introduced a Riemannian (quadratic differential) metric over the space of a parametric family of probability distributions and proposed the geodesic distance induced by the metric as a measure of dissimilarity between probability distributions. The metric was based on the Fisher information matrix and it arose in a natural way through the concepts of statistical discrimination (see also Rao, 1949, 1954, 1973 pp. 329-332, 1982a). Such a choice of the quadratic differential metric, which we will refer to as the information metric, has indeed some attractive properties such as invariance for transformation of the variables as well as the parameters. It also seems to provide an appropriate (informative) geometry on the probability space for studying large sample properties of estimators of parameters in terms of simple loss functions as demonstrated by Amari (1982, 1983), Cencov (1982), Efron (1975, 1982), Eguchi (1983, 1984), Kass (1981) and others. Kass (1980, Ph.D. thesis) explores the possibility of using differential geometric ideas in statistical inference.

The geodesic distances based on the information metric have been computed for a number of parametric family of distributions in recent papers by Atkinson and Mitchell (1981), Burbea (1986), Kass (1981), Mitchell and Krzanowski (1985), and Oller and Cuadras (1985).

In two papers, Burbea and Rao (1982a, 1982b) gave some general methods for constructing quadratic differential metrics on probability spaces, of which the Fisher information metric belonged to a special class. In view of the rich variety of possible metrics, it would be useful to lay down some criteria for the choice of an appropriate metric for a given problem. Amari has stated that a metric should reflect the stochastic and statistical properties of the family of probability distributions. In particular he emphasized the invariance of the metric under transformations of the variables as well as the parameters. Čencov (1972) shows that the Fisher information metric is unique under some conditions including invariance. Burbea and Rao (1982a) showed that the Fisher information metric is the only metric associated with invariant divergence measures of the type introduced by Ciszâr (1967). However, there exist other types of invariant metrics as shown in Section 3 of this paper.

The choice of a metric naturally depends on a particular problem under investigation, and invariance may or may not be relevant. For instance, consider the space of multinomial distributions, $\Delta = \{(p_1,\ldots,p_n)\colon\ p_i > 0, \Sigma p_i = 1\}$, which is a submanifold of the positive orthant, $X = \{(x_1,\ldots,x_n)\colon x_i > 0\}$ of the Euclidean space R^n. A Riemannian metric on X automatically provides a metric on the submanifold Δ. In a study of linkage and selection of gametes in a biological population, Shahshahani (1979) considered the metric

$$ds^2 = \sum_1^n \frac{\Sigma x_i}{x_i}\, dx_i^2 \tag{1.1}$$

which induces the information metric on Δ. This metric provided a convenient framework for a discussion of certain biological problems. However, Nei (1978) considered a distance measure associated with the Euclidean metric

$$ds^2 = \Sigma dx_i^2 \tag{1.2}$$

which he found to be more appropriate for evolutionary studies in biology. The metric induced on Δ by (1.2) is not the Fisher information metric. Rao (1982a, 1982b) has shown that a more general type of metric

$$\Sigma\Sigma a_{ij} dx_i dx_j \tag{1.3}$$

called the quadratic entropy is more meaningful in certain sociometric and biometric studies.

The object of the present paper is to provide some general methods of constructing Riemannian metrics on probability spaces, and discuss in particular the metric generated by the quadratic entropy which is an ideal measure of diversity (see Lau, 1985 and Rao, 1982b), and has properties similar to the information metric, like invariance. We also give a list of geodesic distances based on the information metric computed by various authors (Atkinson and Mitchell, 1981; Burbea, 1986; Mitchell and Krzanowski, 1985; Oller and Cuadras, 1985 and Rao, 1945).

The basic approach adopted in the paper is first to define a measure of divergence or dissimilarity between two probability measures, and then to use it to derive a metric on M, the manifold of parameters, by considering two distributions defined by two contiguous points in M. We thus provide a method for the construction of an appropriate geometry or geometries on the parameter space for discussion of practical problems. Some divergence measures may be more appropriate for discussing properties of estimators using simple loss functions while others may be appropriate in the study of population dynamics in biology. It is not unusual in practice to study a problem under different models for observed data to examine consistency and robustness of results. The variety of metrics reported in the paper would be of some use in this direction.

2. JENSEN DIFFERENCE AND ENTROPY DIFFERENTIAL METRIC

Let ν be a σ-finite additive measure defined on a σ-algebra of subsets of a measurable space $\underline{X}$, and $\underline{P}$ be the usual Lebesgue space of ν measurable density functions,

$$\underline{P} = \{p(x): \ p(x) \geq 0, \ x\varepsilon\underline{X}, \ \int_{\underline{X}} p(x)d\nu(x) = 1\} \ . \tag{2.1}$$

We call H: $\underline{P}\to R$ an entropy (functional) on $\underline{P}$ if

(i) $H(p) = 0$ when p is degenerate,

(ii) $H(p)$ is concave on $\underline{P}$.

In such a case, with $\lambda \geq 0$, $\mu \geq 0$, $\lambda + \mu = 1$, Rao (1982a) defined the Jensen difference between p and $q\varepsilon\underline{P}$ as

$$J(\lambda,\mu; \ p,q) = H(\lambda p + \mu q) - \lambda H(p) - \mu H(q) \ . \tag{2.2}$$

The function J: $\underline{P} \times \underline{P}\to R$ is non-negative and vanishes if $p = q$ (iff $p = q$ when H is strictly concave). If the entropy function H is regarded as a measure of diversity within a population, then the Jensen difference J can be interpreted as a measure of diversity (or dissimilarity) between two populations. For the use of Jensen difference in the measurement, apportionment and analysis of diversity between populations, the reader is referred to Rao (1982a, 1982b).

Let us now consider a subset of probability densities characterized by a vector parameter θ

$$\underline{P}_\theta = \{p(x,\theta): \ p(x,\theta)\varepsilon\underline{P}, \ \theta\varepsilon M, \text{ a manifold in } R^n\}$$

and assume that $p(x,\theta)$ is a smooth function admitting derivatives of a certain order with respect to θ and differention under the integral sign. For convenience of notation, we write

$$p(\cdot,\theta) = p_\theta,\ H(\theta) = H(p_\theta),\ H(\theta,\phi) = H(\lambda p_\theta + \mu p_\phi)$$

$$J(\theta,\phi) = H(\theta,\phi) - \lambda H(\theta) - \mu H(\phi) \tag{2.3}$$

where $\theta,\phi \in M$. Putting $\phi = \theta + d\theta$ and denoting the i-th component of a vector with a subscript i, we consider the formal expansion of $J(\theta,\theta+d\theta)$,

$$\frac{1}{2!}\sum_1^n\sum_1^n \frac{\partial^2 J(\theta,\phi=\theta)}{\partial\phi_i\partial\phi_j} d\theta_i d\theta_j + \frac{1}{3!}\sum_1^n\sum_1^n\sum_1^n \frac{\partial^3 J(\theta,\phi=\theta)}{\partial\phi_i\partial\phi_j\partial\phi_k} d\theta_i d\theta_j d\theta_k + \dots$$

$$= \frac{1}{2!}\Sigma\Sigma\, g^H_{ij}(\theta) d\theta_i d\theta_j + \frac{1}{3!}\ \Sigma\Sigma\Sigma\, c^H_{ijk}(\theta) d\theta_i d\theta_j d\theta_k + \dots \tag{2.4}$$

In (2.4), the coefficients of the first order differentials vanish since $J(\theta,\phi)$ has a minimum at $\phi = \theta$, and the notation such as $\partial^2 J(\theta,\phi=\theta)/\partial\phi_i\partial\phi_j$ is used for replacing ϕ by θ after carrying out the indicated differentiations.

From the definition of the J function, it follows that the (g^H_{ij}) is a non-negative definite matrix and obeys the tensorial law under transformation of parameters. We define the matrix and the associated differential metric

$$(g^H_{ij}) \text{ and } \Sigma\Sigma\, g^H_{ij} d\theta_i d\theta_j \tag{2.5}$$

as the H-entropy information matrix and H-entropy differential metric respectively. We prove the following theorem which provides an alternative computation of the H-information matrix directly from a given entropy H.

<u>Theorem 2.1</u>

$$g^H_{ij}(\theta) = - \left.\frac{\partial^2 H(\lambda p_\theta + \mu p_\phi)}{\partial\theta_i\partial\phi_j}\right|_{\phi=\theta} . \tag{2.6}$$

<u>Proof</u>: By definition

$$\begin{aligned} g^H_{ij}(\theta) &= \frac{\partial^2 J(\theta,\phi=\theta)}{\partial\phi_i\partial\phi_j} \\ &= \frac{\partial^2 H(\theta,\phi=\theta)}{\partial\phi_i\partial\phi_j} - \mu\, \frac{\partial^2 H(\phi=\theta)}{\partial\phi_i\partial\phi_j} . \end{aligned} \tag{2.7}$$

Since $J(\theta,\phi)$ attains a minimum at $\phi = \theta$

$$\frac{\partial H(\theta,\phi=\theta)}{\partial\phi_j} = \mu\, \frac{\partial H(\theta)}{\partial\theta_j} \tag{2.8}$$

Differentiating both sides of (2.8) with respect to θ_i we have

$$\frac{\partial^2 H(\theta,\phi=\theta)}{\partial\theta_i\partial\phi_j} + \frac{\partial^2 H(\theta,\phi=\theta)}{\partial\phi_i\partial\phi_j} = \frac{\partial^2 H(\theta)}{\partial\theta_i\partial\theta_j} \tag{2.9}$$

which gives (2.6), and the desired result is proved.

Let us consider a general entropy function of the type

$$H(p_\theta) = - \int h(p_\theta) d\nu(x) \tag{2.10}$$

where h", the second derivative of h, is a non-negative function. Then using (2.6)

$$\begin{aligned} g_{ij}^H(\theta) = g_{ij}^h(\theta) &= - \frac{\partial^2 H(\theta,\phi=\theta)}{\partial\theta_i \partial\phi_j} \\ &= \int \frac{\partial^2 h(\lambda p_\theta + \mu p_\theta)}{\partial\theta_i \partial\phi_j} \Big|_{\phi=\theta} d\nu(x) \\ &= \lambda\mu \int h''(p_\theta) \frac{\partial p_\theta}{\partial\theta_i} \frac{\partial p_\theta}{\partial\theta_j} d\nu(x) . \end{aligned} \tag{2.11}$$

If $h(x) = x \log x$, leading to Shannon's entropy, then

$$g_{ij}^h = g_{ij}(\theta) = \lambda\mu \int \frac{1}{p_\theta} \frac{\partial p_\theta}{\partial\theta_i} \frac{\partial p_\theta}{\partial\theta_j} d\nu(x) \tag{2.12}$$

become the elements of Fisher's information matrix. If $h(x) = (\alpha-1)^{-1}(x^\alpha - x)$, $\alpha \neq 1$, we have the α-order entropy of Havrda and Charvát (1967) and

$$g_{ij}^h = g_{ij}^\alpha(\theta) = \alpha\lambda\mu \int p^\alpha \frac{\partial \log p_\theta}{\partial\theta_i} \frac{\partial \log p_\theta}{\partial\theta_j} d\nu(x) \tag{2.13}$$

which provide the elements of α-order entropy information matrix, and the corresponding differential metric given in Burbea and Rao (1982a, 1982b).

We prove Theorem 2.2 which gives alternative expressions for the coefficients of the third order differentials in the expansion of $J(\theta,\phi)$.

Theorem 2.2.

$$c_{ijk}^H = - \left[\frac{\partial^3 H(\theta,\phi=\theta)}{\partial\theta_i \partial\theta_j \partial\phi_k} + \frac{\partial^3 H(\theta,\phi=\theta)}{\partial\theta_i \partial\phi_j \partial\phi_k} + \frac{\partial^3 H(\theta,\phi=\theta)}{\partial\theta_j \partial\phi_i \partial\phi_k} \right] \tag{2.14}$$

Proof: By definition

$$\begin{aligned} c_{ijk}^H(\theta) &= \frac{\partial^3 J(\theta,\phi=\theta)}{\partial\phi_i \partial\phi_j \partial\phi_k} \\ &= \frac{\partial^3 H(\theta,\phi=\theta)}{\partial\phi_i \partial\phi_j \partial\phi_k} - \mu \frac{\partial^3 H(\theta)}{\partial\theta_i \partial\theta_j \partial\theta_k} \end{aligned} \tag{2.15}$$

From (2.9), writing i = j and j = k we have

$$\frac{\partial^2 H(\theta,\phi=\theta)}{\partial\theta_j \partial\phi_k} + \frac{\partial^2 H(\theta,\phi=\theta)}{\partial\phi_j \partial\phi_k} = \mu \frac{\partial^2 H(\theta)}{\partial\theta_j \partial\theta_k} .$$

Differentiating with respect to θ_i

$$\frac{\partial^3 H(\theta,\phi=\theta)}{\partial\theta_i \partial\theta_j \partial\phi_k} + \frac{\partial^3 H(\theta,\phi=\theta)}{\partial\phi_i \partial\theta_j \partial\phi_k} + \frac{\partial^3 H(\theta,\phi=\theta)}{\partial\theta_i \partial\phi_j \partial\phi_k} + \frac{\partial^3 H(\theta,\phi=\theta)}{\partial\phi_i \partial\phi_j \partial\phi_k} = \mu \frac{\partial^3 H(\theta)}{\partial\theta_i \partial\theta_j \partial\theta_k}$$

which gives (2.14) as equivalent to (2.15). This proves Theorem 2.2.

Let H be Shannon's entropy. Then, an easy computation gives

$$c_{ijk} = \lambda\mu\{[\Gamma^{(1)}_{ijk} + (1-\lambda)T_{ijk}] + [\Gamma^{(1)}_{jki} + (1-\mu)T_{ijk}] + [\Gamma^{(1)}_{ikj} + (1-\mu)T_{ijk}]\} \tag{2.16}$$

where

$$\Gamma^{(1)}_{ijk} = E(\frac{\partial^2 \log p_\theta}{\partial\theta_i \partial\theta_j} \frac{\partial \log p_\theta}{\partial\theta_k}) , \quad T_{ijk} = E(\frac{\partial \log p_\theta}{\partial\theta_i} \frac{\partial \log p_\theta}{\partial\theta_j} \frac{\partial \log p_\theta}{\partial\theta_k}) . \tag{2.17}$$

Adopting the notation of Amari for α-connexion

$$\Gamma^{(\alpha)}_{ijk} = \Gamma^{(1)}_{ijk} + \frac{1-\alpha}{2} T_{ijk}$$

the expression (2.16) can be written

$$c_{ijk} = \lambda\mu[\Gamma^{(2\lambda-1)}_{ijk} + \Gamma^{(2\mu-1)}_{jki} + \Gamma^{(2\mu-1)}_{ikj}] . \tag{2.18}$$

When $\lambda = \mu = \frac{1}{2}$, (2.18) becomes

$$c_{ijk} = \frac{1}{4} [\Gamma^{(0)}_{ijk} + \Gamma^{(0)}_{jki} + \Gamma^{(0)}_{ikj}] . \tag{2.19}$$

Remark 1. In the definition of the Jensen difference (2.2), we used apriori probabilities λ and μ for the two probability distributions p and q which have some relevance in population studies. But in problems of statistical inference, a symmetric version may be used by taking $\lambda = \mu = \frac{1}{2}$.

3. THE QUADRATIC ENTROPY

The quadratic entropy was introduced in Rao (1982a) as a general measure of diversity of a probability distribution over any measurable space. It is defined as a function Q: $\underline{P} \to R_+$

$$Q(p) = \int_{\underline{X}\times\underline{X}} K(x,y)p(x)p(y)d\nu(x)d\nu(y) \tag{3.1}$$

where $K(x,y)$ is *symmetric, non-negative and conditionally negative definite,* i.e.,

$$\sum_1^n\sum_1^n K(x_i,x_j)a_ia_j \le 0$$

for any choice of $(x_1,\ldots,x_n)$ and of $(a_1,\ldots,a_n)$ such that $a_1+\ldots+a_n = 0$, with the further condition $K(x,y) = 0$ if $x = y$. It was shown in Rao (1982b, 1984) that the quadratic entropy is concave over $\underline{P}$ and its Jensen difference has nice convexity properties which makes it an ideal measure of diversity. In view of its usefulness in statistical applications, we give explicit expressions for the quadratic differential metric and the connection coefficients associated with the quadratic entropy, in the case of the parametric family $\underline{P}_\theta$.

From Theorem 2.1, the (i,j)-th element of the Q-information matrix is

$$g_{ij}^Q(\theta) = -\left.\frac{\partial^2 Q(\lambda p_\theta + \mu p_\phi)}{\partial\theta_i\partial\phi_j}\right|_{\phi=\theta}. \tag{3.2}$$

Observing that

$$Q(\lambda p_\theta + \mu p_\theta) = \int K(x,y)[\lambda p(x,\theta)+\mu p(x,\phi)][\lambda p(y,\theta)+\mu p(y,\phi)]d\nu(x)d\nu(y),$$

we find the explicit expression for (3.2) as

$$g_{ij}^{Q}(\theta) = -2\lambda\mu \int K(x,y)\, \frac{\partial p(x,\theta)}{\partial\theta_i}\, \frac{\partial p(y,\theta)}{\partial\theta_j}\, d\nu(x)\partial\nu(y) \tag{3.3}$$

$$= -2\ \lambda\mu\ E[K(x,y)\, \frac{\partial\ \log p(x,\theta)}{\partial\theta_i}\, \frac{\partial\ \log\ p(y,\theta)}{\partial\theta_j}] \ .$$

Using the expression (2.14), we find on carrying out the necessary computations

$$c_{ijk}^{Q} = -2\lambda\mu(\Gamma_{ijk} + \Gamma_{ikj} + \Gamma_{jki})$$

where

$$\Gamma_{ijk} = \int K(x,y)\, \frac{\partial p(x,\theta)}{\partial\theta_k}\, \frac{\partial^2 p(y,\theta)}{\partial\theta_i\partial\theta_j}\, d\nu(x)d\nu(y)\ . \tag{3.4}$$

It is of interest to note that the expressions (3.3) and (3.4) are invariant for transformations of both the parameters and variables.

For further properties of quadratic entropies, the reader is referred to Lau (1984) and Rao (1984).

4. METRICS BASED ON DIVERGENCE MEASURES

Burbea and Rao (1982a, 1982b), Burbea (1986) and Eguchi (1984) have considered metrics arising out of a variety of divergence measures between probability distributions. A typical divergence measure is of the form

$$D_F(p_\theta, p_\phi) = \int_{\underline{X}} F[p(x,\theta), p(x,\phi)]d\nu(x) \tag{4.1}$$

where F satisfies the following conditions:

(i) $F(\cdot,\cdot)$ is a C^3-function of $R_+ \times R_+$,

(ii) $F(x,\cdot)$ is strictly convex on R_+ for every $x \varepsilon R_+$,

(iii) $F(x,x) = 0$ for every $x \varepsilon R_+$,

(iv) $\frac{\partial F(x,y = x)}{\partial y} = 0$ for every $x \varepsilon R_+$.

Let us consider the expansion

$$D_F(p_\theta, p_{\theta+d\theta}) = \frac{1}{2!} \Sigma\Sigma g_{ij}^F(\theta)d\theta_i d\theta_j + \frac{1}{3!} c_{ijk}^F(\theta)d\theta_i d\theta_j d\theta_k + \cdots \tag{4.2}$$

and obtain explicit expressions for g_{ij}^F and c_{ijk}^F.

Theorem 4.1. Let

$$F_1(x,y) = \frac{\partial F(x,y)}{\partial x}, \quad F_2(x,y) = \frac{\partial F(x,y)}{\partial y}$$

$$F_{11} = \frac{\partial^2 F(x,y)}{\partial x^2}, \quad F_{12} = \frac{\partial^2 F(x,y)}{\partial x \partial y}, \quad F_{22} = \frac{\partial^2 F(x,y)}{\partial y^2}$$

$$F_{222} = \frac{\partial^3 F(x,y)}{\partial y^3}.$$

Then

$$\text{(i)} \quad g_{ij}^F(\theta) = \int F_{22}[p_\theta, p_\theta] \frac{\partial p_\theta}{\partial \theta_i} \frac{\partial p_\theta}{\partial \theta_j} d\nu(x)$$

$$= - \int F_{12}[p_\theta, p_\theta] \frac{\partial p_\theta}{\partial \theta_i} \frac{\partial p_\theta}{\partial \theta_j} d\nu(x) .$$

$$\text{ii)}\quad c^F_{ijk} = \int F_{222}[p_\theta, p_\theta] \frac{\partial p_\theta}{\partial \theta_i} \frac{\partial p_\theta}{\partial \theta_j} \frac{\partial p_\theta}{\partial \theta_k} d\nu(x)$$
$$+ \int F_{22}[p_\theta, p_\theta][\frac{\partial^2 p_\theta}{\partial\theta_i \partial\theta_j} \frac{\partial p_\theta}{\partial \theta_k} + \frac{\partial^2 p_\theta}{\partial\theta_i \partial\theta_k} \frac{\partial p_\theta}{\partial \theta_j} + \frac{\partial^2 p_\theta}{\partial\theta_j \partial\theta_k} \frac{\partial p_\theta}{\partial \theta_i}] d\nu(x) .$$

The results are established by straight forward computations.

Let us consider the directed divergence measure of Csiszár (1967), which plays an important role in problems of statistical inference,

$$D(p_\theta, p_\phi) = \int p(x,\theta)\, f(\frac{p(x,\phi)}{p(x,\theta)})\, d\nu(x) \tag{4.3}$$

where f is a convex function. In this case

$$g^f_{ij}(\theta) = \frac{\partial^2 D}{\partial\phi_i \partial\phi_j}\Big|_{\phi=\theta}$$
$$= f''(1) \int \frac{1}{p} \frac{\partial p}{\partial\theta_i} \frac{\partial p}{\partial\theta_j} d\nu(x) = f''(1) g_{ij}(\theta) \tag{4.4}$$

where g_{ij} are the elements of Fisher's information matrix. Thus a wide class of invariant divergence measures provide the same informative geometry on the parameter manifold. However, the c_{ijk} coefficients may depend on the particular convex function f chosen as shown below.

$$c^f_{ijk}(\theta) = \frac{\partial^3 D}{\partial\phi_i \partial\phi_j \partial\phi_k}\Big|_{\phi=\theta}$$
$$= f''(1)[\Gamma^{(1)}_{ijk} + \Gamma^{(1)}_{ikj} + \Gamma^{(1)}_{jki}] + (f'''(1) + 3f''(1))T_{ijk} \tag{4.5}$$

where $\Gamma^{(1)}_{ijk}$ and T_{ijk} are as defined in (2.17).

The results (4.4) and (4.5) have consequences in estimation theory, specially in the study of second order efficiency. While a large number of estimation procedures lead to first order efficient estimates (i.e., having the same asymptotic variance based on the elements of Fisher information matrix), they are distinguishable by different second order efficiencies of the derived estimators (see Rao, 1962).

If f is a convex function, then

$$f^*(u) = uf(\frac{1}{u})$$

is also convex, and the measure (4.3) associated with f+f* is

$$D^*(p_\theta,p_\phi) = \int [p_\theta f(\frac{p_\phi}{p_\theta}) + p_\phi f(\frac{p_\theta}{p_\phi})]d\nu(x) \tag{4.6}$$

which is symmetric in θ and ϕ. However, we may define (4.5) as a symmetric divergence measure without requiring f to be a convex function but satisfying the condition that $xf(x^{-1}) + f(x)$ is non-negative on R_+. In such a case

$$g_{ij}^{f+f^*}(\theta) = 2f''(1)g_{ij}(\theta)$$

$$c_{ijk}^{f+f^*}(\theta) = 2f''(1)[\Gamma_{ijk}^{(1)} + \Gamma_{ikj}^{(1)} + \Gamma_{jki}^{(1)}] + 3f'''(1)T_{ijk}$$

<u>Remarks on Sections 2, 3 and 4</u>. As pointed out by a referee, a unified treatment of the results in these three sections is possible by considering a general dissimilarity measure $D : p \times p \to \{0,\infty\}$ satisfying

(a) $D(p_\theta,p_\phi)$ is a c^3 function of θ,ϕ,

(b) $D(p,p) = 0$ for every $p \times p$.

Then putting

$$D_{i;jk} = \frac{\partial^3 D}{\partial\theta_i \partial\phi_j \partial\phi_k} \text{ etc.},$$

and differentiating $D_{;j}|_{\theta=\phi} = 0$ yields

$$[D_{i;j} + D_{;ij}]_{\theta=\phi} = 0,$$

$$[D_{ik;j} + D_{i:jk} + D_{k;ij} + D_{;ijk}]_{\theta=\phi} = 0$$

giving expressions for g_{ij}^D and c_{ijk}^D for a general D. However, the approach adopted in the paper enabled a discussion of the construction of the distance measures D through more basic functions like quadratic entropy, general entropy, cross entropy, and divergence between probability measures. The results expressed in terms of the basic functions are of some interest.

It is also possible to regard the dissimilarity measures of Section 3 and 4 as having the common form

$$D(p,q) = \int_{\underline{X} \times \underline{X}} F(p(x),q(x),p(y),q(y))d\nu(x,y)$$

where ν is a symmetric measure on $\underline{X} \times \underline{X}$. However, the expressions for g_{ij} and c_{ijk} are not simple.

5. OTHER DIVERGENCE MEASURES

In the last section, we considered the f-divergence measure which led to the Fisher information metric. A special case of this measure is the city block distance, or the overlap distance (see Rao, 1948, 1982a),

$$D_0(p_\theta, p_\phi) = \int |p(x,\theta)-p(x,\phi)| d\nu(x) \tag{5.1}$$

obtained by choosing $f(x) = 1-\min(x,1)$, which admits a direct interpretation in terms of errors of classification in discrimination problems. However, this is not a smooth function and no formula of the type (4.7) is available to determine the coefficients of the differential metric. But in some cases, it may turn out that

$$D_0(p_\theta, p_\phi) = D_0(\theta,\phi)$$

is a smooth function of θ and ϕ in which case

$$g_{ij} = \frac{\partial^2 D_0(\theta,\phi=\theta)}{\partial\phi_i \partial\phi_j} . \tag{5.2}$$

In the case when $p(x,\theta)$ is a p-variate normal density with mean $\underset{\sim}{\mu}$ and fixed variance covariance matrix $\underset{\sim}{\Sigma}$, the coefficient (5.2) can be easily computed to be proportional to σ^{ij}, the (i,j)-th element of Σ^{-1}, which is indeed the (i,j)-th element of the Fisher information matrix. The same result holds for any elliptical family, as then $D_0(\theta,\phi)$ is a function of the Mahalanobis distance between θ and ϕ (see Mitchell and Krzanowski, 1985).

Let $p(x,\theta)$ be the density of a uniform distribution in the interval $[0,\theta]$. Then it is seen that

$$D_0(\theta,\phi) = 2(1 - \frac{\theta}{\phi}) \quad \text{if } \theta \leq \phi$$
$$= 2(1 - \frac{\phi}{\theta}) \quad \text{if } \theta > \phi. \tag{5.3}$$

Although this is not a differentiable function, it is seen that

$$ds^2 = 4\,\frac{d\theta^2}{\theta^2}$$

is the metric associated with (5.3).

Another general divergence measure which has some practical applications is

$$D_\psi(p_\theta,p_\phi) = \int [\psi(p_\theta)-\psi(p_\phi)]^2 d\nu(x)$$

which is indeed a smooth function if ψ is so. In this case

$$g^\psi_{ij}(\theta) = 2\int [\psi'(p_\theta)]^2 \frac{\partial p_\theta}{\partial\theta_i}\frac{\partial p_\theta}{\partial\theta_j}\, d\nu(x)$$

$$c^\psi_{ijk}(\theta) = 6\int \psi'(p_\theta)\psi''(p_\theta)\frac{\partial p_\theta}{\partial\theta_i}\frac{\partial p_\theta}{\partial\theta_j}\frac{\partial p_\theta}{\partial\theta_k}\, d\nu(x)$$
$$+ 2\int [\psi'(p_\theta)]^2\Big(\frac{\partial^2 p_\theta}{\partial\theta_i\partial\theta_j}\frac{\partial p_\theta}{\partial\theta_k} + \frac{\partial^2 p_\theta}{\partial\theta_i\partial\theta_k}\frac{\partial p_\theta}{\partial\theta_j} + \frac{\partial^2 p_\theta}{\partial\theta_j\partial\theta_k}\frac{\partial p_\theta}{\partial\theta_i}\Big)\, d\nu(x).$$

Another measure of interest is the cross entropy introduced in Rao and Nayak (1985). If H is any entropy function, then the cross entropy of p_ϕ with respect to p_θ was defined as

$$D(p_\theta,p_\phi) = H(p_\phi) - H(p_\theta) - \lim_{\lambda\downarrow 0}\frac{H[p_\phi+\lambda(p_\theta-p_\phi)]-H(p_\phi)}{\lambda}\,. \tag{5.4}$$

Let

$$H(p) = -\int h(p)d\nu(x)$$

as chosen in (2.10). Then (5.4) reduces to

$$D(p_\theta,p_\phi) = -\int h(p_\phi)d\nu(x) - \int h'(p_\phi)(p_\theta-p_\phi)d\nu(x) + \int h(p_\theta)d\nu(x)\,.$$

Then

$$g^h_{ij} = \int h''(p_\theta)\frac{\partial p_\theta}{\partial\theta_j}\frac{\partial p_\theta}{\partial\theta_j}\, d\nu(x)$$

which is the same as the h-entropy information matrix derived in (2.10), apart from a constant. Similarly

$$c^h_{ijk} = \Gamma^{(1)}_{ijk} + \Gamma^{(1)}_{ikj} + \Gamma^{(1)}_{jki} + T_{ijk}$$

where

$$\Gamma^{(1)}_{ijk} = E\{p_\theta h''(p_\theta) \frac{\partial^2 \log p_\theta}{\partial\theta_i \partial\theta_j} \frac{\partial \log p_\theta}{\partial\theta_k}\}$$

$$T_{ijk} = E\{[3p_\theta h''(p_\theta) + 2p_\theta^2 h'''(p_\theta)] \frac{\partial \log p_\theta}{\partial\theta_i} \frac{\partial \log p_\theta}{\partial\theta_j} \frac{\partial \log p_\theta}{\partial\theta_k}\}$$

6. GEODESIC DISTANCES

In Rao (1945) it was suggested that the information metric could be used to obtain the geodesic distances between probability distributions. Given any quadratic differential metric

$$ds^2 = \Sigma\Sigma\, g_{ij}(\theta) d\theta_i d\theta_j \tag{6.1}$$

where the matrix (g_{ij}) is positive definite, the geodesic curve $\theta = \theta(t)$ can in principle be determined from the Euler-Lagrange equations

$$\sum_1^n g_{ij}\, \ddot{\theta}_i + \sum_1^n\sum_1^n \Gamma_{ijk}\dot{\theta}_i\dot{\theta}_j = 0,\ k = 1,\ldots,n \tag{6.2}$$

and from the boundary conditions

$$\theta(t_1) = \theta, \quad \theta(t_2) = \phi \ .$$

In (6.2), the quantity

$$\Gamma_{ijk} = \frac{1}{2}\left[\frac{\partial}{\partial\theta_i} g_{jk} + \frac{\partial}{\partial\theta_j} g_{ki} - \frac{\partial}{\partial\theta_k} g_{ij}\right] \tag{6.3}$$

and is known as the "Christoffel symbol of the first kind."

By definition of the geodesic curve $\theta = \theta(t)$, its tangent vector $\dot{\theta} = \dot{\theta}(t)$ is of constant length with respect to the metric ds^2. Thus

$$\sum_1^n\sum_1^n g_{ij}\dot{\theta}_i\dot{\theta}_j = \text{constant} \ . \tag{6.4}$$

The constant may be chosen to be of value 1 when the curve parameter t is the arc length parameter s, $0 \leq s \leq s_0$, with $\theta(0) = \theta$, $\theta(s_0) = \phi$ and $s_0 = g(\theta,\phi)$ is the geodesic distance between θ and ϕ.

Aitkinson and Mitchell (1981) describe two other methods of deriving geodesic distances starting from a given differential metric. The distances obtained by these authors in various cases are given below. In each case we give the probability function $p(x,\theta)$ and the associated geodesic distance of (θ,ϕ) based on the Fisher information metric.

(1) Poisson distribution

$p(x,\theta) = e^{-\theta}\,\theta^{x}/x!$, $x = 0,1,\ldots,$ $\theta>0$

$g(\theta,\phi) = 2|\sqrt{\theta} - \sqrt{\phi}\,|$

(2) Binomial distribution (n fixed)

$p(x,\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}$, $x = 0,1,\ldots,n$, $0<\theta<1$

$$g(\theta,\phi) = 2\sqrt{n}|\sin^{-1}\sqrt{\theta} - \sin^{-1}\sqrt{\phi}|$$
$$= 2\sqrt{n}\cos^{-1}[\sqrt{\theta\phi} + \sqrt{(1-\theta)(1-\phi)}\,].$$

(3) Exponential distribution

$p(x,\theta) = \theta e^{-x\theta}$, $x \geq 0$

$g(\theta,\phi) = |\log\theta - \log\phi|$.

(4) Gamma distribution (n fixed)

$p(x,\theta) = \theta^{n}[\Gamma(n)]^{-1}x^{n-1}e^{-x\theta}$, $x \geq 0$

$g(\theta,\phi) = \sqrt{n}\,|\log\theta - \log\phi|$

(5) Normal distribution (fixed variance)

$p(x,\mu,\sigma_0^2) = N(\mu,\sigma_0^2;x)$, σ_0 fixed

$g(\mu_1,\mu_2) = |\mu_1 - \mu_2|/\sigma_0$

(6) Normal distribution (fixed mean)

$p(x,\mu_0,\sigma^2) = N(\mu_0,\sigma^2;x)$, μ_0 fixed

$g(\sigma_1^2,\sigma_2^2) = \sqrt{2}\,|\log\sigma_1 - \log\sigma_2|$

(7) Normal distribution

$p(x,\mu;\sigma^2) = N(\mu,\sigma^2;x)$, μ and σ both variable.

The information metric in this case is

$$ds^2 = \frac{d\mu^2}{\sigma^2} + \frac{2d\sigma^2}{\sigma^2} \tag{6.5}$$

and the geodesic distance is

$$g(\mu_1,\sigma_1;\mu_2,\sigma_2) = \sqrt{2}\left|\log\frac{1+\delta(1,2)}{1-\delta(1,2)}\right|$$
$$= 2\sqrt{2}\tanh^{-1}\delta(1,2) \tag{6.6}$$

where $\sigma(1,2)$ is the positive square root of

$$\frac{(\mu_1-\mu_2)^2+2(\sigma_1-\sigma_2)^2}{(\mu_1-\mu_2)^2+2(\sigma_1+\sigma_2)^2}\,.$$

The explicit form (6.6) is given in Burbea and Rao (1982a). From (6.6)

$$g(\mu,\sigma_1^2;\mu,\sigma_2^2) = \sqrt{2}|\log\sigma_1 - \log\sigma_2|$$

which agrees with result (6). However, $g(\mu_1,\sigma^2;\mu_2,\sigma^2)$ does not reduce to result (7) since σ = constant is not a geodesic curve with respect to the metric (6.5)

(8) Multivariate normal distribution

$N_p(\mu,\Sigma;x)$, Σ fixed

$$g(\mu_1,\mu_2) = [(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)]^{\frac{1}{2}}$$

which is Mahalanobis distance.

(9) Multivariate normal distribution

$N(\mu,\Sigma;x)$, μ fixed

$$g(\Sigma_1,\Sigma_2) = [2^{-1}\sum_1^p(\log\lambda_i)^2]^{\frac{1}{2}}$$

where $0 < \lambda_1 \leq \ldots \leq \lambda_p$ are the roots of the determinantal equation $|\Sigma_2-\lambda\Sigma_1| = 0$. The above explicit form is due to S. T. Jensen as mentioned in Atkinson and Mitchell (1981).

(10) Negative binomial distribution

$$p(x,\theta) = [x!\Gamma(r)]^{-1}\Gamma(x+r)\theta^x(1-\theta)^r, \text{ r fixed}$$

$$g(\theta,\phi) = 2\sqrt{r}\cosh^{-1}\frac{1-\sqrt{\theta\phi}}{\sqrt{(1-\theta)(1-\phi)}}$$

$$= 2\sqrt{r}\log\frac{1-\sqrt{\theta\phi}+|\sqrt{\theta}-\sqrt{\phi}|}{\sqrt{(1-\theta)(1-\phi)}}$$

This computation is due to Oller and Cuadras (1985).

(11) Multinomial distribution

$$p(n_1,\ldots,n_k;\pi_1,\ldots,\pi_k) = \frac{n!}{n_1!\ldots n_k!}\pi_1^{n_1}\ldots\pi_k^{n_k}\text{ , n fixed.}$$

Let $\underset{\sim}{\pi}_1 = (\pi_{11},\ldots,\pi_{k1})$ and $\underset{\sim}{\pi}_2 = (\pi_{12},\ldots,\pi_{k2})$. Then

$$g(\underset{\sim}{\pi}_1,\underset{\sim}{\pi}_2) = 2\sqrt{n}\cos^{-1}(\sum_1^k\sqrt{\pi_{i1}\pi_{i2}})$$

The above computation was originally done by Rao (1945), but an easier method of derivation is given by Atkinson and Mitchell (1981).

Recently Burbea (1984) obtained geodesic distances in the case of independent Poisson and Normal distributions which are given below. These results (12) and (13) follow directly from (1) and (7) respectively as the squared geodesic distances behave additively under combination of independent distributions.

(12) Independent Poisson distributions

$$p(x_1,\ldots,x_n;\theta_1,\ldots,\theta_n) = \prod_1^n e^{-\theta_i} \frac{\theta_i^{x_i}}{x_i!}$$

$$g(\theta_1,\ldots,\theta_n;\phi_1,\ldots,\phi_n) = 2[\sum_1^n(\sqrt{\theta_i} - \sqrt{\phi_i})^2]^{1/2}$$

(13) Independent Normal distributions

$$N(x;\mu_1,\sigma_1^2)\ldots N(x_n;\mu_n,\sigma_n^2)$$

$$g[(\mu_{11}\sigma_{11}^2),\ldots,(\mu_{n1},\sigma_{n1}^2);(\mu_{12},\sigma_{12}^2),\ldots,(\mu_{n2},\sigma_{n2}^2)]$$

$$= \sqrt{2}\ [\sum_{k=1}^n \log^2 \frac{1+\delta_k(1,2)}{1-\delta_k(1,2)}]^{1/2}$$

where $\delta_k(1,2)$ is the positive square root of

$$\frac{(\mu_{k1}-\mu_{k2})^2+2(\sigma_{k1}-\sigma_{k2})^2}{(\mu_{k1}-\mu_{k2})^2+2(\sigma_{k1}+\sigma_{k2})^2}\ .$$

(14) Multivariate elliptic distributions

$$p(x|\underset{\sim}{\mu},\Sigma) = |\Sigma|^{-1/2}h[(x-\mu)'\Sigma^{-1}(x-\mu)],$$

for some function h, and Σ is fixed

$$g(\underset{\sim}{\mu}_1,\underset{\sim}{\mu}_2) = c_h(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)$$

where c_h is a constant, which is essentially Mahalanobis distance. This result is due to Mitchell and Krzanowski (1985).

The use of the c_{ijk} coefficients defined in (2.4) and (4.2) in the discussion of statistical problems will be considered in a future communication.

REFERENCES

Amari, S. I. (1982). Differential geometry of curved exponential families - curvature and information loss. Ann. Stat. 10, 357-385.

Amari, S. I. (1983). A foundation of information geometry. Electronics and Communications in Japan 66-A, 1-10.

Atkinson, C. and Mitchell, A. F. S. (1981). Rao's distance measure. Sankhyā 43, 345-365.

Burbea, J. (1986). Informative geometry in probability spaces. Expo. Math. 4, 347-378.

Burbea, J. and Rao, C. Radhakrishna (1982a). Entropy differential metric, distance and divergence measures in probability spaces: a unified approach. J. Multivariate Anal. 12, 575-596.

Burbea, J. and Rao, C. Radhakrishna (1982b). Differential metrics in probability spaces. Probability Math. Statist. 3, 115-132.

Čencov, N. N. (1982). Statistical decision rules and optimal inference. Transactions of Mathematical Monographs 53, Amer. Math. Soc., Providence.

Csiszár, I. (1967). Information-type measures of difference of probability distributions and indirect observations. Studia Scientiarum Mathematicarum Hungrica 2, 299-318.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency, with discussion). Ann. Statist. 3, 1189-1217.

Efron, B. (1982). Maximum likelihood decision theory. Ann. Statist. 10, 340-356.

Eguchi, S. (1983). Second order efficiency of minimum contrast estimators in a curved exponential family. Ann. Statist. 11, 793-803.

Eguchi, S. (1984). A differential geometric approach to statistical inference on the basis of contrast functionals. Tech. Report No. 136, Hiroshima University, Hiroshima, Japan.

Havrda, M. E. and Charvat, F. (1967). Quantification method of classification processes: Concept of α-entropy. Kybernetika 3, 30-35.

Kass, R. E. (1980). The Riemannian structure of model spaces: a geometrical approach to inference. Ph.D. thesis, University of Chicago.

Kass, R. E. (1981). The geometry of asymptotic inference. Tech. Rept. 215. Dept. of Statistics, Carnegie-Mellon University.

Lau, Ka-Sing (1985). Characterization of Rao's quadratic entropy. Sankhyā A 47, 295-309.

Mitchell, A. F. S. and Krzanowski, W. J. (1985). The Mahalanobis distance and elliptic distributions. (To appear in Biometrika).

Nei, M. (1978). The theory of genetic distance and evolution of human races. Japan J. Human Genet. 23, 341-369.

Oller, J. M. and Cuadras, C. M. (1985). Rao's distance for negative multinomial distributions. Sankhyā 47, 75-83.

Rao, C. Radhakrishna (1945). Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. 37, 81-91.

Rao, C. Radhakrishna (1948). The utilization of multiple measurements in problems of biological classification (with discussion). J. Roy. Statist. Soc. B10, 159-203.

Rao, C. Radhakrishna (1949). On the distance between two populations. Sankhyā 9, 246-248.

Rao, C. Radhakrishna (1954). On the use and interpretation of distance functions in statistics. Bull. Inst. Inter. Statist. 34, 90-100.

Rao, C. Radhakrishna (1962). Efficient estimates and optimum inference procedures in large samples (with discussion). J. Roy. Statist. Soc. B 24, 46-72.

Rao, C. Radhakrishna (1973). Linear Statistical Inference and its Applications. (Second edition) Wiley, New York.

Rao, C. Radhakrishna (1982a). Diversity and dissimilarity coefficients: a unified approach. J. Theoret. Pop. Biology 21, 24-43.

Rao, C. Radhakrishna (1982b). Diversity: its measurement, decomposition, apportionment and analysis. Sankhyā A 44, 1-22.

Rao, C. Radhakrishna (1984). Convexity properties of entropy functions and analysis of diversity. In Inequalities in Statistics and Probability, IRS Lecture Notes, Vol. 5, 68-77.

Rao, C. Radhakrishna and Nayak, T. K. (1985). Cross entropy, dissimilarity measures and characterizations of quadratic entropy. IEEE Trans. Information Theory IT 31, 589-593.

Shahshahani, S. (1979). A new mathematical framework for the study of linkage and selection. Memoirs of the American Mathematical Society, No. 211.