HIROSHIMA MATH. J.
15 (1985), 341–391

A differential geometric approach to statistical inference on the basis of contrast functionals

Shinto EGUCHI

(Received September 20, 1984)

Summary

In this paper, we consider contrast functionals on the space of all probability measures equivalent to each other. Many examples of the contrast functional have been proposed and estimation methods based on them, called the minimum contrast estimation methods, have been investigated since the theory of estimation was initiated by R. A. Fisher. It is shown that a contrast functional generates a conjugate metric structure with a Riemannian metric and a conjugate pair of affine connections on the space. We show that this structure explains some properties of the minimum contrast estimator. In particular, the explicit formulas for the limiting information loss for the estimators are given in covariance structure models. Moreover we propose a generalized scoring method for seeking the minimum contrast estimates. It is shown that the convergence of the algorithm is affected by two geometric quantities which can be expressed in the conjugate metric structure.

0. Introduction

The concepts of information, entropy, energy, diversity, discrepancy, and divergence for random phenomena occupy a fundamental position in various fields of mathematical sciences, e.g., statistical mechanics, information theory, system and control theory, evolutional biology and statistics (see Boltzmann [14], Fisher [28], Shannon [54], Wiener [62], Kullback [39] and Simpson [56]). Although various terminologies for the concepts are used, the concepts have a common aspect which can be used for the comparison of random phenomena. In this sense, as a measure of the concepts we consider *a cortrast functional* defined on a pair of probability measures, which is positive except in the case of agreement between the two probability measures. R. A. Fisher [28] was the first to introduce a measure of information, called the Fisher information, into the theory of estimation. Since his work, statisticians have proposed various contrast functionals and investigated inference based on them (see, e.g. Mahalanobis [43], Battarcharya [13], Jeffreys [35], Haldane [30], Chernoff [18] Matusita [44], Reyni [53], Kagan [36], Csiszar [19] and Burbea and Rao [16]).

Closely related to the work mentioned above are geometric approaches to statistics. For example, consider a linear regression model

$$y = X\beta + \varepsilon$$

with a design matrix X of size $n \times m$ and a parameter β, i.e., the n-variate data y are observed from a point on the m-dimensional plane

$$\mu(X) = \{X\beta : \beta \in R^m\}$$

in R^n with a random term ε. The geometric properties of $\mu(X)$ and the stochastic properties of ε together can be used to determine the optimal form for statistical inference. Suppose that ε follows the Gaussian distribution with mean $\mathbf{0}$ and known covariance Σ_0. Then with respect to an inner product $(y_1, y_2) = y_1' \Sigma_0^{-1} y_2$ in R^n, the orthogonally projective estimator

$$\hat{\beta}(y) = (X' \Sigma_0^{-1} X)^{-1} X' \Sigma_0^{-1} y,$$

called Gauss–Markov estimator, is optimal in the sense of minimizing the variace in the class of all unbiased estimators. There is *a dual structure* between the linear regression model and the Gauss–Markov estimator in the sense that the data space R^n can be represented as

$$\cup_{\mu \in \mu(X)} \hat{\beta}^{-1}(\mu)$$

by the pair of the model and the estimator, where $\hat{\beta}^{-1}(\mu) = \{y : X\hat{\beta}(y) = \mu\}$. However, if the data follows a more general stochastic mechanism, then such a geometric structure cannot be explained in terms of Euclidean geometry associated with the linear regression model. To overcome this difficulty, we need a Riemannian geometry.

Rao [47] was the first to point out that a parametric family of probability measures or a statistical model is a Riemannian space equipped with the Fisher information as its metric. Subsequently, the metric associated with this Riemannian space will be called the information metric. Since Rao's work appeared, statisticians have investigated geometric aspects of statistical models (see Yoshizawa [63], Chentsov [17], Atkinson and Mitchell [7] and Skovgaard [57]).

Especially important work in this field was done by Efron [21], who elucidated a dual structure between a one-parameter curved exponential family and the maximum likelihood estimator. In discussion of Efron's paper, Dawid [20] noted two affine connections, called the mixture and exponential connections. These connections are apparently independent of the information metric in statistical models. Amari [2] showed that the two affine connections are conjugate in the sense that the information metric between parallel shifts with respect to these connections is invariant. He also elucidated in the differential geometric framework that two curvatures introduced by Efron are the second fundemental forms with respect to the respective connections. This completes the theory

of second order efficiency for the maximum likelihood estimator by way of the contributions of Fisher [28], [49], [50], Ghosh and Subramanyam [29] and Efron [21]. Thus the conjugacy between the mixture and exponential connections plays an important role in statistics. We call the triple of the information metric, the mixture and the exponential connections *the statistical conjugate metric structure.* For recent developments in differential geometric statistics, see Amari [3] [4], Barndorff–Nielsen and Blæsid [9], Amari and Kumon [5], Kumon and Amari [40], [41], Nagaoka and Amari [45], Eguchi [23], [24], [25], Wei and Tsai [61], Kass [37], Lauritzen [42] and Taneichi, Sato and Kawaguchi [59].

The purpose of this paper is to investigate a geometry generated by a contrast functional on the space of all probability measures equivalent to each other and its applications to statistical inference. The paper is divided into two parts.

Part I contains a differential geometric approach based on a contrast functional. Our discussion is restricted to a finite parametric family of the full space, or a statistical model, except for Section 4. Section 1 is devoted to prepare differential geometric tools employed in Part I. In section 2, we introduce three types of contrast functionals, called *W-type*, *M-type* and *S-type*. In particular, a systematic construction of contrast functionals of W-type based on operations which lead to a relation among classical contrast functionals is given. A notion of *scale invariance* is introduced and its implication is considered in the measure theoretic light. It is shown that scale invariance is obtained only for the class of Chernoff informations. In Section 3, we show that a contrast functional generates *a conjugate metric structure* on the statistical models. This is a triple which consists of a Riemannian metric and a conjugate pair of affine connections. In general, the structure is irrelevant to the statistical conjugate metric structure. However in the case of contrast functionals of W-type, the generated conjugate metric structures are closely related to the statistical conjugate metric structure. Section 4 leads to an extension of the conjugate metric structures on finite parametric families to the full space. Nagaoka and Amari [45] present the α-geodesic curve connecting two probability measures. This extension is defined in terms of the Gâteau differential along the α-geodesic curve. Thus we obtain the *α-representation* of the conjugate metric structure on the full space. It is also shown that the above properties on the contrast functionals of W-type still hold on the full space.

Part II deals with statistical inference based on a contrast functional from the view point of the differential geometric approaches. In Section 5, first we give a general survey of estimation theory in terms of the notion of a summary introduced by Efron [22]. The set of all estimators can be classified into Fisher-consistent, first order efficient and second order efficient classes on the basis of

the measure of limiting information loss. It follows from the result of Eguchi [23] that the conjugate metric structure generated by a contrast functional determines which class the minimum contrast estimator belongs to. We present a construction of second order efficient estimators by using the operations developed in Section 2. *A classification* in the class of all contrast functionals is defined on the basis of their conjugate metric structurs. Finally the equivalence between two classifications of estimators and contrast functionals is shown. Section 6 deals with covariance structure models which are defined by specifying the covariance matrices of Gaussian measures (see Browne [15]). In this model, we investigate the class of estimators proposed by Swain [58] in comparison with other minimum contrast estimators. Explicit formulas for the limiting information loss for the estimators are represented in terms of the trace and the Kronecker product on square matrices. Some examples, intracorrelation model and linear covariance structure model are given along with numerical results. Finally in Section 7, we consider an algorithm for seeking a minimum contrast estimate. *A generalized scoring method* is proposed. The convergence of this algorithm is elucidated by the generated conjugate metric structure.

Part I. Geometry of a space of probability measures on the basis of contrast functionals

1. Differential geometric framework of statistical model

Let μ be a σ-finite measure on a measure space $(\mathcal{X}, \mathcal{B})$ with a σ-algebra $\mathcal{B}$ of subsets of a sample space $\mathcal{X}$. A space $\mathcal{P}_0$ denotes the class of all probability measures equivalent to μ. The density form of $\mathcal{P}_0$ is written as

$$\mathcal{F}_0 = \left\{ f = \frac{dP}{d\mu} : P \in \mathcal{P}_0 \right\}$$

with respect to μ. We often have our interests in a finite parametric family, or a model

$$\mathcal{F} = \{ f_\theta \in \mathcal{F}_0 : \theta \in \Theta \} ,$$

where Θ is an open subset of the n-dimensional Euclidean space $\boldsymbol{R}^n$. The dimension of $\mathcal{F}$ is defined by the dimension of Θ. A typical model is given in the following example.

Example 1.1. A family of all Gaussian distributions on $\boldsymbol{R}^k$ is parametrized as

$$\mathcal{G} = \{ f_\theta(x) = (2\pi)^{-\frac{k}{2}} (\det \Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)' \Sigma^{-1}(x-\mu)} : \theta \in \Theta \}$$

with respect to the k-dimensional Lebesque measure. Here the parameter θ

consists of elements of the mean vector μ and the convariance matrix Σ. The dimension of $\mathcal{G}$ is $\{k+k(k+1)/2\}$. We shall investigate special structures of $\mathcal{G}$ in a subsequent section.

A parametric family $\mathcal{F}$ is said to be regular if the following conditions are satisfied:

A–1. *The mapping $\theta \mapsto f_\theta$ is continuous with the topology generated by the Hellinger distance, which will be defined in* (2.1) *of Section* 2.

A–2. *The function f_θ is C^3-differentiable in θ under the integral sign.*

The condition **A**–2 implies that

$$\int \partial_i f_\theta d\mu = \int \partial_j \partial_i f_\theta d\mu = \int \partial_k \partial_j \partial_i f_\theta d\mu = 0$$

for every θ where $\theta=(\theta^1, \theta^2,\ldots, \theta^n)$ and $\partial_i = \partial/\partial\theta^i$.

Let ϕ be a C^3-diffeomorphism of θ into τ. The family $\mathcal{F}$ is also written as

$$\{\bar{f}_\tau(x): \tau \in T\}$$

in terms of τ, where $T=\phi(\Theta)$ and

$$\bar{f}_\tau(x) = f_{\phi^{-1}(\tau)}(x) .$$

Note that the conditions **A**–1 and **A**–2 are independent of a choice of parameter. Henceforth we treat regular parametric families.

Amai [2] formulated a differential geometric framework for the theory of statistical inference. He focussed on the properties of the model which are invariant under one-to-one transformations on both the sample space and the parameter space. Our results are based on Amari's framework. In the rest of this section we make a concise review, see Amari [2], [3] for detailed discussion.

Let $T_f(\mathcal{F})$ be the tangent space of $\mathcal{F}$ at f. A set of functions on $\mathcal{X}$:

$$e_i = e_i(\theta) = \partial_i \log f_\theta$$

with $i=1, 2,\ldots, n$ is defined as a basis of $T_{f_\theta}(\mathcal{F})$ with respect to θ-coordinates. The tangent space $T_{f_\theta}(\mathcal{F})$ is seen to be a linear subspace of

$$\{s: s \text{ is a } \mathcal{B}\text{-measureable function on } \mathcal{X} \text{ with } \int s f_\theta d\mu = 0\}$$

because of **A**–2. Let t be a one-to-one transformation on the sample space $\mathcal{X}$. A space $\mathcal{F}^t$ denotes the class of the densities induced from the densities in $\mathcal{F}$ by means of t. Then it is noted that the natural basis of $\mathcal{F}^t$ is still equal to that of $\mathcal{F}$. Thus the choice of the natural basis is invariant under one-to-one transformations on $\mathcal{X}$.

A Riemannian metric tensor g on $\mathcal{F}$ is introduced as the components

$$g_{ij}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{e_i(\boldsymbol{\theta})e_j(\boldsymbol{\theta})\}$$

with respect to the parameter $\boldsymbol{\theta}$(cf. Rao [47]), which will be called *the information metric.* Furthermore a pair of affine connections $\overset{m}{\Gamma}$ and $\overset{e}{\Gamma}$ is defined to have their coefficients

$$\overset{m}{\Gamma}_{ij,k}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{\partial_i e_j(\boldsymbol{\theta})e_k(\boldsymbol{\theta})\} + E_{\boldsymbol{\theta}}\{e_i(\boldsymbol{\theta})e_j(\boldsymbol{\theta})e_k(\boldsymbol{\theta})\}$$

and

$$\overset{e}{\Gamma}_{ij,k}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{\partial_i e_j(\boldsymbol{\theta})e_k(\boldsymbol{\theta})\},$$

respectively. Note that g, $\overset{m}{\Gamma}$ and $\overset{e}{\Gamma}$ are invariant under transformations on the sample space. It is natural for us to require that statistical inference based on a sample should be invariant under one-to-one transformations of the sample. The invariance of the geometric quantities g, $\overset{m}{\Gamma}$ and $\overset{e}{\Gamma}$ will play an important role in investigating the structure of statistical inference.

Example 1.2. We raise two typical parametric models. One is a mixture family:

$$\mathcal{F}_m = \{f_\eta(\boldsymbol{x})\eta^1 + \cdots + f_n\eta^n + f_0(1 - \textstyle\sum_{i=1}^n \eta_i) : \boldsymbol{\eta} \in H\}$$

with f_j in $\mathcal{F}_0$ for $j=0, 1,\ldots, n$ and

$$H = \{\boldsymbol{\eta}=(\eta_1,\ldots,\eta_n): \textstyle\sum_{i=1}^n \eta_i < 1,\ \eta_i > 0\ (i=1,\ldots,n)\}.$$

The other is an exponential family:

$$\mathcal{F}_e = \{f_{\boldsymbol{\beta}}(\boldsymbol{x}) = e^{\boldsymbol{x}'\boldsymbol{\beta}-\psi(\boldsymbol{\beta})} : \boldsymbol{\beta} \in B\},$$

where

$$B = \{\boldsymbol{\beta} \in \boldsymbol{R}^n : \int e^{\boldsymbol{x}'\boldsymbol{\beta}} d\mu(\boldsymbol{x}) < \infty\}.$$

Then the coordinate systems $\boldsymbol{\eta}$ of $\mathcal{F}_m$ and $\boldsymbol{\beta}$ of $\mathcal{F}_e$ become the affine parameters of $\overset{m}{\Gamma}$ and $\overset{e}{\Gamma}$, respectively. In other words, the coefficients of $\overset{m}{\Gamma}$ identically vanish with respect to $\boldsymbol{\eta}$ and those of $\overset{e}{\Gamma}$ do with respect to $\boldsymbol{\beta}$.

In the sense of Example 1.2, $\overset{m}{\Gamma}$ and $\overset{e}{\Gamma}$ are called *the mixture connection* and *the exponential connection*, respectively. Here recall the metric connection $\overset{0}{\Gamma}$, or the Levi-Civita connection with respect to g given by the coefficients

$$\overset{0}{\Gamma}_{ij,k} = \frac{1}{2}(\partial_i g_{jk} + \partial_j g_{ki} - \partial_k g_{ij}).$$

The pair of connections $\overset{m}{\Gamma}$ and $\overset{e}{\Gamma}$ has a relation:

$$\overset{0}{\Gamma} = \frac{1}{2}(\overset{m}{\Gamma} + \overset{e}{\Gamma}).$$

We now call a triple

$$\mathscr{C}_s = (g, \overset{m}{\Gamma}, \overset{e}{\Gamma})$$

the statistical conjugate metric structure, of which terminolgy will be justified in Section 3. The α-version of $\mathscr{C}_s$ is defined as

$$\mathscr{C}_\alpha = (g, \overset{-\alpha}{\Gamma}, \overset{\alpha}{\Gamma})$$

with the α-connection

$$\overset{\alpha}{\Gamma} = \frac{1-\alpha}{2}\overset{m}{\Gamma} + \frac{1+\alpha}{2}\overset{e}{\Gamma}.$$

We call $\mathscr{C}_\alpha$ *the α-conjugate metric structure* on $\mathscr{F}$.

We consider the case that the model $\mathscr{F}$ is embedded in an exponential family $\mathscr{F}_e$. Let n and m be the dimensions of $\mathscr{F}_e$ and $\mathscr{F}$, respectively. Suppose that a parametrization of $\mathscr{F}$ is given as follows:

$$\{f_\theta(\boldsymbol{x}) = e^{\boldsymbol{x}'\boldsymbol{\beta}(\boldsymbol{\theta}) - \psi[\boldsymbol{\beta}(\boldsymbol{\theta})]} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$$

with the m-component parameter $\boldsymbol{\theta}$. We note that if

$$\boldsymbol{\beta}(\boldsymbol{\theta}) = A\boldsymbol{\theta} + \boldsymbol{\beta}_0$$

with an $n \times m$ matrix A of full rank, then $\mathscr{F}$ returns to an m-dimensional exponential family with the form

$$f_\theta(x) = e^{t'\theta - \phi(\theta)},$$

with respect to $\tilde{\mu}$, where $\boldsymbol{t} = A'\boldsymbol{x}$, $\phi(\boldsymbol{\theta}) = \psi(A\boldsymbol{\theta} + \boldsymbol{\beta}_0)$ and $\tilde{\mu}$ is defined as

$$\tilde{\mu}(B) = \int_B e^{\boldsymbol{\beta}_0'\boldsymbol{x}} d\mu(\boldsymbol{x}).$$

Thus any exponential family has reproductivity with respect to a flat embedding. Hence we call $\mathscr{F}$ *an (n, m)-curved exponential family* if the image $\boldsymbol{\beta}(\Theta)$ is non-flat in the parameter space B of $\mathscr{F}_e$ and the convex hull of $\boldsymbol{\beta}(\Theta)$ includes an open subset of B. The structure $\mathscr{C}_s$ on $\mathscr{F}_e$ is easily given as follows: The information metric g with $g_{ij} = \partial_i \partial_j \psi$, the mixture connection $\overset{m}{\Gamma}$ with $\overset{m}{\Gamma}_{ij,k} = \partial_i \partial_j \partial_k \psi$ and the

exponential connection $\overset{e}{\Gamma}$ with $\overset{e}{\Gamma}_{ij,k}=0$. We introduce the form of $\mathscr{C}_s$ induced to $\mathscr{F}$ by way of the orthogonal decomposition

$$T_f(\mathscr{F}_e) = T_f(\mathscr{F}) \oplus T_f^{\perp}(\mathscr{F})$$

with respect to the metric g for each f in $\mathscr{F}$. The connecting tensor B form $T_f(\mathscr{F}_e)$ to $T_f(\mathscr{F})$ has the components

$$B_a^i = \frac{\partial \beta^i(\boldsymbol{\theta})}{\partial \theta^a} \qquad (a=1, 2,\ldots, m,\ i=1, 2,\ldots, n),$$

while the tensor $B^{\perp}$ connecting $T_f(\mathscr{F}_e)$ with $T_f^{\perp}(\mathscr{F})$ is defined as the components B_λ^i with $\lambda=m+1,\ldots, n$ and $i=1, 2,\ldots, n$ satisfying

$$B_\lambda^i g_{ij} B_a^j = 0$$

for each $a=1, 2,\ldots, m$. The metric g on $\mathscr{F}_e$ is decomposed into

$$\tilde{g}_{ab} = B_a^i g_{ij} B_b^j \tag{1.1}$$

and

$$\tilde{g}_{\lambda\mu} = B_\lambda^i g_{ij} B_\mu^j, \tag{1.2}$$

i.e., the matrix form of

$$\tilde{G} = \begin{bmatrix} B'GB & O \\ O & B^{\perp\prime}GB^{\perp} \end{bmatrix}.$$

Similarly the connections $\overset{m}{\Gamma}$ and $\overset{e}{\Gamma}$ on $\mathscr{F}_e$ are induced to $\mathscr{F}$. The embedding curvature tensor H_Γ: $T(\mathscr{F})\times T(\mathscr{F})\times T^{\perp}(\mathscr{F})\to \boldsymbol{R}$, or the second fundamental form of $\mathscr{F}$ is given by components

$$(H_\Gamma)_{ab\lambda} = \partial_a B_b^i B_\lambda^j g_{ij} + \Gamma_{ij,k} B_a^i B_b^j B_\lambda^k \tag{1.3}$$

with respect to an affine connection Γ. In particular we write

$$\overset{e}{H} = H_{\overset{e}{\Gamma}}$$

and

$$\overset{m}{H} = H_{\overset{m}{\Gamma}}.$$

In Section 4 we shall introduce a structure of estimation methods on the basis of the second fundamental forms $\overset{e}{H}$ and $\overset{m}{H}$.

2. Contrast functionals on the space of probability densities

In this section we introduce various contrast functionals over the space $\mathscr{F}_0$ of all probability densities with a common support. The product divergence with (α, β)-index is presented, which will lead to relations among classical contrast functionals, *e.g.* the squared Hellinger distance, the Jeffereys divergence, the Chernoff information of order α and the Kagan divergence. We consider some operations on the space of all convex functions, which give systematic construction of contrast functionals.

We call $\rho: \mathscr{F}_0 \times \mathscr{F}_0 \to \boldsymbol{R}$ a contrast functional if $\rho(f, g) \geqq 0$ for all f and g in $\mathscr{F}_0$ with equality if and only if $f=g$. Note that the functional ρ is not assumed to be symmetric. We shall assert through subsequent discussion that non-symmetry of ρ is essential to the optimal structure of the statistical estimation based on ρ. Thus we introduce a symbol $*$ defined by

$$\rho^*(f, g) = \rho(g, f)$$

for each f and g in $\mathscr{F}_0$.

We note that for any $\delta > 0$, ρ^δ is also a contrast functional. Hence we assume that a contrast functional ρ has the same order as the squared Hellinger distance

$$H^2(f, g) = 2\int (\sqrt{f} - \sqrt{g})^2 d\mu, \tag{2.1}$$

i.e., there exists a positive number ε such that

$$\lim_{t \to 0} \frac{\rho(f_t, f)}{H^2(f_t, f)} = \varepsilon \tag{2.2}$$

for every smooth curve $\{f_t: |t| < \delta\}$ through f at $t=0$.

Let w be a function defined on the real half-axis $(0, \infty)$ which satisfies

$$w(t) - w(1) > w'(1)(t-1)$$

for each $t > 0$, $t \neq 1$. Then by the function w, we define

$$\rho_w(f, g) = \boldsymbol{E}_f w\left[\frac{g(\boldsymbol{x})}{f(\boldsymbol{x})}\right] - w(1).$$

The functional ρ_w becomes a contrast functional since

$$\rho_w(f, g) = \rho_{\tilde{w}}(f, g)$$

and $\tilde{w}(1)=0$, $\tilde{w}(t) > 0$ for each $t > 0$, $t \neq 1$, where

$$\tilde{w}(t) = w(t) - w(1) - w'(1)(t-1).$$

From these properties we may assume without loss of generality that $w(1)=w'(1)=0$. Further w is normalized to statisfy $w''(1)=1$ since

$$\rho_{\delta w} = \delta \rho_w$$

for every constant $\delta>0$. Then it is easily seen that the limit ε for ρ_w in (2.2) is equal to 1. Thus $\mathscr{W}$ denotes the space of all C^2-differentiable functions such that $w(1)=w'(1)=0$, $w''(1)=1$ and $w(t)>0$ if $t>0$, $t\neq 1$. Henceforth we call ρ_w *a contrast functional of weighted ratio type* (*of W-type*, for short).

Example 2.1. Most of classical contrast functionals, or divergences can be expressed in terms of W-type as follows:

(1) *Kullback–Leibler information* (Kullback–Leibler [38]):

$$\rho_{KL}(f, g) = \int f(\log f - \log g) d\mu.$$

(2) *squared Hellinger distance*:

$$H^2(f, g) = 2 \int (\sqrt{f} - \sqrt{g})^2 d\mu.$$

(3) *Jeffereys divergence* (Jeffereys [35]):

$$\rho_J(f, g) = \frac{1}{2} \int (f-g)(\log f - \log g) d\mu.$$

(4) *Chernoff information of order* α (Chernoff [18]):

$$\rho_\alpha(f, g) = \frac{4}{1-\alpha^2} \left\{ 1 - \int f^{\frac{1-\alpha}{2}} g^{\frac{1+\alpha}{2}} d\mu \right\}.$$

(5) *exponential divergence*:

$$\rho_e(f, g) = \frac{1}{2} \int f(\log f - \log g)^2 d\mu.$$

(6) *the Kagan divergence* (Kagan [36]):

$$\rho_{\chi^2}(f, g) = \frac{1}{2} \int \frac{(f-g)^2}{f} d\mu.$$

(7) *the product divergence with* (α, β)*-index*

$$\rho_{\alpha\beta}(f, g) = \frac{2}{(1-\alpha)(1-\beta)} \int \left\{ 1 - \left(\frac{g}{f}\right)^{\frac{1-\alpha}{2}} \right\} \left\{ 1 - \left(\frac{g}{f}\right)^{\frac{1-\beta}{2}} \right\} f \, d\mu.$$

All the divergences (1)–(7) are of W-type. For example, the Chernoff infro-

mation of order α can be rewritten as ρ_{w_α} with

$$w_\alpha(t) = \frac{4}{1-\alpha^2}(1-t^{\frac{1+\alpha}{2}}) + \frac{2}{1-\alpha}(t-1),$$

while the product divergence $\rho_{\alpha\beta}$ with (α, β)-index can be similarly expressed by

$$w_{\alpha\beta}(t) = \frac{2}{(1-\alpha)(1-\beta)}(1-t^{\frac{1-\alpha}{2}})(1-t^{\frac{1-\beta}{2}}).$$

The divergence $\rho_{\alpha\beta}$ is newly introduced here, which connects many other classical contrast functionals in the following way:

$$\rho_{00} = H^2$$

$$\rho_{-\alpha\alpha} = \frac{1}{2}(\rho_\alpha + \rho_{-\alpha}),$$

$$\lim_{\alpha\to 1} \rho_{-\alpha\alpha} = \rho_J,$$

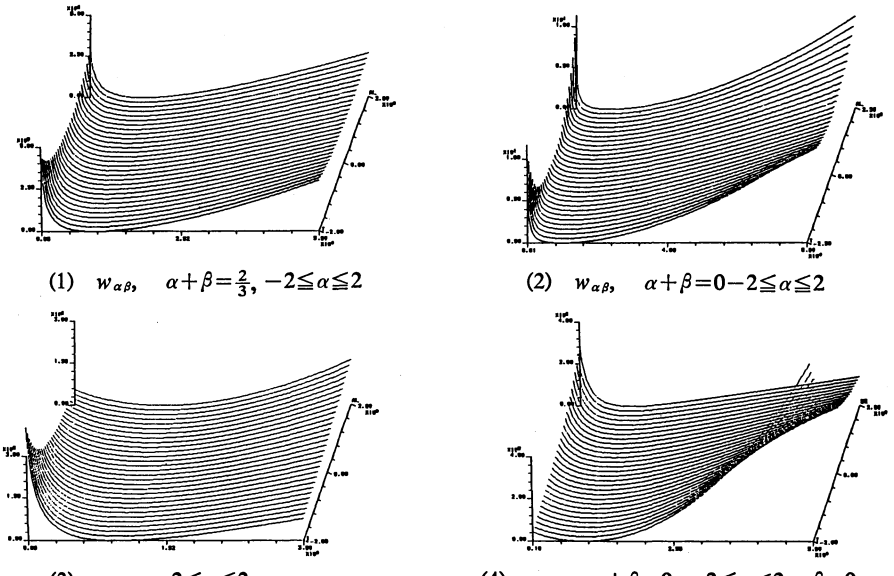
$$\lim_{\alpha\to 1} \rho_{\alpha\alpha} = \rho_m,$$

$$\lim_{\alpha\to -1} \rho_{\alpha\alpha} = \rho_e$$

and

$$\rho_{\alpha\beta} = \frac{1+\alpha}{2(1-\beta)}\rho_{-\alpha} + \frac{1+\beta}{2(1-\alpha)}\rho_{-\beta} - \frac{(\alpha+\beta)(2-\alpha-\beta)}{2(1-\alpha)(1-\beta)}\rho_{\frac{1-\alpha-\beta}{2}}.$$

The graphs of w_α and $w_{\alpha\beta}$ are given in the following figure.



(1) $w_{\alpha\beta}$, $\alpha+\beta=\frac{2}{3}$, $-2\leqq\alpha\leqq 2$ (2) $w_{\alpha\beta}$, $\alpha+\beta=0$ $-2\leqq\alpha\leqq 2$

(3) w_α, $-2\leqq\alpha\leqq 2$ (4) $w_{\alpha\beta}$, $\alpha+\beta=0$, $-2\leqq\alpha\leqq 2$, $\beta=0$.

Fig. 1. The comparison between the graphs of w_α and $w_{\alpha\beta}$

The exponential divergence ρ_e is analogous to the quasi-distance on the family of spectral densities in *stationary Gaussian processes* used by Taniguchi [60]. The Kagan divergence ρ_{χ^2} is reduced to

$$\chi^2 = \sum_{i=1}^{n} \frac{(\pi_i - p_i)^2}{\pi_i}$$

if f and g are multinomial distributions with cell probability vectors $\boldsymbol{p}$ and $\boldsymbol{\pi}$, which justifies the notation of ρ_{χ^2}. We shall give a derivation of ρ_e and ρ_{χ^2} by a functional-analitic approach in Section 4.

The explicit forms of the contrast functionals (1)–(7) for the most familiar model to us:

$$\mathcal{G}_1 = \left\{ f_{\mu}(x) = \frac{1}{(2\pi)^{\frac{1}{2}k}} e^{-\frac{1}{2}\|x-\mu\|^2} : \mu \in R^k \right\}$$

are given as follows:

$$\rho_{KL}(f_{\mu}, f_m) = \rho_J(f_{\mu}, f_m) = \frac{1}{2}\|\mu - m\|^2$$

$$H^2(f_{\mu}, f_m) = 4\left[1 - \exp\left\{-\frac{\|\mu - m\|^2}{8}\right\}\right],$$

$$\rho_{\alpha}(f_{\mu}, f_m) = \frac{4}{1-\alpha^2}\left[1 - \exp\left\{-\frac{1-\alpha^2}{8}\|\mu - m\|^2\right\}\right],$$

$$\rho_e(f_{\mu}, f_m) = \frac{1}{2}\|\mu - m\|^2\left[1 + \frac{\|\mu - m\|^2}{4}\right],$$

$$\rho_m(f_{\mu}, f_m) = \frac{1}{2}\exp\{\|m - \mu\|^2\} - \frac{1}{2}$$

and

$$\rho_{\alpha\beta}(f_{\mu}, f_m) = \frac{2}{(1+\alpha)(1+\beta)}\left[1 - \exp\left\{-\frac{(1-\alpha)(1+\beta)}{8}\|m - \mu\|^2\right\}\right.$$
$$\left. - \exp\left\{-\frac{(1-\beta)(1+\alpha)}{8}\|m - \mu\|^2\right\} + \exp\left\{\frac{(2+\alpha+\beta)(\alpha+\beta)}{8}\|m - \mu\|^2\right\}\right].$$

Notice that functionals (1)–(6) have various forms even for the simple model $\mathcal{G}_1$ while keeping $\varepsilon = 1$ in (2.2).

We now return to the subject of a general functional of W-type. The smoothness of w gaurantees a unique representation of ρ_w.

THEOREM 2.1. *Suppose that the functions w_1 and w_2 in $\mathcal{W}$ are analytic. Then $\rho_{w_1} = \rho_{w_2}$ on $\mathcal{F}_0$ if and only if $w_1 = w_2$.*

PROOF. The sufficiency is clear. We show the necessity. Assume $\rho_{w_1} = \rho_{w_2}$ on $\mathcal{F}_0$. Expand w_1 and w_2 at $t = 1$:

$$w_p(t) = \sum_{k=2}^{\infty} w_p^{(k)}(1) \frac{(t-1)^k}{k!}$$

for $p=1, 2$. Then it holds that

$$\rho_{w_1}(f, tf+(1-t)g) - \rho_{w_2}(f, tf+(1-t)g)$$
$$= \sum_{k=3}^{\infty} \frac{w_1^{(k)}(1)-w_2^{(k)}(1)}{k!} (t-1)^k \int \left(\frac{g-f}{f}\right)^k f \, d\mu$$

for any t, $0 \leq t < 1$ and every f and g in $\mathscr{F}_0$. Therefore we conclude that

$$w_1^{(k)}(1) = w_2^{(k)}(1)$$

for $k=3, 4, \ldots$, since it can be satisfied that

$$0 < \left| \int \left(\frac{g-f}{f}\right)^k f d\mu \right| < M$$

by choosing f and g sufficiently close to each other. This completes the proof.

Let $\mathscr{W}_1$ be the subclass of $\mathscr{W}$ whose members are defined to be convex on $(0, \infty)$. Note that $w_{\alpha\beta}$ does not belong to $\mathscr{W}_1$. Csiszar [19] considered a transformation $*$ on $\mathscr{W}_1$ as

$$w^*(t) = tw(t^{-1})$$

for w in $\mathscr{W}_1$. Note that w^* is in $\mathscr{W}_1$ and $w^{**}=w$. By a brief manipulation, the symbol $*$ yields that $(\rho_w)^* = \rho_{w^*}$ on $\mathscr{F}_0$ for every w in $\mathscr{W}_1$, which leads us to the following theorem on account of Theorem 2.1.

THEOREM 2.2. *Let w be analytic on $(0, \infty)$. Then a contrast functional ρ_w is symmetric if and only if $w=w^*$. Under this assumption, it holds that*

$$w'''(1) = -\frac{3}{2}.$$

PROOF. The former statement is easily proved. So we show the latter statement. From the assumption of symmetry we have

$$w'''(t) = \frac{1}{2}\{w'''(t)+w^{*'''}(t)\}.$$

Differentiating $w^*(t)=tw(t^{-1})$, we have

$$w'''(t) + w^{*'''}(t^{-1}) = -3t^{-4}w''(t^{-1}).$$

These two formulas imply $w'''(1)=-3/2$.

We next introduce an operation $\oplus$ on $\mathscr{W}_1$ as

$$w^{\oplus}(t) = tw(t) - 2\int_1^t w(s)ds.$$

We write

$$\rho_w^{\oplus} = \rho_{w^{\oplus}}.$$

Note that $w^{\oplus}$ belongs to $\mathscr{W}_1$. The classical contrast functionals, or divergences as in Example 2.1 are connected with each other in terms of $\oplus$ and $*$:

$$\rho_{KL}^{\oplus *} = \rho_{KL},$$
$$\rho_{KL}^{\oplus\oplus} = \rho_m,$$
$$\rho_m^{*\oplus} = \rho_{KL},$$
$$\rho_m^{\oplus} = \rho_\alpha \quad (\alpha = 5)$$

and

$$\rho_\alpha^{\oplus} = \rho_{\alpha+2}$$

for any α in $\boldsymbol{R}$. Thus the family of all Chernoff informations is closed with respect to both operations $*$ and $\oplus$. Furthermore let

$$w^{\ominus}(t) = t^{-1}w(t) + 2\int_1^t s^{-2}w(s)ds + 2\int_1^t\int_1^s u^{-2}w(u)duds.$$

Then the operation $\ominus$ yields the following properties:

THEOREM 2.3. *It holds that*

$$w^{\ominus\oplus} = w^{\oplus\ominus} = w,$$

or

$$\rho_w^{\ominus\oplus} = \rho_w^{\oplus\ominus} = \rho_w$$

for any w in $\mathscr{W}_1$.

PROOF. By direct differentiation we have

$$\frac{d^2}{dt^2} w^{\ominus\oplus}(t) = \frac{d^2}{dt^2} w^{\ominus\oplus}(t) = \frac{d^2}{dt^2} w(t)$$

for any w in $\mathscr{W}_1$, which completes the proof because of $w(1) = w'(1) = 0$.

It follows from Theorem 2.3 that the operation $\ominus$ is the inverse mapping of $\oplus$. We note that

$$\lim_{n\to\infty} w^{\overbrace{\oplus\cdots\oplus}^{n\text{-times}}} = w_\infty$$

and

$$\lim_{n\to\infty} w^{\overbrace{\ominus\cdots\ominus}^{n-\text{times}}} = w^*_\infty,$$

where

$$w_\infty(t) = \begin{cases} 0 & \text{if } 0 < t \leq 1. \\ \infty & \text{otherwise.} \end{cases}$$

The operations $*$, $\oplus$ and $\ominus$ are bijective. These are connected with each other in the following way:

$$\rho_w^{*\ominus*\ominus} = \rho_w^{\ominus*\ominus*} = \rho_w^{*\oplus*\oplus} = \rho_w^{\oplus*\oplus*} = \rho_w$$

for any w in $\mathscr{W}_1$.

Now we investigate a scale transformation on $\mathscr{W}$. Let w_σ be

$$w_\sigma(t) = \frac{\sigma^2}{w''(\sigma^{-1})}\{w(\sigma^{-1}t) - w(\sigma^{-1})\}$$

It is easily seen that, for every positive constants σ and v, the function w_σ is in $\mathscr{W}$ and $w_{\sigma v} = (w_\sigma)_v$. We say ρ_w to be *scale invariant* if

$$\rho_w = \rho_{w_\sigma}$$

for any $\sigma > 0$. The scale invariance implies the following: Let $\mathfrak{M}$ be a space of all finite measures which are equivalent to the carrier measure μ. We denote the density form of $\mathfrak{M}$ with respect to μ by $\mathscr{M}$. The space $\mathscr{M}$ is a convex cone in the L_1 space. An equivalent relation $\sim$ on $\mathscr{M}$ is introduced as follows: $m_1 \sim m_2$ if there exists a constant σ such that $m_1(\boldsymbol{x}) = \sigma m_2(\boldsymbol{x})$ for $\mu - a.e.$ $\boldsymbol{x}$. Then it holds that $\mathscr{M}/\sim$ is isomorphic to $\mathscr{F}$. In this context, let

$$\delta(m_1, m_2) = \rho_{w_\sigma}(f_1, f_2),$$

where $f_i = m_i/|m_i|$ and $\sigma = |m_1|/|m_2|$ with $|m_i| = \int m_i \, d\mu$ for $i = 1, 2$. Then it follows that

$$\delta(m_1, m_2) \geqq 0$$

for all m_1 and m_2 in $\mathscr{M}$ with the equality if and only if $m_1 \sim m_2$. The scale invariance is equivalent to the following condition:

$$\delta(\sigma m_1, v m_2) = \delta(m_1, m_2)$$

for every positive scalars σ and v. That is to say, δ is well-defined as a functional over $\mathscr{M}/\sim$. We have a characterization of scale invariance:

THEOREM 2.4. *A contrast functional ρ_w is scale invariant if and only if ρ_w is the Chernoff information of order α.*

PROOF. Assume the scale invariance of ρ_w. Then it holds that

$$w''(st) = w''(s)w''(t) \tag{2.3}$$

for every positive numbers s and t. From the assumption of smoothness for w it follows that the function satisfying (2.3) is uniquely determined up to the form

$$w''(t) = t^k$$

with a constant k since

$$\log w''(e^{u+v}) = \log w''(e^u) + \log w''(e^v)$$

for any u and v in $\boldsymbol{R}$. Therefore the function w is, with relation to $k=(-3+\alpha)/2$, represented as

$$w(t) = \frac{4}{1-\alpha^2}(1-t^{\frac{1+\alpha}{2}}) + \frac{2}{1-\alpha}(t-1)$$

which generates the Chernoff information of order α. The inverse is easily seen. The proof is complete.

Throughout this section, our interests have been focussed on the class of contrast functionals of W-type. However this class may be restrictive in the class of all contrast functionals. We can introduce a variety of representations with respect to a contrast functional ρ: Let Φ be a C^2-differentiable monotone function with $\Phi(0)=0$ and $\Phi'(0)=1$. In terms of Φ, define

$$\rho_{\Phi_\sigma}(f, g) = \sigma^{-1}\Phi[\sigma\rho(f, g)]$$

with a non-zero constant σ. Then ρ_{Φ_σ} is also a contrast functional with the same order as ρ.

The original form of the Chernoff information of order α can be changed into

$$\tilde{\rho}_\alpha(f, g) = \frac{4}{\alpha^2-1}\log\int f^{\frac{1-\alpha}{2}} g^{\frac{1+\alpha}{2}} d\mu \tag{2.4}$$

if we take

$$\Phi_\alpha(t) = \frac{4}{\alpha^2-1}\log\left(1+\frac{4}{\alpha^2-1}t\right)$$

as Φ. The transformation generates the additivity of information:

$$\tilde{\rho}_\alpha(f, g) = \tilde{\rho}_\alpha(f_1, g_1) + \tilde{\rho}_\alpha(f_2, g_2)$$

with $f(x, y)=f_1(x)f_2(y)$ and $g(x, y)=g_1(x)g_2(y)$. On the other hand, the additivity property of ρ_α holds if and only if $\alpha=1$ or -1, *i.e.*, $\rho_\alpha=\rho_{KL}$ or ρ_{KL}^*.

In the following section we shall give a geometry of the parametric family on the basis of the contrast functional ρ. It will be shown that the geometry is independent of the Φ-representation of ρ.

3. Geometry generated by a contrast functional: finite dimensional cases

We return to the investigation of the geometry on a parametric family

$$\mathcal{F} = \{f_\theta : \theta \in \Theta\}$$

in the space $\mathcal{F}_0$ of all probability densities with a common support. Here the dimension of $\mathcal{F}$ is assumed to be finite throughout this section. Let ρ be a contrast functional on $\mathcal{F}_0$. The restiriction of ρ to $\mathcal{F}$ is written as

$$\rho(\theta_1, \theta_2) = \rho(f_{\theta_1}, f_{\theta_2})$$

with respect to θ-coordinates, which will be called a contrast function on $\mathcal{F}$. Then we have an intuition that a contrast function on $\mathcal{F}$ tells us *the statistical intrincity* of the model $\mathcal{F}$ if ρ grasps statistical backgrounds.

We use the symbols ε_i and δ_j for the partial differentials $\partial/\partial\theta_1^i$ and $\partial/\partial\theta_2^j$ respectively. A metric tensor $g^{(\rho)}$ on $\mathcal{F}$ is defined by components

$$g_{ij}^{(\rho)}(\theta) = (\varepsilon_i\varepsilon_j\rho(\theta_1, \theta_2))_{\theta_1=\theta_2=\theta} = \varepsilon_i\varepsilon_j\rho(\theta, \theta)$$

with respect to θ. A pair of affine connections $\Gamma^{(\rho)}$ and ${}^*\Gamma^{(\rho)}$ are introduced by defining the coefficients as

$$\Gamma_{ij,k}^{(\rho)}(\theta) = -\varepsilon_i\varepsilon_j\delta_k\rho(\theta, \theta)$$

and

$${}^*\Gamma_{ij,k}^{(\rho)}(\theta) = -\delta_i\delta_j\varepsilon_k\rho(\theta, \theta),$$

respectively (cf. Eguchi [23]). Note that the geometric objects $g^{(\rho)}$, $\Gamma^{(\rho)}$ and ${}^*\Gamma^{(\rho)}$ satisfy the transformation law for the coordinates transformations.

We give some indentities which will be used in a subsequent discussion. Since the contrast function attains a minimum at $\theta_1=\theta_2$ $(=\theta$, say), it holds that

$$\varepsilon_i\rho(\theta, \theta) = 0 \tag{3.1}$$

and

$$\delta_j\rho(\theta, \theta) = 0 \tag{3.2}$$

for $i, j=1, 2,\ldots, n$, which lead, by differentiating the both sides of (3.1) and (3.2), to

$$\partial_j \varepsilon_i \rho(\boldsymbol{\theta}, \boldsymbol{\theta}) = \varepsilon_j \varepsilon_i \rho(\boldsymbol{\theta}, \boldsymbol{\theta}) + \varepsilon_j \delta_i \rho(\boldsymbol{\theta}, \boldsymbol{\theta}) = 0 \tag{3.3}$$

and

$$\partial_i \delta_j \rho(\boldsymbol{\theta}, \boldsymbol{\theta}) = \varepsilon_i \delta_j \rho(\boldsymbol{\theta}, \boldsymbol{\theta}) + \delta_i \delta_j \rho(\boldsymbol{\theta}, \boldsymbol{\theta}) = 0, \tag{3.4}$$

where $\partial_i = \partial/\partial\theta^i$. Furthermore we have

$$\varepsilon_k \varepsilon_j \varepsilon_i \rho + \delta_k \varepsilon_j \varepsilon_i \rho + \varepsilon_k \delta_j \varepsilon_i \rho + \delta_k \delta_j \varepsilon_i \rho = 0$$

and

$$\varepsilon_k \delta_j \varepsilon_i \rho + \delta_k \delta_j \varepsilon_i \rho + \varepsilon_k \delta_j \delta_i \rho + \delta_k \delta_j \delta_i \rho = 0$$

for $i, j, k = 1, 2, \ldots, n$, where the arguments of ρ are abbreviated. The above identities can be rewritten as

$$g_{ij}^{(\rho)} = \delta_i \delta_j \rho = -\delta_i \varepsilon_j \rho = -\varepsilon_i \delta_j \rho,$$

$$\varepsilon_k \varepsilon_j \varepsilon_i \rho = \Gamma_{ij,k}^{(\rho)} + {}^*\Gamma_{jk,i}^{(\rho)} + \Gamma_{ki,j}^{(\rho)} \tag{3.5}$$

and

$$\delta_k \delta_j \delta_i \rho = {}^*\Gamma_{ij,k}^{(\rho)} + {}^*\Gamma_{jk,i}^{(\rho)} + \Gamma_{ki,j}^{(\rho)}, \tag{3.6}$$

Using these relations, we have the following lemmas.

LEMMA 1. *Let $\overset{0}{\Gamma}{}^{(\rho)}$ be the metric connection with respect to $g^{(\rho)}$. Then it holds that*

$$\overset{0}{\Gamma}{}^{(\rho)} = \frac{1}{2}(\Gamma^{(\rho)} + {}^*\Gamma^{(\rho)}).$$

PROOF. By definition, $\overset{0}{\Gamma}{}^{(\rho)}$ has the coefficients

$$\overset{0}{\Gamma}{}_{ij,k}^{(\rho)} = \frac{1}{2}(\partial_i g_{jk}^{(\rho)} + \partial_j g_{ki}^{(\rho)} - \partial_k g_{ij}^{(\rho)}),$$

for $i, j, k = 1, 2, \ldots, n$, which are expressed as

$$\frac{1}{2}(\varepsilon_k \varepsilon_j \varepsilon_i \rho - \Gamma_{jk,i}^{(\rho)} - \Gamma_{ki,j}^{(\rho)} + \Gamma_{ij,k}^{(\rho)}). \tag{3.7}$$

By inserting (3.5) into (3.7), we have

$$\overset{0}{\Gamma}{}_{ij,k}^{(\rho)} = \frac{1}{2}(\Gamma_{ij,k}^{(\rho)} + {}^*\Gamma_{ij,k}^{(\rho)}).$$

This completes the proof.

Define a tensor $T^{(\rho)}$ of order 3 as

$$T^{(\rho)} = \Gamma^{(\rho)} - {}^*\Gamma^{(\rho)}.$$

The difference between (3.5) and (3.6) leads us to

$$T^{(\rho)}_{ij,k} = \varepsilon_k\varepsilon_j\varepsilon_i\rho - \delta_k\delta_j\delta_i\rho,$$

which shows the following lemma.

LEMMA 2. *The tensor $T^{(\rho)}$ is symmetric.*

Let Γ be an affine connection and let C be a curve in $\mathscr{F}$ with a parameter t, i.e.,

$$C = \{f_{\theta(t)} : |t| < \varepsilon\}.$$

Then a vector field X on $\mathscr{F}$ is said to be parallel along C with respect to Γ if

$$\dot{X}^i(t) + \Gamma^i_{jk}(\theta(t))X^k(t)\dot{\theta}^j(t) = 0,$$

where X is expressed as $X(t) = X^i(t)e_i(\theta(t))$ with the natural basis $\{e_i\}$. The correspondance $\pi: X(0) \to X(t)$ on X is called the parallel shift with respect to Γ. By using Lemmas 1 and 2, we show the conjugacy between $\Gamma^{(\rho)}$ and ${}^*\Gamma^{(\rho)}$.

THEOREM 3.1. *Let π and ${}^*\pi$ be the parallel shifts with respect to $\Gamma^{(\rho)}$ and ${}^*\Gamma^{(\rho)}$, respectively. Then it holds that*

$$g^{(\rho)}(\pi X, {}^*\pi Y) = g^{(\rho)}(X, Y)$$

for all vector fields X and Y on $\mathscr{F}$.

PROOF. Take a curve

$$C = \{f_{\theta(t)} : |t| < \varepsilon\}$$

in $\mathscr{F}$. We write

$$\pi_{0t}X = X^i(t)e_i(\theta(t))$$

and

$${}^*\pi_{0t}Y = {}^*Y^i(t)e_i(\theta(t)),$$

where $\pi_{0,t}$ and ${}^*\pi_{0,t}$ denote the parallel shifts along C with respect to $\Gamma^{(\rho)}$ and ${}^*\Gamma^{(\rho)}$, respectively. Thus it holds that

$$\dot{X}^i + \Gamma^{(\rho)i}_{jk}X^k\dot{\theta}^j = 0$$

and

$${}^*\dot{Y}^i + {}^*\Gamma^{(\rho)i}_{jk}{}^*Y^k\dot{\theta}^j = 0$$

for $i=1, 2,\ldots, n$. Therefore

$$\frac{d}{dt} g^{(\rho)}(\pi_{0t}X, {}^*\pi_{0t}Y) = (\partial_k g^{(\rho)}_{ij} - \Gamma^{(\rho)}_{ki,j} - {}^*\Gamma^{(\rho)}_{jk,i}) X^i {}^*Y^j \dot{\theta}^k.$$

From Lemmas 1 and 2 it follows that

$$\Gamma^{(\rho)} = \overset{0}{\Gamma}{}^{(\rho)} + \frac{1}{2} T^{(\rho)}$$

and

$${}^*\Gamma^{(\rho)} = \overset{0}{\Gamma}{}^{(\rho)} - \frac{1}{2} T^{(\rho)}$$

with the metric connection $\overset{0}{\Gamma}{}^{(\rho)}$ and the symmetric tensor $T^{(\rho)}$. Hence we have the identity:

$$\partial_k g^{(\rho)}_{ij} - \Gamma^{(\rho)}_{ki,j} - {}^*\Gamma^{(\rho)}_{jk,i} = 0$$

for $i, j, k=1, 2,\ldots, n$, which shows that $g^{(\rho)}(\pi_{0t}X, {}^*\pi_{0t}Y)$ is constant in t. This completes the proof.

One notes that if $\overset{0}{\pi}$ is the parallel shift with respect to the metric connection $\overset{0}{\Gamma}{}^{(\rho)}$, then

$$g^{(\rho)}(\overset{0}{\pi}X, \overset{0}{\pi}Y) = g^{(\rho)}(X, Y)$$

Considering π and ${}^*\pi$ as a conjugate version of $\overset{0}{\pi}$, we call a triple

$$\mathscr{C}(\rho) = (g^{(\rho)}, \Gamma^{(\rho)}, {}^*\Gamma^{(\rho)})$$

a conjugate metric structure generated by ρ (cf. Fig. 2).

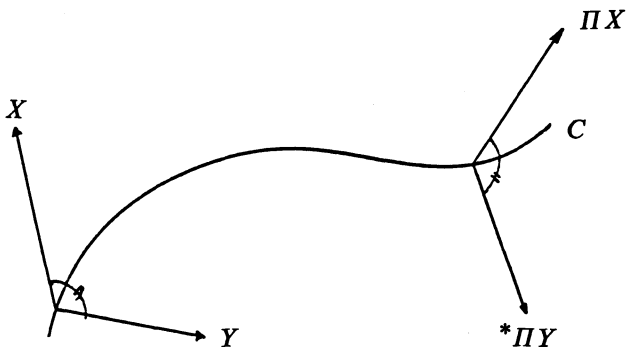


Fig. 2. The compatibility the metric $g^{(\rho)}$ and a pair of parallelisms with respect to $\Gamma^{(\rho)}$ and ${}^*\Gamma^{(\rho)}$.

Amari [2] showed that the structures $\mathscr{C}_s$ and $\mathscr{C}_\alpha$ have the same conjugacy as in $\mathscr{C}(\rho)$. In fact the symmetry of the tensor

$$T = \overset{m}{\Gamma} - \overset{e}{\Gamma}$$

is easily shown on account of

$$T_{ijk}(\boldsymbol{\theta}) = \boldsymbol{E}_{\boldsymbol{\theta}} e_i(\boldsymbol{\theta}) e_j(\boldsymbol{\theta}) e_k(\boldsymbol{\theta}).$$

Lauritzen [42] referred to T as the skewness of the model $\mathscr{F}$. On the other hand, the tensor $T^{(\rho)}$ expreses the non-symmetry of ρ. Symmetry of a contrast function ρ leads to the freeness of $T^{(\rho)}$ although the inverse statement does not hold in general.

Let $\varepsilon(\boldsymbol{\theta})$ be a positive scalar function on $\mathscr{F}$. For a contrast function ρ,

$$\tilde{\rho}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \varepsilon(\boldsymbol{\theta}_2)\rho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$$

becomes a contrast function again. In comparison between $\mathscr{C}(\rho) = (g^{(\rho)}, \Gamma^{(\rho)}, {}^*\Gamma^{(\rho)})$ and $\mathscr{C}(\tilde{\rho}) = (g^{(\tilde{\rho})}, \Gamma^{(\tilde{\rho})}, {}^*\Gamma^{(\tilde{\rho})})$, we have the following relations

$$g^{(\tilde{\rho})}_{ij} = \varepsilon g^{(\rho)}_{ij},$$

$$\Gamma^{(\tilde{\rho})}_{ij,k} = \varepsilon\, \Gamma^{(\rho)}_{ij,k} - g^{(\rho)}_{ij}\partial_k \varepsilon$$

and

$${}^*\Gamma^{(\tilde{\rho})}_{ij,k} = \varepsilon\, {}^*\Gamma^{(\rho)}_{ij,k} + g^{(\rho)}_{jk}\, \partial_i \varepsilon + g^{(\rho)}_{ki}\, \partial_j\, \varepsilon.$$

The first relation shows that the metric tensor $g^{(\rho)}$ and $g^{(\tilde{\rho})}$ are in conformal correspondance. This implies that the angle between any two vectors with respect to $g^{(\rho)}$ agrees with that with respect to $g^{(\tilde{\rho})}$.

The structure $\mathscr{C}(\rho)$ leads to the following expansion by neglecting the fourth and higher order terms:

$$\begin{aligned}(3.8)\qquad \rho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = {} & \frac{1}{2}\, g^{(\rho)}_{ij}(\boldsymbol{\theta})(\bar{\theta}^i_1 - \bar{\theta}^i_2)\,(\bar{\theta}^j_1 - \bar{\theta}^j_2) \\ & + \frac{1}{6}\,\{2\Gamma^{(\rho)}_{ij,k}(\boldsymbol{\theta}) + {}^*\Gamma^{(\rho)}_{ij,k}(\boldsymbol{\theta})\}\bar{\theta}^i_1\bar{\theta}^j_1\bar{\theta}^k_1 \\ & - \frac{1}{2}\,\{\Gamma^{(\rho)}_{ij,k}(\boldsymbol{\theta})\bar{\theta}^i_1\bar{\theta}^j_1\bar{\theta}^k_2 + {}^*\Gamma^{(\rho)}_{ij,k}(\boldsymbol{\theta})\bar{\theta}^i_2\bar{\theta}^j_2\bar{\theta}^k_1\} \\ & + \frac{1}{6}\,\{2{}^*\Gamma^{(\rho)}_{ij,k}(\boldsymbol{\theta}) + \Gamma^{(\rho)}_{ij,k}(\boldsymbol{\theta})\}\bar{\theta}^j_2\bar{\theta}^j_2\bar{\theta}^k_2\end{aligned}$$

for any $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ close to $\boldsymbol{\theta}$ with $\bar{\boldsymbol{\theta}}_p = \boldsymbol{\theta}_p - \boldsymbol{\theta}$ $(p=1, 2)$. Note that the formula holds for any coordinate system. Let Φ be a monotone C^2-function with $\Phi(0)=0$

and $\Phi'(0)=1$. As discussed in Section 2, a contrast functional ρ has the representation $\Phi(\rho)$. Note that the formula (3.8) is independent of this representation, since the difference between ρ and $\Phi(\rho)$ is of the fourth or higher order at least.

Let $\mathscr{P}$ be the space of all contrast functionals over $\mathscr{F}_0$. The space $\mathscr{P}$ is closed under manipulations defind as

$$\rho_{\boxtimes\alpha} = \rho_1^{\frac{1+\alpha}{2}} \rho_2^{\frac{1-\alpha}{2}}$$

and

$$\rho_{\boxplus\alpha} = \frac{1+\alpha}{2}\rho_1 + \frac{1-\alpha}{2}\rho_2$$

for ρ_1 and ρ_2 in $\mathscr{P}$ and for all α, $-1<\alpha<1$.

THEOREM 3.2. *Under the above setup, the conjugate metric structures of* $\rho_{\boxtimes\alpha}$ *and* $\rho_{\boxplus\alpha}$ *satisfy*

$$\mathscr{C}(\rho_{\boxtimes\alpha}) = \mathscr{C}(\rho_{\boxplus\alpha}) = \frac{1+\alpha}{2}\mathscr{C}(\rho_1) + \frac{1-\alpha}{2}\mathscr{C}(\rho_2)$$

if the tensor $g^{(\rho_1)}$ *is equal to* $g^{(\rho_2)}$.

PROOF. The second equality is immediate. We show the first equality. By differentiating $\rho_{\boxtimes\alpha}$, we have

$$\varepsilon_j\varepsilon_k\rho_{\boxtimes\alpha} = \frac{1+\alpha}{2}\left(\frac{\rho_2}{\rho_1}\right)^{\frac{1-\alpha}{2}}\delta_j\varepsilon_k\rho_1 + \frac{1-\alpha}{2}\left(\frac{\rho_1}{\rho_2}\right)^{\frac{1+\alpha}{2}}\delta_j\varepsilon_k\rho_2$$
$$+ \frac{1-\alpha^2}{4}\left\{\left(\frac{\rho_1}{\rho_2}\right)^{\frac{1+\alpha}{2}}\left(\varepsilon_j\frac{\rho_2}{\rho_1}\right)\delta_k\rho_1 + \left(\frac{\rho_2}{\rho_1}\right)^{\frac{1-\alpha}{2}}\varepsilon_j\left(\frac{\rho_1}{\rho_2}\right)\delta_k\rho_2\right\}$$

for $j, k=1, 2,\ldots, n$. It holds that

$$\lim_{\theta_2\to\theta_1}\frac{\rho_1(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{\rho_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)} = 1$$

because of $g^{(\rho_1)}=g^{(\rho_2)}$. Therefore we have

$$g^{(\rho\boxtimes\alpha)} = \frac{1+\alpha}{2}g^{(\rho_1)} + \frac{1-\alpha}{2}g^{(\rho_2)}$$

Similarly it follows that

$$\Gamma^{(\rho\boxtimes\alpha)} = \frac{1+\alpha}{2}\Gamma^{(\rho_1)} + \frac{1-\alpha}{2}\Gamma^{(\rho_2)}$$

and the same relation holds also for $*\Gamma^{(\rho\boxtimes\alpha)}$. The proof is complete.

Note that

$$\mathscr{C}(\rho_{\boxtimes\alpha}) = \mathscr{C}_\alpha$$

for any α, $-1<\alpha<1$ if $(\rho_1, \rho_2)=(\rho_{KL}, \rho_{KL}^*)$ with the Kullback–Leibler divergence ρ_{KL} and

$$\rho_{KL}^*(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \rho_{KL}(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1).$$

Let us see the conjugate metric structures generated by classical contrast functions. We need the following assumption:

A–3. *The definite integral*

$$\rho_w(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \int w\left[\frac{f_{\boldsymbol{\theta}_2}(x)}{f_{\boldsymbol{\theta}_1}(x)}\right] f_{\boldsymbol{\theta}_1}(x)\, d\mu(x)$$

is twice differentiable with respect to $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ under the integral sign for any $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ sufficiently close to each other.

The following proposition is a straightforward extension of Eguchi's results [23] in an exponential family to the regular parametric family $\mathscr{F}$.

THEOREM 3.3. *Let ρ_w be a contrast function of W-type on $\mathscr{F}$. Then under the assumption* A–3 *we have*

$$\mathscr{C}(\rho_w) = \mathscr{C}_{\alpha_w} \tag{3.9}$$

on $\mathscr{F}$ where

$$\alpha_w = 3 + 2w'''(1).$$

PROOF. By the assumption A–3,

$$\varepsilon_i\delta_j\rho_w(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = -E_{\boldsymbol{\theta}_2}\left\{w''\left[\frac{f_{\boldsymbol{\theta}_2}(x)}{f_{\boldsymbol{\theta}_1}(x)}\right]\left(\frac{f_{\boldsymbol{\theta}_2}(x)}{f_{\boldsymbol{\theta}_1}(x)}\right) e_i(\boldsymbol{\theta}_1)\, e_j(\boldsymbol{\theta}_2)\right\}, \tag{3.10}$$

where E_θ denotes the expectation with respect to f_θ. Furthermore it follows from A–3 that

$$\begin{aligned}(3.11)\quad &\varepsilon_i\varepsilon_j\delta_k\rho_w(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\\ &= E_{\boldsymbol{\theta}_2}\left[\left\{w'''\left[\frac{f_{\boldsymbol{\theta}_2}(x)}{f_{\boldsymbol{\theta}_1}(x)}\right]\left(\frac{f_{\boldsymbol{\theta}_2}(x)}{f_{\boldsymbol{\theta}_1}(x)}\right)^2 + w''\left[\frac{f_{\boldsymbol{\theta}_2}(x)}{f_{\boldsymbol{\theta}_1}(x)}\right]\frac{f_{\boldsymbol{\theta}_2}(x)}{f_{\boldsymbol{\theta}_1}(x)}\right\} e_i(\boldsymbol{\theta}_1)e_j(\boldsymbol{\theta}_1)e_k(\boldsymbol{\theta}_2)\right]\\ &\qquad - E_{\boldsymbol{\theta}_2}\left\{w''\left[\frac{f_{\boldsymbol{\theta}_2}(x)}{f_{\boldsymbol{\theta}_1}(x)}\right]\frac{f_{\boldsymbol{\theta}_2}(x)}{f_{\boldsymbol{\theta}_1}(x)}\,\partial_i e_j(\boldsymbol{\theta}_1)e_k(\boldsymbol{\theta}_2)\right\}.\end{aligned}$$

By substituting $\boldsymbol{\theta}_1=\boldsymbol{\theta}$ and $\boldsymbol{\theta}_2=\boldsymbol{\theta}$ in (3.10) and (3.11), we can conclude that

$$g^{(\rho_w)} = g$$

and

$$\Gamma^{(\rho_w)} = (2+w'''(1))\overset{e}{\Gamma} + (-1-w'''(1))\overset{m}{\Gamma}$$

on $\mathscr{F}$. This completes the proof.

Theorem 3.3 shows the every contrast function ρ_w has the α-conjugate metric structure $\mathscr{C}_\alpha$ with $\alpha=\alpha_w$. Amari pointed out through personal communication that a contrast functional of W-type is invariant under one-to-one transformations on the sample space. The invariance between ρ_w and $\mathscr{C}_\alpha$ may lead us to (3.9). Indeed consider

$$\rho(f, g) = \frac{1}{2}\int (f-g)^2 d\mu,$$

which is often adopted as a measurement of optimality for density estimators. The conjugate metric structure $\mathscr{C}(\rho)$ is given in the following way:

$$g_{ij}^{(\rho)} = \int \partial_i f_\theta \partial_j f_\theta d\mu$$

and

$$\Gamma_{ij,k}^{(\rho)} = {}^*\Gamma_{ij,k}^{(\rho)} = \int \partial_i \partial_j f_\theta \partial_k f_\theta d\mu.$$

Thus neither the functional ρ nor $\mathscr{C}(\rho)$ is invariant under transformations on the sample space, which may come from non-existence of $\mathscr{C}(\rho)$ in $\{\mathscr{C}_\alpha\}$.

Example 3.1. Applying theorem 3.2 to the classical contrast functions in Example 2.1, we have the following structures:

(1) *the Kullback–Leibler information*:

$$\mathscr{C}(\rho_{KL}) = \mathscr{C}_S.$$

(2) *the squared Hellinger distance*:

$$\mathscr{C}(H^2) = \mathscr{C}_\alpha \quad (\alpha=0).$$

(3) *the Jeffreys divergence*:

$$\mathscr{C}(\rho_J) = \mathscr{C}_\alpha \quad (\alpha=0).$$

(4) *the Chernoff information of order* α:

$$\mathscr{C}(\rho_\alpha) = \mathscr{C}_\alpha$$

(5) *the exponential divergence*:

$$\mathscr{C}(\rho_e) = \mathscr{C}_\alpha \quad (\alpha = -3)$$

(6) *the Kagan divergence*:

$$\mathscr{C}(\rho_{x^2}) = \mathscr{C}_\alpha \quad (\alpha = 3)$$

(7) *the product divergence with* (α, β)*-index*:

$$\mathscr{C}(\rho_{\alpha\beta}) = \mathscr{C}_\gamma \quad \left(\gamma = \frac{3}{2}(\alpha+\beta)\right).$$

Recall the operations *, $\oplus$ and $\ominus$ on $\mathscr{W}_1$ defined in Section 2. We note that

$$\alpha_{w^*} = -\alpha_w,$$

$$\alpha_{w\oplus} = \alpha_w + 2$$

and

$$\alpha_{w\ominus} = \alpha_w - 2.$$

If ρ_w is symmetric, then $\alpha_w = 0$ on account of Theorem 2.2. That is, the conjugacy of $\mathscr{C}(\rho_w)$ becomes trivial.

We have investigated the restricted version of a contrast functional on $\mathscr{F}_0$. However a contrast function ρ may be defined only on $\mathscr{F}$ when we are concerned with $\mathscr{F}$. On the basis of this idea, a class of contrast functions only on $\mathscr{F}$ is introduced. The application to statistical estimation is reported in Eguchi [27]. We shall note that the conjugate metric structures induced by this class of contrast functions are fairly different from the α-conjugate metric structures.

Let ϕ be a C^3-diffeomorphism of $\boldsymbol{\theta}$ into $\boldsymbol{\tau}$. We define a contrast function

$$\rho_\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{1}{2}(\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2)' G(\boldsymbol{\tau}_1)(\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2)$$

on $\mathscr{F}$ with respect to $\boldsymbol{\tau}$, where $G(\boldsymbol{\tau})$ is the matrix that consists of the components of g and $\boldsymbol{\tau}_p = \phi^{-1}(\boldsymbol{\theta}_p)$ with $p = 1, 2$. The function ρ_ϕ can be considered as an approximation of ρ with $g^{(\rho)} = g$ on $\mathscr{F}$:

$$\rho(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \approx \rho_\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$$

with respect to the coordinate system $\boldsymbol{\tau}$. Note that

$$\rho_\phi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \rho_\psi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$$

if the mapping $\phi \circ \psi^{-1}$ is an affine transformation. Thus there is a one-to-one correspondance between the class of all non-flat transformations of coordinates and the class of all contrast functions ρ_τ.

Mahalanobis [43] introduced the squared distance

$$\frac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)'\,\Sigma_0^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

in the normal family with known covariance Σ_0, which is called Mahalanobis' D^2. The contrast function ρ_ϕ can be considered as an extension of D^2 to the family $\mathscr{F}$, by which reason we call ρ_ϕ of Mahalanobis-type, or of M-type for short. Similarly as in the case of the contrast functions of M-type, a straightforward calculation leads us to the following theorem.

THEOREM 3.4. *Let ρ_ϕ be a contrast function of M-type with respect to ϕ: $\boldsymbol{\theta}\to\boldsymbol{\tau}$. Then the Riemmanian metric due to ρ_ϕ coincides with the information metric g. The pair of affine connections Γ^M and $^*\Gamma^M$ has the following coefficients*

$$\Gamma^M_{ij,k}(\boldsymbol{\tau}) = \partial_i g_{jk}(\boldsymbol{\tau}) + \partial_j g_{ki}(\boldsymbol{\tau})$$

and

$$^*\Gamma^M_{ij,k}(\boldsymbol{\tau}) = -\partial_k g_{ij}(\boldsymbol{\tau})$$

with respect to $\boldsymbol{\tau}$.

Note that the pair of Γ^M and $^*\Gamma^M$ decomposes Christofell's three indices

$$\frac{1}{2}(\partial_i g_{jk}+\partial_j g_{ki}) - \frac{1}{2}\partial_k g_{ij}$$

into the first and second terms, respectively. The contrast function ρ_τ of M-type deeply depends on the parameter $\boldsymbol{\tau}$.

Example 3.2. We consider a mixture model of countably many distributions:

$$\mathscr{F}_m = \{f_\theta(\boldsymbol{x}) = \textstyle\sum_{i=1}^\infty \theta_i f_i(\boldsymbol{x}) : \sum_{i=1}^\infty \theta^i = 1,\ \theta^i > 0 \quad (i=1, 2,\ldots)\},$$

where f_i is in $\mathscr{F}_0$ with $i=1, 2,\ldots$. The contrast function of M-type with the mixture parameter $\boldsymbol{\theta}=(\theta^1, \theta^2,\ldots)$ is given as

$$\rho_\theta(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{1}{2}\sum_{j=1}^\infty \sum_{i=1}^\infty (\theta_1^i-\theta_2^i)(\theta_1^j-\theta_2^j) E_{\boldsymbol{\theta}_1}\{e_i(\boldsymbol{\theta}_1)e_j(\boldsymbol{\theta}_1)\}$$

$$= \int \frac{(f_{\boldsymbol{\theta}_1}-f_{\boldsymbol{\theta}_2})^2}{f_{\boldsymbol{\theta}_1}}\,d\mu,$$

which is nothing but the Kagan divergence in Example 2.1.

Example 3.3. Let $\mathscr{F}_e$ be an exponential family, *i.e.*,

$$\mathscr{F}_e = \{f_\theta(\boldsymbol{x}) = e^{\boldsymbol{x}'\boldsymbol{\theta}-\psi(\boldsymbol{\theta})} : \boldsymbol{\theta}\in\Theta\}$$

The expectation parameter $\boldsymbol{\eta}$ is often used by defining a transformation ϕ of $\boldsymbol{\theta}$ into $\boldsymbol{\eta}$ by

$$\phi(\boldsymbol{\theta}) = \boldsymbol{E}_{\boldsymbol{\theta}}\boldsymbol{x}.$$

Then the contrast functions of M-type with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are

$$\rho_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{1}{2}(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)'G(\boldsymbol{\theta}_1)(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)$$

and

$$\rho_{\boldsymbol{\eta}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{1}{2}(\boldsymbol{\eta}_1-\boldsymbol{\eta}_2)'G^{-1}(\boldsymbol{\theta}_1)(\boldsymbol{\eta}_1-\boldsymbol{\eta}_2),$$

respectively, where $G(\boldsymbol{\theta})=[\partial^2/(\partial\theta^i\partial\theta^j)\psi(\boldsymbol{\theta})_{i,j}]$. By a straightforward calculation it holds that

$$\mathscr{C}(\rho_{\boldsymbol{\theta}}) = \mathscr{C}_{\alpha} \quad (\alpha=-3)$$

and

$$\mathscr{C}(\rho_{\boldsymbol{\eta}}) = \mathscr{C}_{\alpha} \quad (\alpha=3)$$

on $\mathscr{F}_e$. Thus we have the relation

$$\mathscr{C}(\rho^*_{\boldsymbol{\eta}}) = \mathscr{C}(\rho_{\boldsymbol{\theta}})$$

on $\mathscr{F}_e$ with $\rho^*_{\boldsymbol{\eta}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)=\rho_{\boldsymbol{\eta}}(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1)$. For the case of the multinormal family $\mathscr{G}$ with means $\mathbf{0}$:

$$\mathscr{G} = \left\{f_{\boldsymbol{\theta}}(\boldsymbol{x})=\exp\left[-\frac{1}{2}\operatorname{tr}(\boldsymbol{x}\boldsymbol{x}'\Sigma^{-1})-\frac{1}{2}\log(\det\Sigma)\right]\right\}$$

with $\boldsymbol{\theta}=(\sigma^{ij})_{i\geqq j}$ for $\Sigma^{-1}=(\sigma^{ij})_{i,j}$, it holds that

$$\rho_{\boldsymbol{\eta}}(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1)=\rho_{\boldsymbol{\theta}}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = \frac{1}{2}\operatorname{tr}\{(\Sigma_1-\Sigma_2)\Sigma_1^{-1}\}^2$$

which is often called the generalized least squares function (cf. Browne [15]). Thus we have $\rho_{\boldsymbol{\eta}}=\rho^*_{\boldsymbol{\theta}}$ on $\mathscr{G}_0$, which means the same duality as $\rho_{\alpha}=\rho^*_{-\alpha}$ on $\mathscr{G}_0$ with the Chernoff divergence of order α in Example 2.1. By a linear transformation

$$S(\boldsymbol{x}) = B\boldsymbol{x}$$

with a nonsingular matrix B, the family $\mathscr{G}_0$ is transformed into

$$\bar{\mathscr{G}}_0 = \{f_{\bar{\theta}}(\boldsymbol{x})\colon \bar{\Sigma}=B\Sigma B'\}$$

We note that this transformation keeps the contrast function invariant, *i.e.*,

$$\rho_{\bar{\theta}}(\bar{\theta}_1, \bar{\theta}_2) = \frac{1}{2}\mathrm{tr}\{(\bar{\Sigma}_1 - \bar{\Sigma}_2)\,\bar{\Sigma}_1^{-1}\}^2 = \rho_{\theta}(\theta_1, \theta_2)\,.$$

It is seen form Examples 3.2 and 3.3 that a contrast function ρ_τ of M-type has the α-conjugate metric structure if the parameter τ is significant in a statistical sense.

4. Geometry generated by a contrast functional: infinite dimensional cases

Our interests have been so far limited to a finite parametric subfamily $\mathscr{F}$ of $\mathscr{F}_0$, where $\mathscr{F}_0$ is the space of all probability measures with a common support. It may be noted that many contrast functionals can be defined over $\mathscr{F}_0$ as in Example 2.1. If we can directly treat the space $\mathscr{F}_0$ itself as a geometric object, then we may expand a perspective formulation in various fields of statistics, *e.g.*, nonparametric inference and robust method, etc. However, we do not see any theory on infinite dimensional manifolds which is applicable to such fields. As the first step for proceeding a differential geometric approach to such fileds, we attempt to extend the conjugate metric structures on $\mathscr{F}$ to $\mathscr{F}_0$.

Nagaoka and Amari [45] derived a curve $\overset{\alpha}{C}$ connecting f and g in $\mathscr{F}_0$:

$$\overset{\alpha}{C} = \{\widehat{fg}_t^{\alpha} = c_\alpha(t)((1-t)f^{\frac{1+\alpha}{2}} + tg^{\frac{1+\alpha}{2}})^{\frac{2}{1+\alpha}} : 0 \le t \le 1\}$$

for α, $-1 \le \alpha \le 1$, where

$$c_\alpha(t) = \left[\int \{(1-t)f^{\frac{1+\alpha}{2}} + tg^{\frac{1+\alpha}{2}})\}^{\frac{2}{1+\alpha}} d\mu\right]^{-1}.$$

The curve $\overset{\alpha}{C}$ satisfies the α-geodesic equation

$$\ddot{\ell}_t + \frac{1-\alpha}{2}\dot{\ell}^2 + \frac{1+\alpha}{2}E_t\dot{\ell}^2 = 0$$

where $\ell_t = \log(\widehat{fg}_t^{\alpha})$ and E_t denotes the expectation with respect to the density $\widehat{fg}_t^{\alpha}$. Thus the curve $\overset{\alpha}{C}$ is called the α-geodesic curve. Note that the curves $\overset{\alpha}{C}$ with $\alpha = 1$ and -1 are reduced to one parameter mixture and exponential families given by

$$\overset{m}{C} = \{(1-t)f + tg : 0 \le t \le 1\}$$

and

$$\overset{e}{C} = \{c_{-1}(t) f e^{t\log(g/f)} : 0 \le t \le 1\}$$

respectively.

Dawid [20] proposed a tangent space of $\mathscr{F}_0$ at f as

$$T_f(\mathscr{F}_0) = \{s \in L^2(f) : E_f\{s(x)\} = 0\},$$

where $L^2(f)$ denotes the space of squared intergrable functions with respect to f. The metric on $T_f(\mathscr{F}_0)$ was defined as the L^2-metric

$$g_f(t, s) = E_f\{t(x)s(x)\},$$

which is reduced to the information metric if the model is restricted to a finite parametric case.

In order to proceed further with the differential geometry on $\mathscr{F}_0$, we introduce the *α-representation* of the statistical conjugate structure $\mathscr{C}_S=(g, \overset{m}{\Gamma}, \overset{e}{\Gamma})$ over $\mathscr{F}_0$.

The natural basis of $T_f(\mathscr{F}_0)$ can be made in the following expression:

$$e_g^\alpha(f) = \lim_{t\to 0} \frac{1}{t}\{\log(\widehat{fg_t^\alpha}) - \log f\} \tag{4.1}$$

$$= -\left[\frac{2}{1+\alpha}\left\{\left(\frac{f}{g}\right)^{\frac{1+\alpha}{2}} - 1\right\} + \frac{1-\alpha}{2}\rho_\alpha(f, g)\right]$$

for g in $\mathscr{F}_0$, where ρ_α denotes the Chernoff information of order α in Example 2.1. In particular we have the mixture and the exponential expressions given by

$$e_g^m(f) = \frac{f-g}{g}$$

and

$$e_g^e(f) = \log(g/f) + \rho_{KL}(f, g),$$

respectively. The transformation of $e_g^\alpha(f)$ into $e_g^{\alpha'}(f)$ is given as

$$e_g^{\alpha'}(f) = -\frac{2}{1+\alpha'}\left[\left\{-\frac{1+\alpha}{2}\left(e_g^\alpha(f) - \frac{1-\alpha}{2}\rho_\alpha(f, g)\right)+1\right\}^{\frac{1+\alpha'}{1+\alpha}} - 1\right]$$

$$+ \frac{1-\alpha'}{2}\rho_{\alpha'}(f, g).$$

In particular,

$$e_g^e(f) = \log(e_g^m(f)+1) + \rho_{KL}(f, g).$$

The Gâteau derivative of $e_p^\alpha(f)$ along the α-geodesic curve connecting f with q is defined by

$$\partial_q e_p^\alpha(f) = \left(\frac{\partial^2}{\partial t \partial s}\log f_{t,s}\right)_{t=s=0}$$

which can be expressed as

$$\partial_q e_p^\alpha(f) = -\frac{1-\alpha}{2}e_p^\alpha(f)e_q^\alpha(f) - \frac{1+\alpha}{2}g_{pq}(f), \tag{4.2}$$

where

$$f_{t,s} = \{(1-t-s)f^{\frac{1-\alpha}{2}} - tp^{\frac{1-\alpha}{2}} - sq^{\frac{1-\alpha}{2}}\}^{\frac{2}{1-\alpha}} \Big/ \left\{\int \{(1-t-s)f^{\frac{1-\alpha}{2}} - tp^{\frac{1-\alpha}{2}} - sq^{\frac{1-\alpha}{2}}\}^{\frac{2}{1-\alpha}} d\mu\right\}$$

and p and q are sufficiently close to f with respect to the Hellinger topology. These formulas (4.1) and (4.2) generate an extension of the statistical conjugate metric structure

$$\mathscr{C}_S = (g, \overset{m}{\Gamma}, \overset{e}{\Gamma})$$

on $\mathscr{F}$ to the infinite dimensional space $\mathscr{F}_0$ as follows: The information metric g has the *α-representation*

$$g_{pq}(f) = \left(\frac{2}{1-\alpha}\right)^2 E_f\left[\left\{\left(\frac{p}{f}\right)^{\frac{1-\alpha}{2}} - 1\right\}\left\{\left(\frac{q}{f}\right)^{\frac{1-\alpha}{2}} - 1\right\}\right] - \left(\frac{1+\alpha}{2}\right)^2 \rho_{-\alpha}(f, p)\rho_{-\alpha}(f, q) \tag{4.3}$$

Furthermore the mixture connection and the exponential connection have *the α-representations*

$$\overset{m}{\Gamma}_{pq,r}(f) = \frac{1+\alpha}{2} T_{pqr}(f)$$

and

$$\overset{e}{\Gamma}_{pq,r}(f) = -\frac{1-\alpha}{2} T_{pqr}(f),$$

respectively, where

$$T_{pqr}(f) = E_f\{e_p^\alpha(f) e_q^\alpha(f) e_r^\alpha(f)\}.$$

Accordingly the α-representation of the α′-connection is

$$\overset{\alpha'}{\Gamma}_{pqr}(f) = \frac{\alpha-\alpha'}{2} T_{pqr}(f).$$

Note that the α-representation of α-connection vanishes over $\mathscr{F}_0$.

Recalling the contrast functionals ρ_{χ^2} and ρ_e in Example 2.1, these have the following correspondance with the α-representation: For the α-representation $g_{pq}(f)$ defined in (4.3),

$$\rho_{\chi^2}(p, q) = \lim_{\alpha \to -1} g_{p,q}(q)$$

and

$$\rho_e(p, q) = \lim_{\alpha\to 1} g_{p,q}(q) + \{\rho_{KL}(p, q)\}^2.$$

We similarly define the conjugate metric structure generated by a contrast functional in terms of the α-geodesic: The metric $g^{(\rho)}$ is

$$g^{(\rho)}_{pq}(f) = (\varepsilon^{\alpha}_p \varepsilon^{\alpha}_q \rho(f_1, f_2))_{f_1=f_2=f}$$

and the conjugate pair of the affine connections is

$$\Gamma^{(\rho)}_{pq,r}(f) = (-\varepsilon^{\alpha}_p \varepsilon^{\alpha}_q \delta^{\alpha}_r \rho(f_1, f_2))_{f_1=f_2=f}$$

and

$$^*\Gamma^{(\rho)}_{pq,r}(f) = (-\delta^{\alpha}_p \delta^{\alpha}_q \varepsilon^{\alpha}_r \rho(f_1, f_2))_{f_1=f_2=f},$$

where ε^{α}_p and δ^{α}_p denote the Gâteau differentials along the α-geodesic curve connecting f with p with respect to f_1 and f_2, respectively.

Let ρ_w be a contrast functional of W-type. We suppose the following condition for ρ_w:

A–4. *The integral*

$$\rho_w(f_1, f_2) = \int w\left(\frac{f_2}{f_1}\right) f_1 d\mu$$

is twice Gâteau-dereiffntiable with respect to f_1 and f_2 under the integral sign for any f_1 and f_2 with a sufficiently small Hellinger distance between f_1 and f_2.

We give a formal extension of Theorem 3.3 to infinite dimensional cases.

THEOREM 4.1. *Under the assumption* A–4, *we have*

$$\mathscr{C}(\rho_w) = \mathscr{C}_{\alpha_w}$$

with $\alpha_w = 3 + 2w'''(1)$.

PROOF. We write

$$f_{t,u} = \{(1-t-u)f^{\frac{1+\alpha}{2}} + tp^{\frac{1+\alpha}{2}} + up^{\frac{1+\alpha}{2}}\}^{\frac{2}{1+\alpha}} \Big/ \left[\int \{(1-t-u)f^{\frac{1+\alpha}{2}} + tp^{\frac{1+\alpha}{2}} + uq^{\frac{1+\alpha}{2}}\}^{\frac{2}{1+\alpha}} d\mu\right]$$

and

$$g_s = \{(1-s)f^{\frac{1+\alpha}{2}} + sr^{\frac{1+\alpha}{2}}\}^{\frac{2}{1+\alpha}} \Big/ \left\{\int \{(1-s)f^{\frac{1+\alpha}{2}} + sr^{\frac{1+\alpha}{2}}\}^{\frac{2}{1+\alpha}} d\mu\right\}.$$

Then it follows from A–4 that

$$\frac{\partial^2}{\partial s \partial t} \rho (f_{t,u}, g_s) = - \int w'' \left(\frac{g_s}{f_{t,u}} \right) \frac{g_s}{f_{t,u}} \frac{\dot{f}_{t,u}}{f_{t,u}} \dot{g}_s d\mu,$$

and

$$\frac{\partial^3}{\partial u \partial s \partial t} \rho (f_{t,u}, g_s) = \int w''' \left(\frac{g_s}{f_{t,u}} \right) \left[\frac{g_s}{f_{t,u}} \right]^2 \left(\frac{\dot{f}_{t,u}}{f_{t,u}} \right)^2 \dot{g}_s d\mu$$
$$+ \int w'' \left(\frac{g_s}{f_{t,u}} \right) \frac{g_s}{f_{t,u}} \left(\frac{\dot{f}_{t,u}}{f_{t,u}} \right)^2 \dot{g}_s d\mu - \int w'' \left(\frac{g_s}{f_{t,u}} \right) \frac{g_s}{f_{t,u}} \left\{ \frac{\partial}{\partial u} \left(\frac{\dot{f}_{t,u}}{f_{t,u}} \right) \right\} \dot{g}_s d\mu .$$

By an argument similar to the proof of Theorem 3.2, we can conclude

$$g^{(\rho_w)} = g$$

and

$$\Gamma^{(\rho_w)} = (2 + w'''(1)) \overset{e}{\Gamma} + (-1 - w'''(1)) \overset{m}{\Gamma}$$

on $\mathscr{F}_0$. This completes the proof.

Part II. Applications to statistical estimation

5. Classification of estimation methods

Fisher [28] presented a notion of efficiency in terms of information loss. Many statisticians have contributed to the classification of estimation methods on the basis of this notion and established a field of asymptotic theory in statistical inference (cf. Chernoff [18], Rao [48], [49], [50], Bahadule [8], Ghosh and Subramanyam [29], Efron [21], Hosoya [32], Amari [2], Phanzagl [46] and Akahira and Takeuchi [1]). In particular, we discuss the classification of estimation methods in terms of limiting information loss, which consists of three classes having Fisher-consistency, first order efficiency and second order efficiency, respectively.

We, along this stream, investigate estimation methods based on contrast functions, which will be called the minimum contrast methods or estimators. In this section we present a classification of contast functions on the basis of the conjugate metric structures established in Section 3. First the results in Eguchi [23] are reviewed in the light of the classification of contrast functions. Next we show that every method of estimation based on a contrast functinal of *W-type* becomes second order efficient by using the operations $*$, $\oplus$ and $\ominus$ developed in Section 2. Finally we consider a characterization of three classes of estimation methods.

We now look at another view of estimation in the light of the idea "*summary*"

introduced by Efron [22]. Let $\mathscr{F}_0$ be the family of all probability density functions with a common support on a sample space $\mathscr{X}$. We are concerned with a prametric subfamily $\mathscr{F}$ of $\mathscr{F}_0$ with the information metric g on $\mathscr{F}$. If a random vector $\boldsymbol{x}$ has a density f in $\mathscr{F}$, then $\boldsymbol{x}$ is said to have the information g. Suppose that a random sample $(\boldsymbol{x}_1, \boldsymbol{x}_2,..., \boldsymbol{x}_N)$ from a density f in $\mathscr{F}$ is given. Then the sample has the information Ng.

Let $t_N=t(\boldsymbol{x}_1, \boldsymbol{x}_2,..., \boldsymbol{x}_N)$ be a statistic with the information $g(t_N)$, i.e., the information metric on the family of induced densities f^{t_N} of $\prod_{i=1}^N f(\boldsymbol{x}_i)$, where f is in $\mathscr{F}$. We call the tensor

$$\delta_{t_N} = Ng - g(t_N)$$

the information loss in reducing from the sample to the statistics t_N. The following property is well-known, see e.g. §5a.4 in Rao [51]:

Proposition. Let t_N be a statistic with sample size N. Then the information loss due to t_N is nonnegative, *i.e.*,

$$\delta_{t_N}(A, A) \geqq 0$$

for any tangent vector A at f, where f denotes the true density.

A mapping $\hat{f}_N$: $\prod_{i=1}^N \mathscr{X} \to \mathscr{F}_0$ is called *the maximum likelihood summary* if

$$\sum_{i=1}^N \log \hat{f}_N(\boldsymbol{x}_i) = \max_{f \in \mathscr{F}_0} \sum_{i=1}^N \log f(\boldsymbol{x}_i)$$

for the sample $(\boldsymbol{x}_1, \boldsymbol{x}_2,..., \boldsymbol{x}_N)$. Note that $\hat{f}_N$ is often called a nonparametric maximum likelihood estimator.

Henceforth we restrict estimators of the true density f to mappings of $\hat{f}_N$. Thus an estimator T: $\mathscr{F}_0 \to \mathscr{F}$ is said to be Fisher-consistent if T is idempotent, *i.e.*,

$$T^2(f) = T(f)$$

for every f in $\mathscr{F}_0$. Let ρ be a contrast functional on $\mathscr{F}_0$. We call T_ρ *the minimum contrast estimator by* ρ if there exists a tubular neighbourhood $\mathscr{N}(\mathscr{F})$ of $\mathscr{F}$ in $\mathscr{F}_0$ such that

$$\rho[f, T_\rho(f)] = \min_{g \in \mathscr{F}} \rho(f, g)$$

for every f in $\mathscr{N}(\mathscr{F})$. Every minimum contrast estimator is Fisher-consistent since $\rho[T_\rho(f), g]=0$ if and only if $g=T_\rho(f)$. Beran [12] discussed the minimum Hellinger distance estimator, *i.e.*, T_ρ with $\rho=H^2$ according to our terminology, in the light of robustness.

A Fisher-consistent estimator $T_N=T(\hat{f}_N)$ is said to be optimal if the information loss due to T_N uniformly vanishes for each sample size N. The following fact is well-known under mild conditions:

If the model $\mathscr{F}$ is an exponential family, then the maximum likelihood estimator is optimal. In other words, the likelihood principle is compatible with the sufficiency principle under the exponential family. However, to our regret, it is impossible to obtain a uniform solution for the optimality of estimation under a general regular parametric family. Therefore we restrict our situation to the following one:

The model $\mathscr{F}$ is assumed to be an (n, m)-curved exponential family. Moreover a measurement of the optimality is adopted the limiting information loss

$$\delta_T = \lim_{N\to\infty} \delta_{T(\hat{f}_N)}$$

in place of $\delta_{T(\hat{f}_N)}$ with sample size N.

On the basis of this changing, let us classify the class $\mathscr{E}$ of all estimators as follows:

$$\mathscr{E}_0 = \{T\in\mathscr{E}: T \text{ is idempotent on } \mathscr{F}_0\},$$

$$\mathscr{E}_1 = \{T\in\mathscr{E}_0: \lim_{N\to\infty} N^{-1}\delta_{T(\hat{f}_N)} = 0\}$$

and

$$\mathscr{E}_2 = \{T\in\mathscr{E}_1: \delta_T \leqq \delta_S \text{ for every } S \text{ in } \mathscr{E}_1\}.$$

The subclasses $\mathscr{E}_0$, $\mathscr{E}_1$ and $\mathscr{E}_2$ of $\mathscr{E}$ are called the Fisher-consistent, the first order efficient and the second order efficient classes, respectively.

The following theorem has been established through a half of a centrury by Fisher [28], Rao [48] [49] [50], Ghosh and Sabramanyam [29], and Efron [21]. Finally Amari [2] has presented an elegant form in terms of differential geometric terminology, which we review in Section 1.

THEOREM 5.1 *Let T be in $\mathscr{E}_1$. Then the limiting information loss due to T is decomposed into*

$$\delta_T = \langle \overset{e}{H}(\mathscr{F})\rangle^2 + \langle \overset{m}{H}(A_T)\rangle^2,$$

where

$$A_T = \{g\in\mathscr{F}_e: T(g) = f\}$$

In particular the maximum likelihood estimator belongs to $\mathscr{E}_2$.

We note that the term $\langle \overset{e}{H}(\mathscr{F})\rangle^2$ is independent of the estimator T. Therefore we call $\langle \overset{e}{H}(\mathscr{F})\rangle^2$ the model part of δ_T and call $\langle \overset{m}{H}(A_T)\rangle^2$ the estimation part of δ_T.

Eguchi [23] investigated which class a minimum contrast estimator T_ρ belongs to. Recall the conjugate metric structure generated by ρ:

$$\mathscr{C}(\rho) = (g^{(\rho)}, \Gamma^{(\rho)}, {}^*\Gamma^{(\rho)})$$

on $\mathscr{F}_e$, which is defined in Section 3. Let $\mathscr{P}_0$ be the class of all contrast functions over $\mathscr{F}_e$ and

$$\mathscr{P}_1 = \{\rho \in P_0 : g^{(\rho)} = \varepsilon g \quad \text{on} \quad \mathscr{F}_e\},$$
$$\mathscr{P}_2 = \{\rho \in P_1 : \mathscr{C}(\rho) = \mathscr{C}_S \quad \text{on} \quad \mathscr{F}_e\}.$$

Then we have the following correspondance of these classes with the classes $\mathscr{E}_0$, $\mathscr{E}_1$, and $\mathscr{E}_2$.

THEOREM 5.2 *It holds that T_ρ belongs to $\mathscr{E}_k$ if ρ belongs to $\mathscr{P}_k$ for $k=0$, 1 and 2. In particular, for a contrast function ρ_w of W-type, T_{ρ_w} always belongs to $\mathscr{E}_1$ and it holds*

$$\delta(T_{\rho_w}) = \langle \overset{e}{H}(\mathscr{F})\rangle^2 + \{2+w'''(1)\}^2\langle T\rangle^2. \tag{5.1}$$

It follows from Theorem 5.2 that T_{ρ_w} belongs to $\mathscr{E}_2$ if and only if $w_1'''(1) = -2$. Accordingly we define a mapping on $\mathscr{W}_1$ as

$$\kappa(w) = \frac{3}{8}(w+w^*)^{\oplus} + \frac{1}{8}(w+w^*)^{\ominus},$$

where the operations $*$, $\oplus$ and $\ominus$ on $\mathscr{W}_1$ are defined in Section 2. Then we have the following theorem.

THEOREM 5.3 *Setting up the above, the minimum contrast estimator by $\rho_{\kappa(w)}$ belongs to $\mathscr{E}_2$ for every w in $\mathscr{W}_1$.*

PROOF is easily seen from the fact that $(\kappa(w))'''(1) = -2$ for every w in $\mathscr{W}_1$.

In addition to this result, it follows that the product divergence $\rho_{\alpha\beta}$ defined in Section 2 is in $\mathscr{P}_2$ if $\alpha+\beta = -2/3$. Thus in this case the minimum contrast estimator by $\rho_{\alpha\beta}$ is in $\mathscr{E}_2$.

Returning the subject to the classes of estimators, let us consider the relations between the classes of estimators and the classes of contrast functions.

THEOREM 5.3. *Let $\hat{f}$ be the maximum likelihood summary of the sample $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)$ from a density f in an (n, m)-curved exponential family $\mathscr{F}$. Then*

$$\lim_{N\to\infty} NE\{\rho[f, T_\rho(\hat{f})]\} \leqq \frac{m}{2} \tag{5.2}$$

for any ρ in $\mathscr{F}_0$.

PROOF. Take a parametrization

$$\mathscr{F} = \{f_{\beta(\theta)}(\boldsymbol{x}) = e^{\boldsymbol{x}'\beta(\theta)-\psi[\beta(\theta)]} : \boldsymbol{\theta}\varepsilon\Theta\}$$

by a parameter space Θ in $\boldsymbol{R}^m$. We write

$$\rho(\boldsymbol{\beta}, \boldsymbol{\theta}) = \rho[f_{\boldsymbol{\beta}}, f_{\boldsymbol{\beta}(\boldsymbol{\theta})}]$$

and

$$\rho(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}}) = \rho[\hat{f}, T_\rho(\hat{f})]$$

with $\hat{f}=f_{\hat{\beta}}$. Since $\hat{\boldsymbol{\theta}}$ is a minimizer of $\rho(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$,

$$(\partial_a \rho(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}))_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0 \tag{5.3}$$

for $a=1, 2,\ldots, m$ with $\partial_a=\partial/\partial\theta^a$. By the first order expansion of (5.3), we have

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = (B'G^{(\rho)}B)^{-1}BG^{(\rho)}\{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}(\boldsymbol{\theta})\},$$

where the true density f is expressed as $f_{\boldsymbol{\beta}(\boldsymbol{\theta})}$ and $B=\partial\boldsymbol{\beta}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and $G^{(\rho)}$ denotes the matrix composed of the components of the metric $g^{(\rho)}$. From the first approaximation

$$\rho[f, T_\rho(\hat{f})] = \frac{1}{2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})'G^{(\rho)}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}),$$

it follows that

$$\lim_{N\to\infty} N\boldsymbol{E}\{\rho[f, T_\rho(\hat{f})]\} = \frac{1}{2}\operatorname{tr}\{(B'G^{(\rho)}B)^{-1}B'G^{(\rho)}G^{-1}G^{(\rho)}B\},$$

since the limiting distribution of $\sqrt{N}\{\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}(\boldsymbol{\theta})\}$ follows the Gaussian law with mean $\mathbf{0}$ and covariance G^{-1}, where G denotes the Fisher information matrix of $\boldsymbol{\beta}$. By Cramer-Rao's inequality, we have

$$(B'GB)^{-1} \leqq (B'G^{(\rho)}B)^{-1}B'G^{(\rho)}G^{-1}G^{(\rho)}B(B'G^{(\rho)}B)^{-1},$$

which implies the inequality (5.2). This completes the proof.

Eguchi [25] gave a characterization of classes $\mathscr{E}_1$ and $\mathscr{E}_2$, which is given in the following theorem.

THEOREM 5.4. *The following statements are equivalent to each other*:

(i) *An estimator $T(\hat{f})$ belongs to $\mathscr{F}_1$.*

(ii) $\lim_{N\to\infty} N\boldsymbol{E}\{\rho[f, T(\hat{f})]\}=\dfrac{m}{2}$.

(iii) $\lim_{N\to\infty} N\boldsymbol{E}\{\rho[f, T(\hat{f})]\} = \dfrac{1}{2}(n-m)$.

Furthermore the estimator $T(\hat{f})$ belongs to $\mathscr{E}_2$ if and only if

$$\lim_{N\to\infty} NE\{\rho[\hat{f}, S(\hat{f})] - \rho[\hat{f}, T(\hat{f})]\} \geqq 0 \tag{5.4}$$

for each $S(f)$ in $\mathscr{E}_1$ and each ρ in $\mathscr{P}_2$.

We can conclude that the class $\mathscr{P}_2$ completely discriminates the class $\mathscr{E}_2$ from the class $\mathscr{E}_1$ in the sense of (5.4). Thus Theorems 5.2 and 5.3 lead the equivalence between classifications $\mathscr{P}_0 \supset \mathscr{P}_1 \supset \mathscr{P}_2$ and $\mathscr{E}_0 \supset \mathscr{E}_1 \supset \mathscr{E}_2$.

6. Covariance structure model

Let $\mathscr{G}$ be the space of all Gaussian measures on $\boldsymbol{R}^k$ with known mean vector and let $\mathscr{S}$ be the space of all positive definite matrices of order k. Then the space $\mathscr{G}$ has the density form of

$$\{f_\Sigma(x) = (2\pi)^{-\frac{1}{2}k}(\det \Sigma)^{-\frac{1}{2}} e^{-\frac{1}{2}x'\Sigma^{-1}x} : \Sigma \in \mathscr{S}\}.$$

as in Example 1.1. Skovgaard [57] intensively investigated the Riemannian geometric properties of $\mathscr{G}$. We consider a subfamily $\mathscr{F}$ of $\mathscr{G}$ which is specified by

$$\mathscr{F} = \{f_{\Sigma(\theta)} : \theta \in \Theta\} \tag{6.1}$$

with respect to Σ-coordinates of $\mathscr{G}$, where Θ denotes an open subset of $\boldsymbol{R}^n$ with $n < k(k+1)/2$. Here we call $\mathscr{F}$ a covariance struture model, which includes some important statistical models, *e.g.*, factor analysis model, linear covariance model and intraclass correlation model (*cf.* Anderson [6], Browne [15] and Shapiro [55]). We apply the results in Section 5 to this model. As a result, explicit formulas for the limiting information losses due to minimum contrast estimators are given in addition to their asymptotic mean square errors and biases.

Let S be the sample covariance matrix based on a sample with size N from a density of the model $\mathscr{F}$. Swain [58] introduced a class of contrast functions

$$\rho_v(f_S, f_\Sigma) = \rho_v(S, \Sigma) = \sum_{r=1}^k v(\lambda_r),$$

where λ_r's denote the eigenvalues of $S^{1/2}\Sigma^{-1}S^{1/2}$ and $v(\lambda)$ is a C^3-function on $(0, \infty)$ satisfying $v(1)=v'(1)=0$, $v''(1)=\frac{1}{2}$ and $v(\lambda)>0$ if $\lambda>0$, $\lambda\neq 1$. We call ρ_v *a contrast function of spectral type*, or *of S-type* for short. Note that $\rho_{v^\circ}(S, \Sigma) = \rho_v(\Sigma, S)$ with $v^\circ(\lambda)=v(\lambda^{-1})$. On the other hand, we have considered a contrast function of *W*-type

$$\rho_w(S, \Sigma) = \int w\left(\frac{f_\Sigma}{f_S}\right) f_S \, d\mu,$$

where the function w is in the space $\mathscr{W}$ defined in Section 2. The class of contrast functions of S-type is closely related with that of contrast functions of W-type. For example, the Kullback-Leibler information ρ_{KL} can be expressed as both ρ_v and ρ_w with $v(t)=\frac{1}{2}w(t)=\frac{1}{2}(-\log t+t-1)$. Further the Chernoff information of order α has both expressions of W-type and of S-type, i.e., $\rho_\alpha=\rho_{w_\alpha}$ with

$$w_\alpha(t) = \frac{4}{1-\alpha^2}(1-t^{\frac{1+\alpha}{2}}) + \frac{2}{1-\alpha}(t-1)$$

and $\tilde{\rho}_\alpha=\rho_{v_\alpha}$ with

$$v_\alpha(\lambda) = \frac{4}{1-\alpha^2}\left\{\log\left(\frac{1-\alpha}{2}\lambda+\frac{1+\alpha}{2}\right)-\frac{1-\alpha}{2}\log\lambda\right\},$$

where $\tilde{\rho}_\alpha$ is defined in (2.4).

Let ϕ be a parameter transformation of Σ into $\boldsymbol{\tau}$. With respect to the parameter $\boldsymbol{\tau}$, the contrast function of M-type is introduced by

$$\rho_\phi(S, \Sigma) = \frac{1}{2}(\boldsymbol{\tau}_1-\boldsymbol{\tau}_2)'G(\boldsymbol{\tau}_1)(\boldsymbol{\tau}_1-\boldsymbol{\tau}_2),$$

where $\boldsymbol{\tau}_1=\phi(S)$ and $\boldsymbol{\tau}_2=\phi(\Sigma)$ and $G(\boldsymbol{\tau}_1)$ denotes the Fisher information matrix of $\boldsymbol{\tau}_1$. For example, consider a transformation ϕ: $\mathscr{S}\to\mathscr{S}$ defined by

$$\phi(\Sigma) = \Sigma^{\frac{1}{2}}.$$

Then we have the contrat function of M-type:

$$\rho_\phi(S, \Sigma) = \mathrm{tr}\,\{S^{-\frac{1}{2}}(\Sigma^{\frac{1}{2}}-S^{\frac{1}{2}})\}^2.$$

This function ρ_ϕ can be also of S-type, i.e.,

$$\rho_\phi(S, \Sigma) = \sum_{r=1}^{k}(\sqrt{\lambda_r}-1)^2.$$

However there is generally no relation of inclusion among the classes of contrast functions of S-type, W-type and M-type. A common property is the invariance under linear transformations on $\boldsymbol{R}^k$, i.e.,

$$\rho(P'SP, P'\Sigma P) = \rho(S, \Sigma)$$

for every non-singular matrix P. In addition to this invariance, a contrast function of S-type has the following invariance:

THEOREM 6.1. *Let ρ_v be a contrast function of S-type. Then*

$$\rho_v(P_1SP_2, P_1\Sigma P_2) = \rho_v(S, \Sigma)$$

for every non-singular matricies P_1 and P_2.

PROOF. We write the eigenvalues of a square matrix M of order k as

$$\lambda_1(M) \leqq \lambda_2(M) \leqq \cdots \leqq \lambda_k(M).$$

Then the proof follows from

$$\lambda_r[(P_1 \Sigma P_2)^{-1}(P_1 S P_2)] = \lambda_r(S^{\frac{1}{2}} \Sigma^{-1} S^{\frac{1}{2}})$$

for $r=1, 2,\ldots, k$.

Let v be analytic on $(0, \infty)$. Define

$$\phi_v(S, \Sigma) = \sum_{k=1}^{\infty} a_k \{S(\Sigma^{-1} - S^{-1})\}^k$$

with $v(\lambda)=\sum_k a_k(\lambda-1)^k$. Then it holds

$$\rho_v(S, \Sigma) = \operatorname{tr} \phi_v(S, \Sigma), \tag{6.2}$$

for every S and Σ in $\mathscr{S}$. We note that $\phi_v(S, \Sigma)$ is postitive definte for every distinct S and Σ in $\mathscr{S}$. James [34] showed that the squared geodesic distance with repect to the information metric is given as $\rho_v(S, \Sigma)$ with $v(\lambda)=\frac{1}{2}(\log \lambda)^2$. By using the relation (6.2), the squared geodesic distance can be expressed as

$$\frac{1}{2} \operatorname{tr} \{\log (S^{\frac{1}{2}} \Sigma^{-1} S^{\frac{1}{2}})\}^2,$$

where log denotes the logarithmic mapping defined on $\mathscr{S}$.

The generalized least squares function

$$\frac{1}{2} \operatorname{tr} \{S^{-1}(\Sigma - S)\}^2$$

is often used in covariance structual models (cf. Anderson [6]). This function has the form ρ_v with $v(\lambda)=\frac{1}{4}(\lambda^{-1}-1)^2$ but can not be expressed as the form of W-type. In general, every contrast function of S-type can be easily calculated, while every contrast function of W-type is not so, *e.g.*, the Chernoff information of order α is generally integrable only for α, $|\alpha|<1$. Thus a contrast function of S-type is more applicable than that of W-type in the space $\mathscr{G}$. In addition to these aspects, we shall show that this function has the additivity property. Let f_S and f_Σ be in $\mathscr{G}$ with statistically independent marginals of the common size, that is,

$$f_S(\boldsymbol{x}_1, \boldsymbol{x}_2) = f_{S_1}(\boldsymbol{x}_1) f_{S_2}(\boldsymbol{x}_2)$$

and

$$f_{\Sigma}(x_1, x_2) = f_{\Sigma_1}(x_1) f_{\Sigma_2}(x_2)$$

where

$$S = \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}.$$

Then in the above setup, we state the following theorem.

THEOREM 6.2. *Let ρ_v be a contrast function of S-type. Then ρ_v has an additivity of information, i.e.,*

$$\rho_v(S, \Sigma) = \rho_v(S_1, \Sigma_1) + \rho_v(S_2, \Sigma_2). \tag{6.3}$$

PROOF. According to the notation in the proof of Theorem 6.1, it holds that

$$\{\lambda_r(S^{\frac{1}{2}} \Sigma^{-1} S^{\frac{1}{2}})\}_{r=1,\ldots,k} = \{\lambda_s (S_1 \Sigma_1^{-1} S_1^{\frac{1}{2}})\}_{s=1,\ldots,\ell},$$
$$\cup \{\lambda_t(S_2^{\frac{1}{2}} \Sigma_2^{-1} S_2^{\frac{1}{2}})\}_{t=1,2,\ldots k-\ell}$$

where ℓ denotes the order of $S_1^{\frac{1}{2}} \Sigma_1^{-1} S_1^{\frac{1}{2}}$. Therefore we have

$$\sum_{r=1}^{k} v[\lambda_r(S^{\frac{1}{2}} \Sigma^{-1} S^{\frac{1}{2}})] = \sum_{s=1}^{\ell} v[\lambda_s(S_1^{\frac{1}{2}} \Sigma_1^{-1} S_1^{\frac{1}{2}})]$$
$$+ \sum_{t=1}^{k-\ell} v[\lambda_t(S_2^{\frac{1}{2}} \Sigma_2^{-1} S_2^{\frac{1}{2}})],$$

which imples (6.3). This completes the proof.

Now we define an estimator $T: \mathcal{G} \to \mathcal{F}$ as

$$\rho_v[f_S, T(f_S)] = \min_{f \in \mathcal{F}} \rho_v(f_S, f),$$

which is called the minimum contrast estimator by ρ_v according to the terminology in Section 5.

From now on, we investigate the properties of this estimator. First we represent the α-conjugate metric structure on $\mathcal{F}$ with respect to the convariance parameter Σ in the following way. The information metric has the matrix form of components

$$\frac{1}{2} \Sigma^{-1} \otimes \Sigma^{-1}$$

and the α-affine connection has the matrix form of coefficients

$$\frac{1+\alpha}{2} \Sigma^{-1} \otimes \Sigma^{-1} \otimes \Sigma^{-1}$$

with respect to Σ-coordinates, where $\otimes$ denotes the Kronecker product. On the

other hand, we obtain the conjugate metric structure $\mathscr{C}(\rho_v)$ by the arguments similar to $\mathscr{C}_\alpha$ as follows: The metric $g^{(\rho_v)}$ is

$$\left[\frac{\partial}{\partial S}\otimes\frac{\partial}{\partial S}\rho_v(S,\Sigma)\right]_{S=\Sigma}=\frac{1}{2}\Sigma^{-1}\otimes\Sigma^{-1}$$

and the affine connection $\Gamma^{(\rho_v)}$ is

$$\left[-\frac{\partial}{\partial S}\otimes\frac{\partial}{\partial S}\otimes\frac{\partial}{\partial\Sigma}\rho_v(S,\Sigma)\right]_{S=\Sigma}=(v'''(1)+1)\,\Sigma^{-1}\otimes\Sigma^{-1}\otimes\Sigma^{-1},$$

since

$$\rho_v(S,\Sigma)=\frac{1}{4}\operatorname{tr}\{S(\Sigma^{-1}-S^{-1})\}^2+\frac{1}{6}v'''(1)\operatorname{tr}\{S(\Sigma^{-1}-S^{-1})\}^3$$
$$+\text{(higher order terms)}.$$

From these results, we immediately have

$$\mathscr{C}(\rho_v)=\mathscr{C}_{\alpha_v}$$

with $\alpha_v=-3-2v'''(1)$, which leads the following theorem because of Theorems 3.2 and 5.2.

THEOREM 6.3. *Let ρ_v be a contrast function of S-type on $\mathscr{G}$. Then the minimum contrast estimator by ρ_v has the limiting information loss*

$$\langle\overset{e}{H}(\mathscr{F})\rangle^2+(1+v'''(1))^2\langle T\rangle^2, \tag{6.4}$$

where $\langle H(\mathscr{F})\rangle^2$ and $\langle T\rangle^2$ are defined in (5.1) *for the covariance structure model $\mathscr{F}$.*

Let ρ_w be a contrast function of W-type. From Theorem 6.2, it follows that the minimum contrast estimator by ρ_v has the same limiting information loss as the estimator by ρ_w if $w'''(1)+v'''(1)=0$

Recall the operations $*$, $\oplus$ and $\ominus$ on $\mathscr{W}_1$ in Section 2. These operations can be also defined on the space $\mathscr{V}$ of all convex functions v's with $v(1)=v'(1)=0$ and $v''(1)=\frac{1}{2}$. Thus define $\nu\colon\mathscr{V}\to\mathscr{V}$ by

$$\nu(v)=\left(\frac{v+v^\circ}{2}\right)^\ominus$$

Then the minimum contrast estimator by $\rho_{\nu(v)}$ is second order efficient for every v in $\mathscr{V}$ by a similar reason as in the proof of Theorem 5.3.

Next by using the commutative relation between the trace and the differential operations on matrices, we present explicit formulas of the limiting information loss of the minimum contrast estimator by ρ_v or ρ_w.

THEOREM 6.4. *Let $\mathscr{F}$ be a covariance structure model of dimension n as in* (6.4). *Then the Gramian forms of $\langle \overset{e}{H}(\mathscr{F})\rangle^2$ and $\langle T\rangle^2$ as in* (6.4) *have the following components.*

$$
\begin{aligned}
(6.5)\quad \langle \overset{e}{H}(\mathscr{F})\rangle^2_{ab} &= \operatorname{tr}\{(\partial_a\Sigma\Sigma^{-1}\partial_c\Sigma\Sigma^{-1} + \partial_c\Sigma\Sigma^{-1}\partial_a\Sigma\Sigma^{-1} - \partial_c\partial_a\Sigma\Sigma^{-1}) \\
&\quad (\partial_b\Sigma\Sigma^{-1}\partial_d\Sigma\Sigma^{-1} + \partial_d\Sigma\Sigma^{-1}\partial_b\Sigma\Sigma^{-1} - \partial_d\partial_b\Sigma\Sigma^{-1})\}\tilde{g}^{cd} \\
&\quad - \operatorname{tr}\{(\partial_a\Sigma\Sigma^{-1}\partial_c\Sigma\Sigma^{-1} + \partial_c\Sigma\Sigma^{-1}\partial_a\Sigma\Sigma^{-1} - \partial_c\partial_a\Sigma\Sigma^{-1})\partial_e\Sigma\Sigma^{-1}\} \\
&\quad \times \operatorname{tr}\{(\partial_b\Sigma\Sigma^{-1}\partial_d\Sigma\Sigma^{-1} + \partial_d\Sigma\Sigma^{-1}\partial_b\Sigma\Sigma^{-1} - \partial_d\partial_b\Sigma\Sigma^{-1}) \\
&\quad \times \partial_h\Sigma\Sigma^{-1}\}\tilde{g}^{cd}\tilde{g}^{eh},
\end{aligned}
$$

and

$$
\begin{aligned}
(6.6)\quad \langle T\rangle^2_{ab} &= 2\operatorname{tr}(\partial_a\Sigma\Sigma^{-1})\operatorname{tr}(\partial_b\Sigma\Sigma^{-1}) + 2k\operatorname{tr}(\partial_a\Sigma\Sigma^{-1}\partial_b\Sigma\Sigma^{-1}) \\
&\quad - \operatorname{tr}(\partial_a\Sigma\Sigma^{-1}\partial_c\Sigma\Sigma^{-1}\partial_b\Sigma\Sigma^{-1}\partial_d\Sigma\Sigma^{-1}\tilde{g}^{cd} + \operatorname{tr}(\partial_a\Sigma\Sigma^{-1}\partial_c\Sigma\Sigma^{-1}\partial_e\Sigma\Sigma^{-1}) \\
&\quad \times \operatorname{tr}(\partial_b\Sigma\Sigma^{-1}\partial_d\Sigma\Sigma^{-1}\partial_h\Sigma\Sigma^{-1})\tilde{g}^{cd}\tilde{g}^{eh}
\end{aligned}
$$

for a, b=1, 2,..., n. Here $\partial_a\Sigma=(\partial/\partial\theta^a)\Sigma(\boldsymbol{\theta})$ and $\tilde{g}^{ab}$ denotes the (a, b)-entry of the inverse matrix of

$$
\tilde{G} = \left\{\frac{1}{2}\operatorname{tr}(\partial_a\Sigma\Sigma^{-1}\partial_b\Sigma\Sigma^{-1})\right\}_{a=b=1,2,\ldots,n}.
$$

PROOF. We show the formula (6.5). For the decomposition

$$
T_f(\mathscr{G}) = T_f(\mathscr{F}) \oplus T_f^{\perp}(\mathscr{F}),
$$

we define B and $B^{\perp}$ by the matrix representations of the connecting tensors of $T_f(\mathscr{G})$ with $T_f(\mathscr{F})$ and $T_f^{\perp}(\mathscr{F})$, respectively. That is, $B=\{\operatorname{vec}(\partial_a\Sigma)\}_{a=1,2,\ldots,n.}$ and $B^{\perp}$ satisfying

$$
B^{\perp}(\Sigma^{-1}\otimes\Sigma^{-1})B = O,
$$

where $\operatorname{vec}\Sigma=(\sigma_{11}, \sigma_{12},\ldots, \sigma_{1k},\ldots, \sigma_{k1}, \sigma_{k2},\ldots, \sigma_{kk})'$. It follows from the decomposition that

$$
(6.7)\qquad B^{\perp}\{B^{\perp\prime}(\Sigma^{-1}\otimes\Sigma^{-1})B^{\perp}\}^{-1}B^{\perp\prime} = \Sigma\otimes\Sigma - B\{B'(\Sigma^{-1}\otimes\Sigma^{-1})B\}^{-1}B'.
$$

By the definition of $\overset{e}{H}(\mathscr{F})$ as in (1.3), it holds

$$
\begin{aligned}
(6.8)\quad &\langle \overset{e}{H}(\mathscr{F})\rangle^2_{ab} \\
&\quad = \{\operatorname{vec}(\partial_a\partial_c\Sigma^{-1})\}'B^{\perp}\{B^{\perp\prime}(\Sigma^{-1}\otimes\Sigma^{-1})B^{\perp}\}B^{\perp}\operatorname{vec}(\partial_b\partial_d\Sigma^{-1})\tilde{g}^{cd}.
\end{aligned}
$$

for $a, b=1, 2,\ldots, n$ with respect to $\boldsymbol{\theta}$. By inserting (6.7) into the right-hand side of (6.8), we have

$$\langle \overset{e}{H}(\mathscr{F})\rangle_{ab}^{2} = 2\,\mathrm{tr}\,(\Sigma\partial_a\partial_c\Sigma^{-1}\Sigma\partial_b\partial_d\Sigma^{-1})\tilde{g}^{cd} - \mathrm{tr}\,(\Sigma\partial_a\partial_c\Sigma^{-1}\partial_e\Sigma\Sigma^{-1})\mathrm{tr}\,(\Sigma\partial_b\partial_d\Sigma^{-1}\partial_f\Sigma\Sigma^{-1})\tilde{g}^{cd}\tilde{g}^{ef}$$

since it holds

$$(\mathrm{vec}\,A)'B\otimes C\,\mathrm{vec}\,D = \mathrm{tr}\,(ABCD)$$

for each square matrix A, B, C and D. Because of an identity

$$\Sigma\partial_a\partial_c\Sigma^{-1} = \partial_a\Sigma\Sigma^{-1}\partial_c\Sigma\Sigma^{-1} + \partial_c\Sigma\Sigma^{-1}\partial_a\Sigma\Sigma^{-1} - \partial_a\partial_c\Sigma\Sigma^{-1},$$

the formula (6.5) holds. We can show the formula (6.6) by the same way as the derivation of (6.5) (see, Eguchi [26] for detailed derivation). The proof is complete.

We remark that the derivation of $\overset{e}{H}(\mathscr{F})$ can be also given by a straightforward extension of the generalized definition of curvatures by Efron [21] to multidimensional cases:

$$\boldsymbol{E}\{\partial_a\partial_c\ell\tilde{g}^{cd}\partial_d\partial_b\ell\} - \boldsymbol{E}\{\partial_c\partial_a\ell\partial_e\ell\}\tilde{g}^{cd}\tilde{g}^{eh}\boldsymbol{E}\{\partial_d\partial_b\ell\partial_h\ell\},$$

where $\ell = -\mathrm{tr}\,(\Sigma^{-1}\boldsymbol{x}\boldsymbol{x}')/2 - \log(\det\Sigma)/2$.

We, up to now, disscussed the properties of estimators independent of the parametrization (6.1), *e.g.*, Fisher-consistency, first order and second order efficiency. This discussion is justified since the estimators in our formulation are independent of (6.1). Nevertheless we often have our interests in the parameter $\boldsymbol{\theta}$, itself (cf. Efron [22] and Berkson [12]). Thus, from now on, we intestigate the $\boldsymbol{\theta}$-version of the minimum contrast estimator $T(f_S)$ by ρ_v, i.e.,

$$\hat{\boldsymbol{\theta}}(S) = \phi^{-1}(T(f_S))$$

with the parametrization ϕ from Θ to $\mathscr{F}$ as in (6.1). The asymptotic bias of the estimator $\hat{\boldsymbol{\theta}}$ is expressed as

$$\boldsymbol{E}\,\hat{\theta}^a = \theta^a + \frac{1}{2N}[b_1^a(\boldsymbol{\theta}) - (v'''(1)+1)b_2^a(\boldsymbol{\theta})] + 0(N^{-2})$$

for $a=1, 2,\ldots, n$, with the sample size N where

$$b_1^a(\boldsymbol{\theta}) = \mathrm{tr}\,(\partial_b\partial_c\Sigma\Sigma^{-1}\partial_d\Sigma\Sigma^{-1})\tilde{g}^{ab}\tilde{g}^{cd}$$

and

$$b_2^a(\boldsymbol{\theta}) = \mathrm{tr}\,(\partial_b\Sigma\Sigma^{-1} - \partial_b\Sigma\Sigma^{-1}\partial_c\Sigma\Sigma^{-1}\partial_d\Sigma\Sigma^{-1}\tilde{g}^{cd})\tilde{g}^{ab}.$$

Hence the bias-corrected estimator is defined by

$$\hat{\theta}^{*a} = \hat{\theta}^a - \frac{1}{2N}[b_1^a(\hat{\boldsymbol{\theta}}) - b_2^a(\hat{\boldsymbol{\theta}})] .$$

For the bias-corrected estimator $\hat{\boldsymbol{\theta}}^*$,

$$E(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta})' = \frac{1}{N}\tilde{G}^{-1} + \frac{1}{N^2}\{\langle \overset{m}{\Gamma} \rangle^2 + \langle \overset{e}{H} \rangle^2 + (v'''(1)+1)^2\langle T \rangle^2\} + O(N^{-3})$$

where

$$\langle \overset{m}{\Gamma} \rangle^2 = \{\mathrm{tr}\,(\partial_a\partial_c \Sigma\Sigma^{-1}\partial_e \Sigma\Sigma^{-1})\,\mathrm{tr}\,(\partial_b\partial_d \Sigma\Sigma^{-1}\partial_c\partial_f \Sigma\Sigma^{-1})\tilde{g}^{cd}\tilde{g}^{ef}\}_{a,b=1,2,\ldots,n}.$$

We apply the above formulas to two pratical models:

Example 6.1. (*Intraclass correlation model*) Let S be sample covariance matrix from a p-variate normal density with covariance matrix

$$\Sigma = (\sigma_{rs}), \qquad \begin{aligned} \sigma_{rs} &= 1 \quad \text{if} \quad r = s, \\ &= \theta \quad \text{otherwise.} \end{aligned}$$

Then a generalized least squares estimate $\hat{\theta}$ based on $v(\lambda)=(\lambda-1)^2/4$ is

$$\mathrm{tr}\,\{S^{-1}J(I-S)\}/\mathrm{tr}\,(S^{-1}J)^2,$$

where $J=\mathbf{1}\mathbf{1}'-I$ with $\mathbf{1}=(1, 1,\ldots, 1)'$. The limiting information loss by $\hat{\theta}$ is of form $\overset{e}{H}_0^2+16T_0^2$, where

$$\overset{e}{H}_0^2 = 16\,\mathrm{tr}\,(J\Sigma^{-1})^4/\mathrm{tr}\,(J\Sigma^{-1})^2 - \{4\,\mathrm{tr}\,(J\Sigma^{-1})^3/\mathrm{tr}\,(J\Sigma^{-1})^2\}^2$$

and

$$T_0^2 = 4(\mathrm{tr}\,J\Sigma^{-1})^2 + 2\,\mathrm{tr}\,(J\Sigma^{-1})^2 - 8\,\mathrm{tr}\,(J\Sigma^{-1})^4/\mathrm{tr}\,(J\Sigma^{-1})^2 + 4(\mathrm{tr}\,J\Sigma^{-1})^3/\mathrm{tr}\,(J\Sigma^{-1}))^2.$$

In particular $\overset{e}{H}_0^2$ has a simple form

$$16(p-1)/\{1+(p-1)\theta^2\}^2.$$

It is noted that the term $\overset{m}{\Gamma}$ vanishes at every θ. The mean squared error of the bias-corrected estimator

$$\hat{\theta}^* = \hat{\theta} - \frac{8}{N}\mathrm{tr}\,\{J\hat{\Sigma}^{-1} - 2(J\hat{\Sigma}^{-1})^4/\mathrm{tr}\,(J\hat{\Sigma}^{-1})^2\}/\mathrm{tr}\,(J\hat{\Sigma}^{-1})^2$$

with $\hat{\Sigma}=\Sigma(\hat{\theta})$ is

$$2/(N\,\mathrm{tr}\,(J\Sigma^{-1})^2) + \frac{4}{N^2}(H_0+T_0)/\mathrm{tr}\,(J\Sigma^{-1})^2$$

$$\left(=\frac{1}{N}U + \frac{1}{N^2}(V_1+V_2), \text{ say}\right) + 0(N^{-3})$$

Table shows that the influence V_2 in terms of the estimator $\hat{\theta}^*$ is fairly larger than the term V_1 caused by the model.

Table. Values of U, V_1 and V_2 at true vaues θ's in the case of $k=6$.

θ	U	V_1	V_2
0.7	0.0171	0.0021	5.5572
0.6	0.0305	0.0095	9.2169
0.5	0.0454	0.0325	12.9174
0.4	0.0600	0.0889	15.6444
0.3	0.0704	0.1886	16.1355
0.2	0.0711	0.2809	13.6252
0.1	0.0579	0.2429	9.0412
0.0	0.0333	0.0889	4.5511
−0.1	0.0096	0.0067	1.3379

Example 6.2. (*Linear covariance model*) A linear covariance structure is described as

$$\Sigma = V_1\theta^1 + V_2\theta^2 + \cdots + V_q\theta^q.$$

for $\boldsymbol{\theta}=(\theta^1,\ldots,\theta^q)$, where positive definte matrices $V_1, V_2,\ldots, V_q$ are linearly independent in the space of all symmetric matrices. A generalized least squares estimates $\hat{\boldsymbol{\theta}}$ based on $v(\lambda)=(\lambda-1)^2/4$ is given as $\hat{\boldsymbol{\theta}}=\tilde{G}^{-1}(S)\boldsymbol{r}(S)$, where

$$\tilde{G}(S) = \left[\frac{1}{2}\operatorname{tr}(S^{-1}V_aS^{-1}V_b)\right]_{a,b=1,\ldots,q},$$

$$\boldsymbol{r}(S) = \left[\frac{1}{2}\operatorname{tr}(S^{-1}V_c)\right]_{c=1,\ldots,q}.$$

The information loss due to $\hat{\boldsymbol{\theta}}$ is represented as $\langle \overset{e}{H}_1\rangle^2+16\langle T_1\rangle^2$, where

$$\langle \overset{e}{H}_1\rangle^2 = 2\operatorname{tr}(V_a\Sigma^{-1}V_c\Sigma^{-1}V_b\Sigma^{-1}V_d\Sigma^{-1})\tilde{g}^{cd} - \operatorname{tr}(V_a\Sigma^{-1}V_c\Sigma^{-1}V_e\Sigma^{-1})\operatorname{tr}(V_b\Sigma^{-1}V_d\Sigma^{-1}V_f\Sigma^{-1})\tilde{g}^{cd}\tilde{g}^{ef}$$

and

$$\langle T_1\rangle^2 = \operatorname{tr}(V_a\Sigma^{-1}V_c\Sigma^{-1}V_e\Sigma^{-1})\operatorname{tr}(V_b\Sigma^{-1}V_d\Sigma^{-1}V_f\Sigma^{-1})\tilde{g}^{cd}\tilde{g}^{ef} + 2\operatorname{tr}(V_a\Sigma^{-1})\operatorname{tr}(V_b\Sigma^{-1}) + 2p\operatorname{tr}(V_a\Sigma^{-1}V_b\Sigma^{-1}) - 4\operatorname{tr}(V_a\Sigma^{-1}V_c\Sigma^{-1}V_b\Sigma^{-1}V_d\Sigma^{-1})\tilde{g}^{cd}$$

with $\tilde{g}^{cd}$ is the inverse element of the matrix

$$\{\operatorname{tr}(V_c\Sigma^{-1}V_d\Sigma^{-1})\}_{c,d=1,2,\ldots q}.$$

7. A generalized scoring method for a minimum contrast estimator

Let $\mathcal{F}$ be an (n, m)-curved subfamily of an exponential family $\mathcal{F}_e$ and let ρ be a contrast function on $\mathcal{F}_e$. In many practical models, the minimum contrast estimator by ρ cannot be obtained as a general solution. This aspect is reduced to a nonlinear optimization problem for the minimization. We now consider an algorithm for seeking the minimum contrast estimator $T(\hat{f})$ by ρ, *i.e.*,

$$\rho[\hat{f}, T(\hat{f})] = \min_{g\in\mathcal{F}} \rho(\hat{f}, g)$$

with the maximum likelihood summary $\hat{f}$. Take a parametrization

$$\mathcal{F} = \{f_{\beta(\theta)}(\boldsymbol{x}) = e^{\beta(\theta)'\boldsymbol{x}-\psi[\beta(\theta)}: \boldsymbol{\theta}\in\Theta\}\,. \tag{7.1}$$

and let the parametric form of ρ on $\mathcal{F}_e\times\mathcal{F}$ denote by

$$\rho(\boldsymbol{\beta}, \boldsymbol{\theta}) = \rho[f_{\beta}, f_{\beta(\theta)}]\,.$$

Under this parametrization we define an algorithm in terms of the following mapping S from $\mathcal{F}$ to $\mathcal{F}$ defined by

$$\hat{S}(f_{\theta}) = f_{\hat{s}(\theta)}\,,$$

where

$$\hat{s}(\boldsymbol{\theta}) = \boldsymbol{\theta} + (G^{(\rho)}(\boldsymbol{\theta}))^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}\rho(\boldsymbol{\beta}, \boldsymbol{\theta}) \tag{7.2}$$

with $\partial/\partial\boldsymbol{\theta}=(\partial/\partial\theta^1,\ldots, \partial/\partial\theta^m)'$. Then the algorithm constritutes a sequence

$$\hat{S}_k(f) = \overbrace{\hat{S}\circ\cdots\circ\hat{S}}^{k-\text{times}}(f) \tag{7.3}$$

from a starting point $f=f_{\theta}$. The algorithm is nothing but the Fisher scoring method in the case that ρ is the Kullback-Leibler information. In this sense, we call the algorithm a generalized scoring method. We note that if $\hat{s}(\boldsymbol{\theta})$ is defined by

$$H^{(\rho)}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\rho(\boldsymbol{\beta}, \boldsymbol{\theta})$$

in place of $G^{(\rho)}(\boldsymbol{\theta})$ in (7.2), then this algorithm is reduced to the Newton-Raphson method. As is well-known, the sequence has quadratic convergence

$$\|\hat{S}_{k+1}(f) - T(\hat{f})\| \leqq \varepsilon\|\hat{S}_k(f) - T(\hat{f})\|^2 \tag{7.4}$$

for the minimum contrast estimator $T(\hat{f})$, where $\|\cdot\|$ denotes the m-dimensional Euclidean norm and ε is some positive constant.

What we should pay attention is that the generalized scoring method deeply depends on the prametrization (7.1), while the minimum contrast estimator $T(\hat{f})$ is independent of the parametrization. Since our purpose is to seek $T(\hat{f})$, it is necessary to know how the parametrization affects the convergence of the squence (7.3). Thus, in this situation, the quadratic convergence property of (7.4) has no significance because this property relates only to the parameter space Θ. Therefore we now investigate the convergence of this algorithm in terms of the conjugate metric structure $\mathscr{C}(\rho)$.

As is introduced in Section 1, let us give the formula of

$$\mathscr{C}(\rho) = (g^{(\rho)}, \Gamma^{(\rho)}, {}^*\Gamma^{(\rho)})$$

induced from $\mathscr{F}_e$ to $\mathscr{F}$. The orthogonal decomposition

$$T_f(\mathscr{F}_e) = T_f(\mathscr{F}) \oplus T_f^{\perp}(\mathscr{F})$$

with respect to the metric $g^{(\rho)}$ leads the components $\{g_{ij}^{(\rho)}\}$ of $g^{(\rho)}$ on $T_f(\mathscr{F}_e)$ to the components $\{\tilde{g}_{ab}^{(\rho)}\}$ and $\{\tilde{g}_{\mu\lambda}^{(\rho)}\}$ on $T_f(\mathscr{F})$ and $T_f^{\perp}(\mathscr{F})$ respectively, as is similar to (1.1) and (1.2) for the metric $g^{(\rho)}$ in place of the information metric g. Similarly the connection $\Gamma^{(\rho)}$ over $\mathscr{F}_e$ is induced as

$$\tilde{\Gamma}_{ab,c}^{(\rho)} = \partial_a B_b^i B_c^j g_{ij}^{(\rho)} + B_a^i B_b^j B_c^k \Gamma_{ij,k}^{(\rho)}$$

with the coefficients $\{\Gamma_{ij,k}^{(\rho)}\}$ of $\Gamma^{(\rho)}$. We denote the embedding curvature tensor with respect to the connection ${}^*\Gamma^{(\rho)}$ by ${}^*H^{(\rho)}$, i.e.,

$${}^*H_{ab\lambda}^{(\rho)} = \partial_a B_b^i B_\lambda^j g_{ij}^{(\rho)} + B_a^i B_b^j B_\lambda^k {}^*\Gamma_{ij,k}^{(\rho)}$$

with the coefficients $\{{}^*\Gamma_{ij,k}^{(\rho)}\}$ of ${}^*\Gamma^{(\rho)}$. We write

$$e^a = e^a(\boldsymbol{\theta}) = g^{(\rho)ab}(\boldsymbol{\theta}) B_b^i(\boldsymbol{\theta}) e_i[\boldsymbol{\beta}(\boldsymbol{\theta})]$$

and

$$e^\lambda = e^\lambda(\boldsymbol{\theta}) = g^{(\rho)\lambda\mu}(\boldsymbol{\theta}) B_\mu^i(\boldsymbol{\theta}) e_i[\boldsymbol{\beta}(\boldsymbol{\theta})]$$

where $e_i(\boldsymbol{\beta}) = (\partial/\partial\beta^i)\rho(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta})$ with $f_{\hat{\beta}} = \hat{f}$.

On the basis of the above setup, we state the following property without reference to the parametrization.

THEOREM 7.1. *Let $\{\hat{S}_k(f)\}_{k=1,2,\ldots}$ be a sequence starting from f in $\mathscr{F}$ which is defined in* (7.3). *Then it holds,*

$$\rho[\hat{f}, \hat{S}_k(f)] = \rho[\hat{f}, \hat{S}_{k+1}(f)] + \rho[\hat{S}_{k+1}(f), \hat{S}_k(f)] + O(\|\mathbf{e}(\boldsymbol{\beta})\|^3), \tag{7.5}$$

for any integer k with $f = f_\beta$. Furthermore the third term can be expressed as

$$-\frac{1}{2}\Gamma^{(\rho)}_{ab,c}\, e^a e^b e^c + \frac{1}{2} {}^*H^{(\rho)}_{ab\lambda} e^a e^b e^\lambda. \tag{7.6}$$

PROOF. Because of the formula (3.8), the proof follows from a straightforward claculation.

We remark that the convergence of the generalized scoring method is attainable as

$$\rho(\hat{f}, f) > \rho[\hat{f}, \hat{S}(f)] > \cdots > \rho[\hat{f}, \hat{S}_k(f)] > \\ > \rho[\hat{f}, \hat{S}_{k+1}(f)] > \cdots > \rho[\hat{f}, T(\hat{f})]$$

through a kind of the Pythagoras theorem (7.5). However this mechanism is strongly affected by the terms of (7.6). The first term depends on the parametrization, while the second is invariant under transformations on the parmaeter space. If we restrict the model $\mathscr{F}$ to a nonlinear regression model and consider the maximum likelihood estimator, then the components of $\Gamma^{(\rho)}$ and ${}^*H^{(\rho)}$ are reduced to the parameter effect curvatures $A^T_{\cdot\cdot}$ and the intrinsic curvature $A^N_{\cdot\cdot}$ by Bates and Watts [10]. Furtheremore we can extend the results by Hamilton, Watts and Bates [31] to a minimum contrast estimator in a general curved exponental model:

THEOREM 7.2. *Let $T(\hat{f})$ be a minimum contrast estimator by ρ. Then it holds that*

$$\rho(\hat{f}, f) - \rho[\hat{f}, T(\hat{f})] - \rho[T(\hat{f}), f] \\ = \frac{1}{2} {}^*H^{(\rho)}_{ab\lambda} e^a e^b e^\lambda + (\textit{higher order terms}). \tag{7.7}$$

for each f in $\mathscr{F}$.

PROOF. The proof is immediate.

Theorem 7.2 shows that the invariance of $T(\hat{f})$ for parameter transformations leads to the invariance of the right-hand side of (7.7).

Now we propose a modification of the generalized scoring method by

$$\hat{s}^*(\boldsymbol{\theta}) = \hat{s}(\boldsymbol{\theta}) + \frac{1}{2}\Gamma^{(\rho)a}_{b,c} e^b e^c - {}^*H^{(\rho)a}_{b\lambda} e^b e^\lambda$$

in place of $\hat{s}(\boldsymbol{\theta})$ in (7.2), by which we define an algorithm

$$\hat{S}^*(f_\theta) = f_{\hat{s}^*(\theta)} \tag{7.8}$$

By Theorems 7.1 and 7.2, the following theorem holds.

THEOREM 7.3. For the modified method S^* as in (7.8), we have

$$\rho[\hat{f}, S^*(f_\beta)] = \rho[\hat{f}, T(\hat{f})] + O(\|e(\beta)\|^4]$$

with $\beta = \beta(\theta)$.

Acknowledgements

I wish to express my gratitude to Professor Y. Fujikoshi of Hiroshima University for his initial recommendation to write this paper, consistent encouragement and also to Drs. M. Taniguchi and R. Nishii for suggestions leading to substantial improvements of the manuscript. I am grateful to Professor S. Amari of University of Tokyo for helpful suggestions and essential comments. Finally I am indebted to Professors M. Okamoto and N. Inagaki and Dr. Y. Toyooka of Osaka University for guidance to many fields of mathematical statistics with instructive encouragement.

References

[1] M. Akahira and K. Takeuchi, Asymptotic efficiency of statistical estimators: Concepts and higher order efficiency, Lecture Notes in Statistics, No. **7** (1981), Springer, New York.

[2] S. Amari, Differential geometry of curved exponential families — curvatures and information loss, Ann. Statist. **10** (1982), 357–387.

[3] S. Amari, Differential geometry of statistical inference, Probability Theory and Mathematical Statistics (ed. K. Ito and J. V. Prokhorov), Springer Lecture Notes in Math. (1983), 26–40.

[3] S. Amari, Geometric theory of asymptotic ancillality and conditional inference, Biometrika **69** (1982), 1–17.

[5] A. Amari and M. Kumon, Differential geometry of Edgeworth expansions in curved exponential family, Ann. Inst. Statist. Math. **35A** (1983), 1–24.

[6] T. W. Anderson, Asymptotically efficient estimation of covariance matrices with linear structure, Ann. Statist. **1** (1971), 135–141.

[7] C. Atkinson and A. F. Mitchell, Rao's distance measure, Sankyā **A43** (1981), 345–365.

[8] R. R. Bahadule, On asymptotic efficiency of tests and estimates, Sankyā **22** (1960), 229–252.

[9] O. E. Barndorff-Nielsen and Blæsid, Exponential models with affine dual foliations, Ann. Statist. **11** (1983), 753–769.

[10] D. M. Bates and D. G. Watts, Relative curvature measure of non-linearity, J. Roy. Statist. Soc. **B40** (1980), 1–25.

[11] J. Berkson, Minimum chi-square, not maximum likelihood! (with discussion), Ann. Statist. **8** (1980), 457–487.

[12] A. Bhattacharya, On a measure of divergece of two multinomial populations, Sankyā **7** (1946), 401.

[13] L. Boltzmann, Vorlesungen über Gastheorie, 2 Bde, J. A. Barth (1912).

[14] M. W. Browne, Generalized least squres estimation in the analysis of covariance structures, South African Statist. J. **8** (1974), 1–24.

[15] J. Burbea and C. R. Rao, Entropy differential metric, distance and divergence measures in probability spaces: A unified apprach, J. Multi. Var. Analys. **12** (1982), 575–596.

[16] N. N. Chentsov, Statistical Decision Rules and Optimal Inference (In Russian), (1972), Nauka, Moscow, translated in English (1982), AMS, Rhode Island.
[17] H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, Ann. Math. Statist. **23** (1952), 493–507.
[18] I. Csiszar, I-divergence geometry of probability distributions and minimization ploblems, Ann. Prob. **3** (1975), 146–158.
[19] A. P. Dawid, Discussion to Efron's paper, Ann. Statist. **3** (1975), 1231–1234.
[20] A. P. Dawid, Further comments on a paper by Bradley Efron, Ann. Statist. **8** (1977), 1249.
[21] B. Efron, Defining the curvature of a statistical problem (with discussion) Ann. Statist. **6** (1975), 1189–1242.
[22] B. Efron, Maximum likelihood and decision theory, Ann. Statist. **10** (1982), 340–356.
[23] S. Eguchi, Second order efficiency of minimum contrast estimators in a curved exponential family, Ann. Statist. **11** (1983), 793–803.
[24] S. Eguchi, Model-fidelity of the maximum likelihood estimator in a curved exponential family, In Statistical Theory and Data Analysis, e.d. K. Matusita (1985) North Holland, Amsterdam.
[25] S. Eguchi, A characterization of second order efficiency in a curved exponential family, Ann. Inst. Statist. Math. **36A** (1984), 199–206.
[26] S. Eguchi, Asymptotic theory of covariance structure models. Tech. Rep. Statist. Research Group, Hiroshima University (1984).
[27] S. Eguchi, Projective Mahalanobis estimator in a curved exponential family, Tech. Rep. Statist. Research Group, Hiroshima University (1984).
[28] R. A. Fisher, Theory of Statistical estimation, Proc. Cambridge Philos. **122** (1925), 700–725.
[29] J. K. Ghosh and K. Subramanyam, Second order efficiency of the maximum likelihood estimators, Sankyā **A36** (1974), 325–358.
[30] J. B. S. Haldane, A class of efficient estimators of a parameter, Bull. Int. Statist. Inst. **33** (1951), 231.
[31] D. C. Hamilton, D. G. Watts and D. M. Bates, Accounting for intrinsic nonlinear regression parameter inference regions, Ann. Statist. **10** (1982), 386–393.
[32] Y. Hosoya, High-order efficiency in the estimation of linear processes, Ann. Statist **7** (1979), 516–530.
[33] P. Hougaard, Parametrization of non-linear models, J. R. Statist. Soc. **44** (1983), 244–252.
[34] A. T. James, The variance information manifold and functions on it, In Multivariate analysis III, ed. P. K. Krishnaiah (1973), 157–169, Academic Press, New York.
[35] H. Jeffreys, Theory of Probability theory (2nd ed.) (1948) Oxford Univ. Press.
[36] A. M. Kagan, On the theory of Fisher's amount of information, Dokl. Akad. Nauk SSSR **151** (1963), 277–278.
[37] R. Kass, Canonical parametrization and zero parameter-effect curvature, J. Roy. Statist. Soc. **B45** (1984), 86–92.
[38] S. Kullback and R. A. Leibler, On information and sufficiency, Ann. Math. Statist. **22** (1951), 79.
[39] S. Kullback, Information theory and statistics, Wiley, New York (1959).
[40] M. Kumon and S. Amari, Geometrical theory of higher-order asymptotics of test, interval estimator and conditional inference, Proc. Roy. Soc. London, **A387** (1983), 429–458.

[41] M. Kumon and S. Amari, Estimation of structural parameter in the presence of a large number of nuisance parameters. To appear in Biometrika (1984).
[42] S. L. Lauritzen, Statistical manifolds, Tech. Rep. of Aalborg University Center (1984).
[43] P. C. Mahalanobis, On the generalized distance in statistics, Proc. Nat. Inst. of Sciences of India, **2** (1936), 49–55.
[44] K. Matusita, Decision rule based on the distance of the classification problem, Ann. Inst. Statist. Math. **8** (1955), 67–77.
[45] H. Nagaoka and S. Amari, Differential geometry of smooth families of probability distributions, METR 82–7, Univ. Tokyo (1982).
[46] J. Pfanzagl, Contributions to General Asymptotic Statistical Theory. Lecture Notes in Statistics **13**, Springer (1982).
[47] C. R. Rao, Information and accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. **37** (1945), 81–91.
[48] C. R. Rao, Asymptotic efficiency and limiting information, (ed. J. Neyman) Proc. Fourth Berkeley Syp. Math. Statist. Prob. **1** (1961), 531–545, Univ. of California Press.
[49] C. R. Rao, Effcient estimates and optimum inference rocedures in large samples, J. Roy. Statist. Soc. B **24** (1962), 46–72.
[50] C. R. Rao, Criteria of estimation in large samples, Sankhyā **25** (1963), 189–206.
[51] C. R. Rao, Linear Statistical Inference and its Applications. Wiley, New York, (1963).
[52] C. R. Rao, Diversity: its measurement, decomposition, apportionment and analysis, Sankhyā **A44** (1982), 1–22.
[53] A. Renyi, On measures of entropy and information, Proc. Fourth Berkeley Symp. Math. Statisti. Prob. **1** (1961), 541–561.
[54] C. E. Shannon, A mathematical theory of communication. Bell System Tech. J. **27** (1948), 623–656.
[55] A. Shapiro, Asymptotic distribution theory in the analysis of cavariance structures (A unified approach), South African Statist. J. **17** (1983), 33–81.
[56] E. H. Simpson, Measurement of diversity, Nature **163** (1949), 688.
[57] L. T. Skovgaard, A Riemannian geometry of the multivariate normal model, To appear in Scand. J. Statist. (1984).
[58] A. J. Swain, A class of factor analysis estimation procedures with common asymptotic sampling properties, Psychometrika **40** (1975), 315–335.
[59] N. Taneichi, Y. Sato and M. Kawaguchi, On an extension of the parameter space of the multinomial distrubution, Bull. Fac. Eng. Hokkaido University **121** (1984), 59–67.
[60] M. Taniguchi, An estimation procedure of parameters of a certain spectral density model, J. Roy. Statist. Soc. B **43** (1981), 34–40.
[61] B. C. Wei and C. L. Tsai, Geometrical method of asymptotic conditional inference based on the subset parameters, Tech. Rep, 417, Univ. Minnesota, (1983).
[62] N. Wiener, Cybernetics, Wily, New York, (1948).
[63] T. Yoshizawa, A geometrical interpretation of location and scale parameters, Memo TYH–2, Harvard Univ., (1971).

Department of Mathematics,
Faculty of Science,
Hiroshima University