DIFFERENTIAL GEOMETRICAL THEORY OF STATISTICS

Shun-ichi Amari*

*Department of Mathematical Engineering and Instrumentation Physics, University of Tokyo, Tokyo, JAPAN

1. INTRODUCTION

Statistics is a science which studies methods of inference, from observed data, concerning the probabilistic structure underlying such data. The class of all the possible probability distributions is usually too wide to consider all its elements as candidates for the true probability distribution from which the data were derived. Statisticians often assume a statistical model which is a subset of the set of all the possible probability distributions, and evaluate procedures of statistical inference assuming that the model is faithful, i.e., it includes the true distribution. It should, however, be remarked that a model is not necessarily faithful but is approximately so. In either case, it should be very important to know the shape of a statistical model in the whole set of probability distributions. This is the geometry of a statistical model. A statistical model often forms a geometrical manifold, so that the geometry of manifolds should play an important role. Considering that properties of specific types of probability distributions, for example, of Gaussian distributions, of Wiener processes, and so on, have so far been studied in detail, it seems rather strange that only a few theories have been proposed concerning properties of a family itself of distributions. Here, by the properties of a family we mean such geometric relations as mutual distances, flatness or curvature of the family, etc. Obviously it is not a trivial task to define such geometric structures in a natural, useful and invariant manner.

Only local properties of a statistical model are responsible for the asymptotic theory of statistical inference. Local properties are represented by the geometry of the tangent spaces of the manifold. The tangent space has a

natural Riemannian metric given by the Fisher information matrix in the regular case. It represents only a local property of the model, because the tangent space is nothing but local linearization of the model manifold. In order to obtain larger-scale properties, one needs to define mutual relations of the two different tangent spaces at two neighboring points in the model. This can be done by defining a one-to-one affine correspondence between two tangent spaces, which is called an affine connection in differential geometry. By an affine connection, one can consider local properties around each point beyond the linear approximation. The curvature of a model can be obtained by the use of this connection. It is clear that such a differential-geometrical concept provides a tool convenient for studying higher-order asymptotic properties of inference. However, by connecting local tangent spaces further, one can obtain global relations. Hence, the validity of the differential-geometrical method is not limited within the framework of asymptotic theory.

It was Rao (1945) who first pointed out the importance in the differential-geometrical approach. He introduced the Riemannian metric by using the Fisher information matrix. Although a number of researches have been carried out along this Riemannian line (see, e.g., Amari (1968), Atkinson and Mitchell (1981), Dawid (1977), James (1973), Kass (1980), Skovgaard (1984), Yoshizawa (1971), etc.), they did not have a large impact on statistics. Some additional concepts are necessary to improve its usefulness. A new idea was developed by Chentsov (1972) in his Russian book (and in some papers prior to the book). He introduced a family of affine connections and proved their uniqueness from the point of view of categorical invariance. Although his theory was deep and fundamental, he did not discuss the curvature of a statistical model. Efron (1975, 1978), independently of Chentsov's work, provided a new idea by pointing out that the statistical curvature plays an important role in higher-order properties of statistical inference. Dawid (1975) pointed out further possibilities. Efron's idea was generalized by Madsen (1979) (see also Reeds (1975)). Amari (1980, 1982a) constructed a differential-geometrical method in

statistics by introducing a family of affine connections, which however turned out to be equivalent to Chentsov's. He further defined α-curvatures, and pointed out the fundamental roles of the exponential and mixture curvatures played in statistical inference. The theory has been developed further by a number of papers (Amari (1982b, 1983a, b), Amari and Kumon (1983), Kumon and Amari (1983, 1984, 1985), Nagaoka and Amari (1982), Eguchi (1983), Kass (1984)). The new developments were also shown in the NATO Research Workshop on Differential Geometry in Statistical Inference (see Barndorff-Nielsen (1985) and Lauritzen (1985)). They together seem to prove the usefulness of differential geometry as a fundamental method in statistics. (See also Csiszár (1975), Burbea and Rao (1982), Pfanzagl (1982), Beale (1960), Bates and Watts (1980), etc., for other geometrical work.)

The present article gives not only a compact review of various achievements up to now by the differential geometrical method most of which have already been published in various journals and in Amari (1985) but also a preview of new results and half-baked ideas in new directions, most of which have not yet been published. Chapter 2 provides an introduction to the geometrical method, and elucidates fundamental geometrical properties of statistical manifolds. Chapter 3 is devoted to the higher-order asymptotic theory of statistical inference, summarizing higher-order characteristics of various estimators and tests in geometrical terms. Chapter 4 discusses a higher-order theory of asymptotic sufficiency and ancillarity from the Fisher information point of view. Refer to Amari (1985) for more detailed explanations in these chapters; Lauritzen (1985) gives a good introduction to modern differential geometry. The remaining Chapters 5, 6, and 7 treat new ideas and developments which are just under construction. In Chapter 5 is introduced a fibre bundle approach, which is necessary in order to study properties of statistical inference in a general statistical model other than a curved exponential family. A Hilbert bundle and a jet bundle are treated in a geometrical framework of statistical inference. Chapter 6 gives a summary of a theory of estimation of a structural parameter

in the presence of nuisance parameters whose number increases in proportion to the number of observations. Here, the Hilbert bundle theory plays an essential role. Chapter 7 elucidates geometrical structures of parametric and non-parametric models of stationary Gaussian time series. The present approach is useful not only for constructing a higher-order theory of statistical inference on time series models, but also for constructing differential geometrical theory of systems and information theory (Amari, 1983 c). These three chapters are original and only sketches are given in the present paper. More detailed theoretical treatments and their applications will appear as separate papers in the near future.

2. GEOMETRICAL STRUCTURE OF STATISTICAL MODELS

Metric and α-connection

Let $S = \{p(x,\theta)\}$ be a statistical model consisting of probability density functions $p(x,\theta)$ of random variable $x \in X$ with respect to a measure P on X such that every distribution is uniquely parametrized by an n-dimensional vector parameter $\theta = (\theta^i) = (\theta^1,\ldots,\theta^n)$. Since the set $\{p(x)\}$ of all the density functions on X is a subset of the L_1 space of functions in x, S is considered to be a subset of the L_1 space. A statistical model S is said to be geometrically regular, when it satisfies the following regularity conditions $A_1 \sim A_6$, and S is regarded as an n-dimensional manifold with a coordinate system θ.

A_1. The domain Θ of the parameter θ is homeomorphic to an n-dimensional Euclidean space R^n.

A_2. The topology of S induced from R^n is compatible with the relative topology of S in the L_1 space.

A_3. The support of $p(x,\theta)$ is common for all $\theta \in \Theta$, so that $p(x,\theta)$ are mutually absolutely continuous.

A_4. Every density function $p(x,\theta)$ is a smooth function in θ uniformly in x, and the partial derivative $\partial/\partial\theta^i$ and integration of $\log p(x,\theta)$ with respect to the measure $P(x)$ are always commutative.

A_5. The moments of the score function $(\partial/\partial\theta^i)\log p(x,\theta)$ exist up to the third order and are smooth in θ.

A_6. The Fisher information matrix is positive definite.

Condition 1 implies that S itself is homeomorphic to R^n. It is

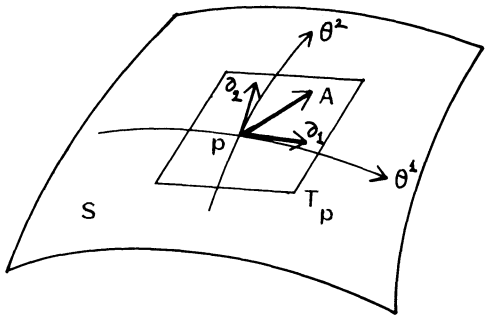

Figure 1

possible to weaken Condition 1. However, only local properties are treated here so that we assume it for the sake of simplicity. In a later section, we assume one more condition which guarantees the validity of Edgeworth expansions.

Let us denote by $\partial_i = \partial/\partial\theta^i$ the tangent vector e_i of the i-th coordinate curve θ^i (Fig. 1) at point θ. Then, n such tangent vectors $e_i = \partial_i$, $i = 1,\ldots, n$, span the tangent space T_θ at point θ of the manifold S. Any tangent vector $A \in T_\theta$ is a linear combination of the basis vectors ∂_i,

$$A = A^i \partial_i,$$

where A^i are the components of vector A and Einstein's summation convention is assumed throughout the paper, so that the summation Σ is automatically taken for those indices which appear twice in one term once as a subscript and once as a superscript. The tangent space T_θ is a linearized version of a small neighborhood at θ of S, and an infinitesimal vector $d\theta = d\theta^i \partial_i$ denotes the vector connecting two neighboring points θ and $\theta + d\theta$ or two neighboring distributions $p(x,\theta)$ and $p(x, \theta + d\theta)$.

Let us introduce a metric in the tangent space T_θ. It can be done by defining the inner product $g_{ij}(\theta) = \langle \partial_i, \partial_j \rangle$ of two basis vectors ∂_i and ∂_j at θ. To this end, we represent a vector $\partial_i \in T_\theta$ by a function $\partial_i \ell(x,\theta)$ in x, where $\ell(x,\theta) = \log p(x,\theta)$ and ∂_i (in $\partial_i \ell$) is the partial derivative $\partial/\partial\theta^i$. Then, it is natural to define the inner product by

$$g_{ij}(\theta) = \langle \partial_i, \partial_j \rangle = E_\theta[\partial_i \ell(x,\theta) \partial_j \ell(x,\theta)], \tag{2.1}$$

where E_θ denotes the expectation with respect to $p(x,\theta)$. This g_{ij} is the Fisher information matrix. Two vectors A and B are orthogonal when

$$\langle A,B\rangle = \langle A^i\partial_i, B^j\partial_j\rangle = A^iB^jg_{ij} = 0.$$

It is sometimes necessary to compare a vector $A\varepsilon T_\theta$ of the tangent space T_θ at one point θ with a vector $B\varepsilon T_{\theta'}$ belonging to the tangent space $T_{\theta'}$ at another point θ'. This can be done by comparing the basis vectors ∂_i at T_θ with the basis vectors ∂_i' at $T_{\theta'}$. Since T_θ and $T_{\theta'}$ are two different vector spaces, the two vectors ∂_i and ∂_i' are not directly comparable, and we need some way of identifying T_θ with $T_{\theta'}$ in order to compare the vectors in them. This can be accomplished by introducing an affine connection, which maps a tangent space $T_{\theta+d\theta}$ at $\theta + d\theta$ to the tangent space T_θ at θ. The mapping should reduce to the identity map as $d\theta\to 0$. Let $m(\partial_j')$ be the image of $\partial_j'\varepsilon T_{\theta+d\theta}$ mapped to T_θ. It is slightly different from $\partial_j\varepsilon T_\theta$. The vector

$$\nabla_{\partial_i}\partial_j = \lim_{d\theta\to 0}\frac{d}{d\theta^i}\{m(\partial_j') - \partial_j\}$$

represents the rate at which the j-th basis vector $\partial_j\varepsilon T_\theta$ "intrinsically" changes as the point θ moves from θ to $\theta+d\theta$ (Fig. 2) in the direction ∂_i. We call $\nabla_{\partial_i}\partial_j$ the covariant derivative of the basis vector ∂_j in the direction ∂_i. Since it is a vector of T_θ, its components are given by

$$\Gamma_{ijk} = \langle\nabla_{\partial_i}\partial_j, \partial_k\rangle \ , \tag{2.2}$$

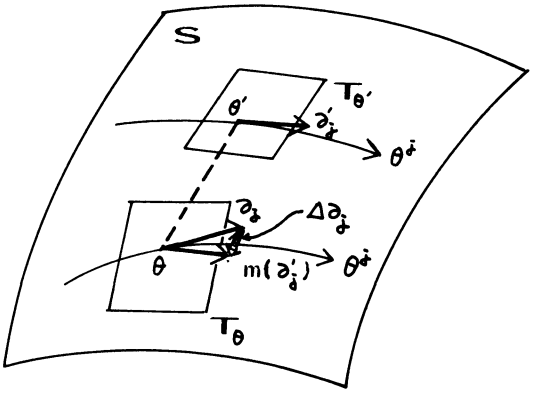

Figure 2

and

$$\nabla_{\partial_i}\partial_j = \Gamma_{ij}{}^{k}\partial_k ,$$

where $\Gamma_{ijk} = \Gamma_{ij}{}^{m}g_{mk}$. We call Γ_{ijk} the components of the affine connection. An affine connection is specified by defining $\nabla_{\partial_i}\partial_j$ or Γ_{ijk}. Let $A(\theta)$ be a vector field, which assigns to every point $\theta\varepsilon S$ a vector $A(\theta) = A^{i}(\theta)\partial_i \varepsilon T_\theta$. The intrinsic change of the vector $A(\theta)$ as the position θ moves is now given by the covariant derivative in the direction ∂_i of $A(\theta) = A^{j}(\theta)\partial_j$, defined by

$$\nabla_{\partial_i}A = (\partial_i A^{j})\partial_j + A^{j}(\nabla_{\partial_i}\partial_j) = (\partial_i A^{j} + \Gamma_{ik}{}^{j}A^{k})\partial_j,$$

in which the change in the basis vectors as well as that in the components $A^{i}(\theta)$ is taken into account. The covariant derivative in the direction $B = B^{i}\partial_i$ is given by

$$\nabla_B A = B^{i}\nabla_{\partial_i}A.$$

We have defined the covariant derivative by the use of the basis vectors ∂_i which are associated with the coordinate system or the parametrization θ. However, the covariant derivative $\nabla_B A$ is invariant under any parametrization, giving the same result in any coordinate system. This yields the transformation law for the components of a connection Γ_{ijk}. When another coordinate system (parametrization) $\theta' = \theta'(\theta)$ is used, the basis vectors change from $\{\partial_i\}$ to $\{\partial'_{i'}\}$, where

$$\partial'_{i'} = B^{i}_{i'}\partial_i,$$

and $B^{i}_{i'} = \partial\theta^{i}/\partial\theta'^{i'}$ is the inverse matrix of the Jacobian matrix of the coordinate transformation. Since the components $\Gamma'_{i'j'k'}$ of the connection are written as

$$\Gamma'_{i'j'k'} = \langle\nabla\partial_{i'}, \partial_{j'}, \partial_{k'}\rangle$$

in this new coordinate system, we easily have the transformation law

$$\Gamma'_{i'j'k'} = B^{i}_{i'}B^{j}_{j'}B^{k}_{k'}\Gamma_{ijk} + B^{i}_{i'}B^{k}_{k'}g_{kj}(\partial_i B^{j}_{j'}).$$

We introduce the α-connection, where α is a real parameter, in the statistical manifold S by the formula

$$\Gamma^{(\alpha)}_{ijk} = E_\theta[\{\partial_i\partial_j\ell(x,\theta) + \frac{1-\alpha}{2}\partial_i\ell(x,\theta)\partial_j\ell(x,\theta)\}\partial_k\ell(x,\theta)]. \qquad (2.3)$$

It is easily checked that the connection defined by (2.3) satisfies the transformation law. In particular, the 1-connection is called the exponential connection, and the -1-connection is called the mixture connection.

2.2 Imbedding and α-curvature

Let us consider an m-dimensional regular statistical model $M = \{q(x,u)\}$, which is imbedded in $S = \{p(x,\theta)\}$ by

$$q(x,u) = p\{x,\theta(u)\}.$$

Here, $u = (u^a) = (u^1,\ldots,u^m)$ is a vector parameter specifying distributions of M, and defines a coordinate system of M. We assume that $\theta = \theta(u)$ is smooth and its Jacobian matrix has a full rank. Moreover, it is assumed that M forms an m-dimensional submanifold in S. We identify a point $u\in M$ with the point $\theta = \theta(u)$ imbedded in S. The tangent space $T_u(M)$ at u of M is spanned by m vectors ∂_a, $a = 1,\ldots, m$, where $\partial_a = \partial/\partial u^a$ denotes the tangent vector of the coordinate curve u^a in M. The basis ∂_a can be represented by a function $\partial_a\ell(x,u)$ in x as before, where $\ell(x,u) = \log q(x,u)$. Since M is imbedded in S, the tangent space $T_u(M)$ of M is regarded as a subspace of the tangent space $T_{\theta(u)}(S)$ of S at $\theta = \theta(u)$. The basis vector $\partial_a\in T_u(M)$ is written as a linear combination of ∂_i,

$$\partial_a = B^i_a(u)\partial_i,$$

where $B^i_a = \partial\theta^i(u)/\partial u^a$. This can be understood from the relation

$$\partial_a\ell(x,u) = B^i_a\partial_i\ell\{x,\theta(u)\}.$$

Hence, the tangential directions of M at u is represented by m vectors ∂_a, $(a = 1,\ldots,m)$ or $B_a = (B^i_a)$ in the component form with respect to the basis ∂_i of $T_{\theta(u)}(S)$.

It is convenient to define $n - m$ vectors ∂_κ, $\kappa = m + 1,\ldots,n$ in $T_{\theta(u)}(S)$ such that n vectors $\{\partial_a,\partial_\kappa\}$, $a = 1,\ldots,m$; $\kappa = m + 1,\ldots,n$, together form a basis of $T_{\theta(u)}(S)$ and moreover ∂_κ's are orthogonal to ∂_a's, (Fig. 3),

$$g_{a\kappa}(u) = \langle\partial_a,\partial_\kappa\rangle = 0.$$

The vectors ∂_κ span the orthogonal complement of $T_u(M)$ in $T_{\theta(u)}(S)$. We denote the components of ∂_κ with respect to the basis ∂_i by $\partial_\kappa = B_\kappa^i(u)\partial_i$. The inner products of any two basis vectors in $\{\partial_a, \partial_\kappa\}$ are given by

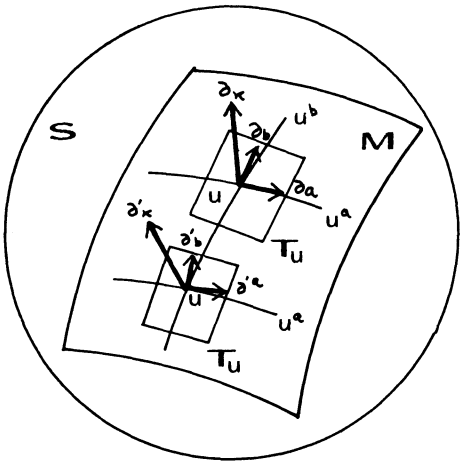

Figure 3

$$g_{ab}(u) = \langle \partial_a, \partial_b \rangle = B_a^i B_b^j g_{ij} ,$$

$$g_{\kappa\lambda}(u) = \langle \partial_\kappa, \partial_\lambda \rangle = B_\kappa^i B_\lambda^k g_{ij} , \qquad (2.4)$$

$$g_{a\kappa}(u) = \langle \partial_a, \partial_\kappa \rangle = B_a^i B_\kappa^j g_{ij} .$$

The basis vector ∂_a may change in its direction as point u moves in M. The change is measured by the α-covariant derivative $\nabla^{(\alpha)}_{\partial_b}\partial_a$ of ∂_a in the direction ∂_b, where the notion of a connection is necessary, because we need to compare two vectors ∂_a and ∂'_a belonging to different tangent spaces $T_{\theta(u)}(S)$ and $T_{\theta(u')}(S)$. The α-covariant derivative $\nabla^{(\alpha)}_{\partial_b}\partial_a$ is calculated in S as

$$\nabla^{(\alpha)}_{\partial_b}\partial_a = B_b^i \nabla^{(\alpha)}_{\partial_i}(B_a^j \partial_j)$$

$$= (\partial_b B_a^j + B_b^i B_a^k \Gamma^{(\alpha)j}_{ik})\partial_j .$$

When the directions of the tangent space $T_u(M)$ of M do not change as point u moves in M, the manifold M is said to be α-flat in S, where the tangent directions are compared by the α-connection. Otherwise, M is curved in the sense of the α-connection. The α-covariant derivative $\nabla^{(\alpha)}_{\partial_b}\partial_a$ is decomposed into the

tangential component belonging to $T_u(M)$ and the normal component perpendicular to $T_u(M)$. The former component represents the way ∂_a changes within $T_u(M)$, while the latter represents the change of ∂_a in the directions perpendicular to $T_u(M)$, as u moves in M. The normal component is measured by

$$H^{(\alpha)}_{ab\kappa} = \langle \nabla^{(\alpha)}_{\partial_a} \partial_b, \partial_\kappa \rangle = (\partial_b B^j_a + B^i_b B^k_a \Gamma^{(\alpha)}_{ik,j}) B^m_\kappa g_{mj}, \tag{2.5}$$

which is a tensor called the α-curvature of submanifold M in S. It is usually called the imbedding curvature or Euler-Shouten curvature. This tensor represents how M is curved in S. A tensor is a multi-linear mapping from a number of tangent vectors to the real set. In the present case, for $A = A^a \partial_a \in T_u(M)$ $B = B^b \partial_b \in T_u(M)$ and $C = C^\kappa \partial_\kappa$ belonging to the orthogonal complement of $T_u(M)$, we have the multi-linear mapping $H^{(\alpha)}$,

$$H^{(\alpha)}(A,B,C) = H^{(\alpha)}_{ab\kappa} A^a B^b C^\kappa.$$

This $H^{(\alpha)}$ is the α-curvature tensor, and $H^{(\alpha)}_{ab\kappa}$ are its components. The submanifold M is α-flat in S when $H^{(\alpha)}_{ab\kappa} = 0$ holds. The $m \times m$ matrix

$$[H^{(\alpha)}_M]^2_{ab} = H^{(\alpha)}_{ac\kappa} H^{(\alpha)}_{bd\lambda} g^{\kappa\lambda} g^{cd}$$

represents the square of the α-curvature of M, where $g^{\kappa\lambda}$ and g^{cd} are the inverse matrix of $g_{\kappa\lambda}$ and g_{cd}, respectively. Efron called the scalar

$$\gamma^2 = [H^{(1)}_M]^2_{ab} g^{ab}$$

the statistical curvature in a one-dimensional model M, which is the trace of the square of the exponential- or 1-curvature of M in our terminology.

Let $\theta = \theta(t)$ be a curve in S parametrized by a scalar t. The curve c: $\theta = \theta(t)$ forms a one-dimensional submanifold in S. The tangent vector ∂_t of the curve is represented in the component form as

$$\partial_t = \dot{\theta}^i(t) \partial_i$$

or shortly by $\dot{\theta}$, where $\cdot$ denotes d/dt. When the direction of the tangent vector $\partial_t = \dot{\theta}$ does not change along the curve in the sense of the α-connection, the curve is called an α-geodesic. By choosing an appropriate parameter, an α-geodesic $\theta(t)$ satisfies the geodesic equation

$$\nabla^{(\alpha)}_{\dot\theta}\dot\theta = 0$$

or in the component form

$$\ddot\theta^i + \Gamma^{(\alpha)i}_{jk}\dot\theta^j\dot\theta^k = 0 \ . \tag{2.6}$$

2.3 Duality in α-flat manifold

Once an affine connection is defined in S, we can compare two tangent vectors $A\varepsilon T_\theta$ and $A'\varepsilon T_{\theta'}$, belonging to different tangent spaces T_θ and $T_{\theta'}$, by the following parallel displacement of a vector. Let c: $\theta = \theta(t)$ be a curve connecting two points θ and θ'. Let us consider a vector field $A(t) = A^i(t)\partial_i\varepsilon T_{\theta(t)}$ defined on each point $\theta(t)$ on the curve. If the vector A(t) does not change along the curve, i.e., the covariant derivative of A(t) in the direction $\dot\theta$ vanishes identically

$$\nabla_{\dot\theta}A(t) = \dot A(t) + \Gamma_{jk}{}^i A^k(t)\dot\theta^j = 0 \ ,$$

the field A(t) is said to be a parallel vector field on c. Moreover, $A(t')\varepsilon T_{\theta(t')}$ at $\theta(t')$ is said to be a parallel displacement of $A(t)\varepsilon T_{\theta(t)}$ at $\theta(t)$. We can thus displace in parallel a vector $A\varepsilon T_\theta$ at θ to another point θ' along a curve $\theta(t)$ connecting θ and θ', by making a vector field A(t) which satisfies the differential equation $\nabla_{\dot\theta}A(t) = 0$, with the boundary conditions $\theta = \theta(0)$, $\theta' = \theta(1)$, and $A(0) = A\varepsilon T_\theta$. The vector $A' = A(1)\varepsilon T_{\theta'}$ at $\theta' = \theta(1)$ is the parallel displacement of A from θ to θ' along the curve c: $\theta = \theta(t)$. We denote it by $A' = \pi_c A$. When the α-connection is used, we denote the α-parallel displacement operator by $\pi^{(\alpha)}_c$. The parallel displacement of A from θ to θ' in general depends on the path c: $\theta(t)$ connecting θ and θ'. When this does not depend on paths, the manifold is said to be flat. It is known that a manifold is flat when, and only when, the Riemann-Christoffel curvature vanishes identically (see textbooks of differential geometry). A statistical manifold S is said to be α-flat, when it is flat under the α-connection.

The parallel displacement does not in general preserve the inner product, i.e., $\langle\pi_c A,\pi_c B\rangle = \langle A,B\rangle$ does not necessarily hold. When a manifold has two affine connections with corresponding parallel displacement operators π_c

and π_C^* and moreover when

$$\langle \pi_C A, \pi_C^* B \rangle = \langle A, B \rangle \tag{2.7}$$

holds, the two connections are said to be mutually dual. The two operators π_C and π_C^* are considered to be mutually adjoint. We have the following theorem in this regard (Nagaoka and Amari (1982)).

Theorem 2.1. The α-connection and $-\alpha$-connection are mutually dual. When S is α-flat, it is also $-\alpha$-flat.

When a manifold S is α-flat, there exists a coordinate system (θ^i) such that

$$\nabla^{(\alpha)}_{\partial_i} \partial_j = 0 \quad \text{or} \quad \Gamma^{(\alpha)}_{ijk}(\theta) = 0$$

identically holds. In this case, a basis vector ∂_i is the same at any point θ in the sense that $\partial_i \varepsilon T_\theta$ is mapped to $\partial_i \varepsilon T_{\theta'}$ by the α-parallel displacement irrespective of the path connecting θ and θ'. Since all the coordinate curves θ^i are α-geodesics in this case, θ is called an α-affine coordinate system. A linear transformation of an α-affine coordinate system is also α-affine.

We give an example of a 1-flat (i.e., $\alpha = 1$) manifold S. The density functions of exponential family $S = \{p(x,\theta)\}$ can be written as

$$p(x,\theta) = \exp\{\theta^i x_i - \psi(\theta)\}$$

with respect to an appropriate measure, where $\theta = (\theta^i)$ is called the natural or canonical parameter. From

$$\partial_i \ell(x,\theta) = x_i - \partial_i \psi(\theta), \qquad \partial_i \partial_j \ell(x,\theta) = -\partial_i \partial_j \psi(\theta) ,$$

we easily have

$$g_{ij}(\theta) = \partial_i \partial_j \psi(\theta), \quad \Gamma^{(\alpha)}_{ijk}(\theta) = \frac{1-\alpha}{2} \partial_i \partial_j \partial_k \psi .$$

Hence, the 1-connection $\Gamma^{(1)}_{ijk}$ vanishes identically in the natural parameter, showing that θ gives a 1-affine coordinate system. A curve $\theta^i(t) = a^i t + b^i$, which is linear in the θ-coordinates, is a 1-geodesic, and conversely.

Since an α-flat manifold is $-\alpha$-flat, there exists a $-\alpha$-flat coordinate system $\eta = (\eta_i) = (\eta_1, \ldots, \eta_n)$ in an α-flat manifold S. Let $\partial^i = \partial/\partial\eta_i$ be the tangent vector of the coordinate curve η_i in the new coordin-

ate system η. The vectors $\{\partial^i\}$ form a basis of the tangent space T_η (i.e. at T_θ where $\theta = \theta(\eta)$) of S. When the two bases $\{\partial_i\}$ and $\{\partial^i\}$ of the tangent space T_θ satisfy

$$\langle \partial_i, \partial^j \rangle = \delta_i^j$$

at every point θ (or η), where δ_i^j is the Kronecker delta (denoting the unit matrix), the two coordinate systems θ and η are said to be mutually dual. (Nagoaoka and Amari (1982)).

Theorem 2.2. When S is α-flat, there exists a pair of coordinate systems $\theta = (\theta^i)$ and $\eta = (\eta_i)$ such that i) θ is α-affine and η is $-\alpha$-affine, ii) θ and η are mutually dual, iii) there exist potential functions $\psi(\theta)$ and $\phi(\eta)$ such that the metric tensors are derived by differentiation as

$$g_{ij}(\theta) = \langle \partial_i, \partial_j \rangle = \partial_i \partial_j \psi(\theta) ,$$

$$g^{ij}(\eta) = \langle \partial^i, \partial^j \rangle = \partial^i \partial^j \phi(\eta) ,$$

where g_{ij} and g^{ij} are mutually inverse matrices so that

$$\partial_i = g_{ij} \partial^j , \quad \partial^i = g^{ij} \partial_j$$

holds, iv) the coordinates are connected by the Legendre transformation

$$\theta^i = \partial^i \psi(\eta), \quad \eta_i = \partial_i \psi(\theta) \tag{2.8}$$

where the potentials satisfy the identity

$$\psi(\theta) + \phi(\eta) - \theta \cdot \eta = 0, \tag{2.9}$$

where $\theta \cdot \eta = \theta^i \eta_i$.

In the case of an exponential family S, ψ becomes the cumulant generating function, the expectation parameter $\eta = (\eta_i)$

$$\eta_i = E_\theta[x_i] = \partial_i \psi(\theta)$$

is -1-affine, θ and η are mutually dual, and the dual potential $\phi(\eta)$ is given by the negative entropy,

$$\phi(\eta) = E[\log p] ,$$

where the expectation is taken with respect to the distribution specified by η.

2.4 α-divergence and α-projection

We can introduce the notion of α-divergence $D_\alpha(\theta,\theta')$ in an α-flat manifold S, which represents the degree of divergence from distribution $p(x,\theta)$ to $p(x,\theta')$. It is defined by

$$D_\alpha(\theta,\theta') = \psi(\theta) + \phi(\eta') - \theta\cdot\eta' \ , \tag{2.10}$$

where $\eta' = \eta(\theta')$ are the η-coordinates of the point θ', i.e., the -α-coordinates of the distribution $p(x,\theta')$. The α-divergence satisfies $D_\alpha(\theta,\theta') \geq 0$ with the equality when and only when $\theta = \theta'$. The -α-divergence satisfies $D_{-\alpha}(\theta,\theta') = D_\alpha(\theta',\theta)$. When S is an exponential family, the -1-divergence is the Kullback-Leibler information,

$$D_{-1}(\theta,\theta') = I[p(x,\theta') : p(x,\theta)] = \int p(x,\theta)\log \frac{p(x,\theta)}{p(x,\theta')}\, dP.$$

As a preview of later discussion, we may also note that, when $S = \{p(x)\}$ is the function space of a non-parametric statistical model, the α-divergence is written as

$$D_\alpha\{p(x),q(x)\} = \frac{4}{1-\alpha^2}\left(1 - \int p(x)^{(1-\alpha)/2}\, q(x)^{(1+\alpha)/2}\, dP\right)$$

when $\alpha \neq \pm 1$, and is the Kullback information or its dual when $\alpha = -1$ or 1.

When θ and $\theta' = \theta + d\theta$ are infinitesimally close,

$$D_\alpha(\theta,\theta + d\theta) = \frac{1}{2} g_{ij}(\theta) d\theta^i d\theta^j \tag{2.11}$$

holds, so that it can be regarded as a generalization of a half of the square of the Riemannian distance, although neither symmetry nor the triangular inequality holds for D_α. However, the following Pythagorean theorem holds (Efron (1978) in an exponential family, Nagaoka and Amari (1982) in a general case).

Theorem 2.3. Let c be an α-geodesic connecting two points θ and θ', and let c' be a -α-geodesic connecting two points θ' and θ'' in an α-flat S. When the two curves c and c' intersect at θ' with a right angle such that θ, θ' and θ'' form a right triangle, the following Pythagorean relation holds,

$$D_\alpha(\theta,\theta') + D_\alpha(\theta',\theta'') = D_\alpha(\theta,\theta'') \ . \tag{2.12}$$

Let $M = \{q(x,u)\}$ be an m-dimensional submanifold imbedded in an α-flat n-dimensional manifold $S = \{p(x,\theta)\}$ by $\theta = \theta(u)$. For a distribution $p(x,\theta_0)\in S$, we search for the distribution $q(x,u)\in M$, which is the closest distribution in M to $p(x,\theta_0)$ in the sense of the α-divergence (Fig. 4a),

$$\min_{u\in M} D_\alpha\{\theta_0,\theta(u)\} = D_\alpha\{\theta_0,\theta(\hat{u})\} \ .$$

We call the resulting $\hat{u}(\theta_0)$ the α-approximation of $p(x,\theta_0)$ in M, assuming such exists uniquely. It is important in many statistical problems to obtain the α-approximation, especially the -1-approximation. Let c(u) be the α-geodesic connecting a point $\theta(u)\in M$ and θ_0, $c(u) : \theta = \theta(t;u)$, $\theta(u) = \theta(0,u)$, $\theta_0 = \theta(1,u)$ (Fig. 4b). When the α-geodesic $c(\tilde{u})$ is orthogonal to M at $\theta(\tilde{u})$, i.e.,

$$\langle \dot{\theta}(0;\tilde{u}), \partial_a \rangle = 0$$

where $\partial_a = \partial/\partial u^a$ are the basis vectors of $T_u(M)$, we call the $\tilde{u}$ the α-projection of θ_0 on M. The existence and the uniqueness of the α-approximation and the α-projection are in general guaranteed only locally. The following theorem was first given by Amari (1982a) and by Nagaoka and Amari (1982) in more general form.

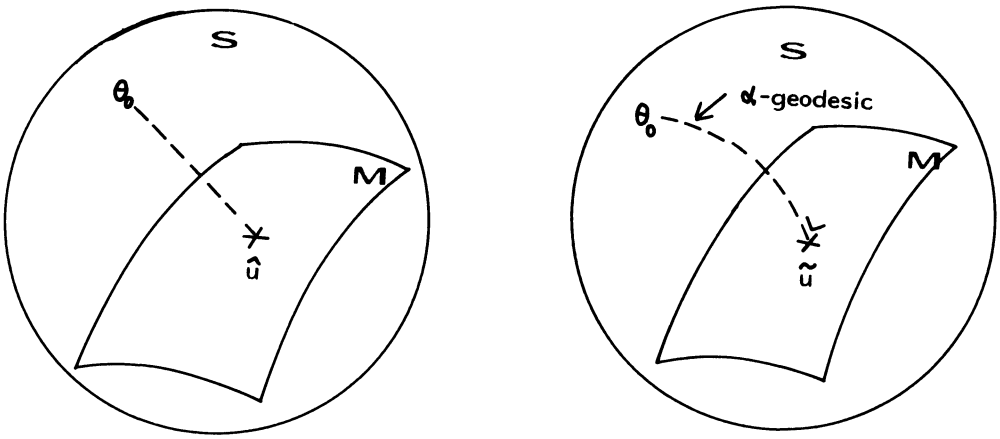

Figure 4

Theorem 2.4. The α-approximation $\hat{u}(\theta_0)$ of θ_0 in M is given by the α-projection $u(\tilde{\theta}_0)$ of θ_0 on M.

3. HIGHER-ORDER ASYMPTOTIC THEORY OF STATISTICAL INFERENCE IN CURVED EXPONENTIAL FAMILY

Ancillary family

Let S be an n-dimensional exponential family parametrized by the natural parameter $\theta = (\theta^i)$ and let $M = \{q(x,u)\}$ be an m-dimensional family parametrized by $u = (u^a)$, $a = 1,\ldots, m$. M is said to be an (n,m)-curved exponential family imbedded in $S = \{p(x,\theta)\}$ by $\theta = \theta(u)$, when $q(x,u)$ is written as

$$q(x,u) = \exp[\theta^i(u)x_i - \psi\{\theta(u)\}].$$

The geometrical structures of S and M can easily be calculated as follows. The quantities in S in the θ-coordinate system are

$$g_{ij}(\theta) = \partial_i\partial_j\psi(\theta) , \qquad \Gamma^{(\alpha)}_{ijk} = \frac{1-\alpha}{2} T_{ijk} ,$$

$$T_{ijk} = \partial_i\partial_j\partial_k\psi(\theta) .$$

The quantities in M are

$$g_{ab}(u) = \langle\partial_a,\partial_b\rangle = B^i_a B^j_b g_{ij} ,$$

$$\Gamma^{(\alpha)}_{abc} = \langle\nabla^{(\alpha)}_{\partial a}\partial_b,\partial_c\rangle = (\partial_a B^i_b)B^j_c g_{ij} + \frac{1-\alpha}{2} T_{abc},$$

$$T_{abc} = B^i_a B^j_b B^k_c T_{ijk} , \qquad B^i_a = \partial_a\theta^i(u) .$$

Here, the basis vector ∂_a of $T_u(M)$ is a vector

$$\partial_a = B^i_a\partial_i$$

in $T_{\theta(u)}(S)$. If we use the expectation coordinate system η in S, M is represented by $\eta = \eta(u)$. The components of the tangent vector ∂_a are given by

$$B_{ai} = \partial_a \eta_i(u) = B_a^j g_{ji} ,$$

where $\partial_a = B_{ai}\partial^i$, $\partial^i = \partial/\partial\eta_i$.

Let $x_{(1)}, x_{(2)},\ldots,x_{(N)}$ be N independent observations from a distribution $q(x,u)\in M$. Then, their arithmetic mean

$$\bar{x} = (\sum_{j=1}^{N} x_{(j)})/N$$

is a minimal sufficient statistic. Since the joint distribution $q(x_{(1)},\ldots, x_{(N)}; u)$ can be written as

$$\prod_{j=1}^{N} q(x_{(j)},u) = \exp[N\{\theta^i(u)\bar{x}_i - \psi\{\theta(u)\}\}],$$

the geometrical structure of M based on N observations is the same as that based on one observation except for a constant factor N. We treat statistical inference based on $\bar{x}$. Since a point x in the sample space X can be identified with a point $\eta = x$ in S by using the expectation parameter η, the observed sufficient statistic $\bar{x}$ defines a point $\hat{\eta}$ in S whose η-coordinates are given by $\bar{x}$, $\hat{\eta} = \bar{x}$. In other words, we regard $\bar{x}$ as the point (distribution) $\hat{\eta}$ in S whose expectation parameter is just equal to $\bar{x}$. Indeed, this $\hat{\eta}$ is the maximum likelihood estimator in the exponential family S.

Let us attach an (n-m)-dimensional submanifold A(u) of S to each point $u\in M$, such that all the A(u)'s are disjoint (at least in some neighborhood of M, which is called a tubular neighborhood) and the union of A(u)'s covers S (at least the tubular neighborhood of M). This is called a (local) foliation of S. Let $v = (v^\kappa)$, $\kappa = m + 1,\ldots, n$ be a coordinate system in A(u). We assume that the pair (u,v) can be used as a coordinate system of the entire S (at least in a neighborhood of M). Indeed, a pair (u,v) specifies a point in S such that it is included in the A(u) attached to u and its position in A(u) is given by v (see Fig. 5). Let $\eta = \eta(u,v)$ be the η-coordinates of the point specified by (u,v). This is the coordinate transformation of S from $w = (u,v)$ to η, where $w = (u,v) = (w^\beta)$ is an n-dimensional variable, $\beta = 1,\ldots, n$, such that its first m components are $u = (u^a)$ and the last n - m components are $v = (v^\kappa)$.

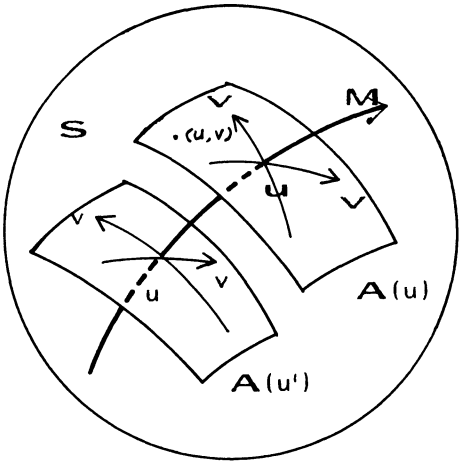

Figure 5

Any point η (in some neighborhood of M) in S can be represented uniquely by $w = (u,v)$. We assume that A(u) includes the point $\eta = \eta(u)$ on M and that the origin $v = 0$ of A(u) is put at the point $u \in M$. This implies that $\eta(u,0)$ is the point $\eta(u) \in M$. We call $A = \{A(u)\}$ an ancillary family of the model M.

In order to analyze the properties of a statistical inference method, it is helpful to use the ancillary family which is naturally determined by the inference method. For example, an estimator $\hat{u}$ can be regarded as a mapping from S to M such that it maps the observed point $\hat{\eta} = \bar{x}$ in S determined by the sufficient statistic $\bar{x}$ to a point $\hat{u}(\bar{x}) \in M$. Its inverse image $\hat{u}^{-1}(u)$ defines an (n-m)-dimensional subspace A(u) attached to $u \in M$,

$$A(u) = \hat{u}^{-1}(u) = \{\eta \in S \mid \hat{u}(\eta) = u\} .$$

Obviously, the estimator $\hat{u}$ takes the value u when and only when the observed $\bar{x}$ is included in A(u). These A(u)'s form a family $A = \{A(u)\}$ which we will call the ancillary family associated with the estimator $\hat{u}$. As will be shown soon, large-sample properties of an estimator $\hat{u}$ are determined by the geometrical features of the associated ancillary submanifolds A(u). Similarly, a test T can be regarded as a mapping from S to the binary set $\{r,\bar{r}\}$, where r and $\bar{r}$ imply, respectively, rejection and acceptance of a null hypothesis. The

inverse image $T^{-1}(r) \subset S$ is called the critical region, and the hypothesis is rejected when and only when the observed point $\hat{\eta} = \bar{x} \in S$ is in $T^{-1}(r)$. In order to analyze the characteristics of a test, it is convenient to use an ancillary family $A = \{A(u)\}$ such that the critical region is composed of some of the $A(u)$'s and the acceptance region is composed of the other $A(u)$'s. Such an ancillary family is said to be associated with the test T.

In order to analyze the geometrical features of ancillary submanifolds, let us use the new coordinate system $w = (u,v)$. The tangent of the coordinate curve w^{β} is given by $\partial_{\beta} = \partial/\partial w^{\beta}$. The tangent space $T_{\eta}(S)$ at point $\eta = \eta(w)$ of S is spanned by $\{\partial_{\beta}\}$, $\beta = 1,\ldots, n$. They are decomposed into two parts $\{\partial_{\beta}\} = \{\partial_a, \partial_{\kappa}\}$, $\beta = 1,\ldots, n$; $a = 1,\ldots, m$; $\kappa = m+1,\ldots,n$. The former part $\partial_a = \partial/\partial u^a$ spans the tangent space $T_u(M)$ of M at u and the latter $\partial_{\kappa} = \partial/\partial v^{\kappa}$ spans the tangent space $T_u(A)$ of $A(u)$. Their components are given by $B_{\beta i} = \partial_{\beta}\eta_i(w)$ in the basis ∂^i. They are decomposed as

$$\partial_a = B_{ai}\partial^i \quad , \quad \partial_{\kappa} = B_{\kappa i}\partial^i \quad ,$$

with $B_{ai} = \partial_a \eta_i(u,v)$, $B_{\kappa i} = \partial_{\kappa}\eta_i(u,v)$. The metric tensor in the w-coordinate system is given by

$$g_{\alpha\beta} = \langle \partial_{\alpha}, \partial_{\beta} \rangle = B_{\alpha i} B_{\beta j} g^{ij} = B^i_{\alpha} B^j_{\beta} g_{ij} \tag{3.1}$$

where

$$B^i_{\alpha} = g^{ij} B_{\alpha j} = \partial\theta^i(u,v)/\partial w^{\alpha} \quad .$$

The metric tensor is decomposed into three parts:

$$g_{ab}(u) = \langle \partial_a, \partial_b \rangle = B_{ai} B_{bj} g^{ij} \tag{3.2}$$

is the metric tensor of M,

$$g_{\kappa\lambda}(u) = \langle \partial_{\kappa}, \partial_{\lambda} \rangle = B_{\kappa i} B_{\lambda j} g^{ij} \tag{3.3}$$

is the metric tensor of $A(u)$, and

$$g_{a\kappa} = \langle \partial_a, \partial_{\kappa} \rangle = B_{ai} B_{\kappa j} g^{ij} \tag{3.4}$$

represents the angles between the tangent spaces of M and $A(u)$. When $g_{a\kappa}(u,0) = 0$, M and $A(u)$ are orthogonal to each other at M. The ancillary family

$A = \{A(u)\}$ is said to be orthogonal, when $g_{a\kappa}(u) = 0$, where $f(u)$ is the abbreviation of $f(u,0)$ when a quantity $f(u,v)$ is evaluated on M, i.e., at $v = 0$. We may treat an ancillary family A_N which depends on the number N of observations. In this case $g_{\alpha\beta}$ also depends on N. When $g_{a\kappa} = \langle \partial_a, \partial_\kappa \rangle$ is a quantity of order $N^{-1/2}$ converging to 0 as N tends to infinity, the ancillary family is said to be asymptotically orthogonal.

The α-connection in the w-coordinate system is given by

$$\Gamma^{(\alpha)}_{\alpha\beta\gamma} = \langle \nabla^{(\alpha)}_{\partial_\alpha} \partial_\beta, \partial_\gamma \rangle = (\partial_\alpha B_{\beta i}) B^i_\gamma - \frac{1+\alpha}{2} T_{\alpha\beta\gamma}$$

$$= (\partial_\alpha B^i_\beta) B_{\gamma i} + \frac{1-\alpha}{2} T_{\alpha\beta\gamma} , \tag{3.5}$$

where $T_{\alpha\beta\gamma} = B^i_\alpha B^j_\beta B^k_\gamma T_{ijk}$. The M-part $\Gamma^{(\alpha)}_{abc}$ gives the components of the α-connection of M and the A-part $\Gamma^{(\alpha)}_{\kappa\lambda\mu}$ gives those of the α-connection of A(u). When A is orthogonal, the α-curvatures of M and A(u) are given respectively by

$$H^{(\alpha)}_{ab\kappa} = \Gamma^{(\alpha)}_{ab\kappa} , \qquad H^{(\alpha)}_{\kappa\lambda a} = \Gamma^{(\alpha)}_{\kappa\lambda a} . \tag{3.6}$$

The quantities $g_{a\kappa}(u)$, $H^{(\alpha)}_{ab\kappa}$ and $H^{(\alpha)}_{\kappa\lambda a}$ are fundamental in evaluating asymptotic properties of statistical inference procedures. When $\alpha = 1$, the 1-connection is called the exponential connection, and we use suffix (e) instead of (1). When $\alpha = -1$, the -1-connection is called the mixture connection, and we use suffix (m) instead of (-1).

3.2 Edgeworth expansion

We study higher-order asymptotic properties of various statistics with the help of Edgeworth expansions. To this end, let us express the point $\hat{\eta} = \bar{x}$ defined by the observed sufficient statistic in the w-coordinate system. The w-coordinates $\hat{w} = (\hat{u},\hat{v})$ are obtained by solving

$$\bar{x} = \eta(\hat{w}) = \eta(\hat{u},\hat{v}) . \tag{3.7}$$

The sufficient statistic $\bar{x}$ is thus decomposed into two parts $(\hat{u},\hat{v})$ which together are also sufficient. When the ancillary family A is associated with an estimator or a test, $\hat{u}$ gives the estimated value or the test statistic,

respectively. We calculate the Edgeworth expansion of the joint distribution of $(\hat{u},\hat{v})$ in geometrical terms. Here, it is necessary further to assume a condition which guarantees the Edgeworth expansion. We assume that Cramér's condition is satisfied. See, for example, Bhattacharya and Ghosh (1978).

When u_0 is the true parameter of distribution, $\bar{x}$ converges to $\eta(u_0, 0)$ in probability as the number N of observations tends to infinity, so that the random variable $\hat{w}$ also converges to $w_0 = (u_0,0)$. Let us put

$$\tilde{x} = \sqrt{N}\{\bar{x} - \eta(u_0,0)\} , \quad \tilde{w} = \sqrt{N}(\hat{w} - w_0) ,$$

$$\tilde{u} = \sqrt{N}(\hat{u} - u_0) , \quad \tilde{v} = \sqrt{N}\,\hat{v} . \tag{3.8}$$

Then, by expanding (3.7), we can express $\tilde{w}$ in the power series of $\tilde{x}$. We can obtain the Edgeworth expansion of the distribution $p(\tilde{w};u_0)$ of $\tilde{w} = (\tilde{u},\tilde{v})$. However, it is simpler to obtain the distribution of the one-step bias-corrected version $\tilde{w}^*$ of $\tilde{w}$ defined by

$$\tilde{w}^* = \tilde{w} - E_{\hat{w}}[\tilde{w}] ,$$

where E_w denotes the expectation with respect to $p(x,w)$. The distribution of $\tilde{w}$ is obtained easily from that of $\tilde{w}^*$. (See Amari and Kumon (1983).)

Theorem 3.1. The Edgeworth expansion of the probability density $p(\tilde{w}^*,u_0)$ of $\tilde{w}^*$, where $q(x,u_0)$ is the underlying true distribution, is given by

$$p(\tilde{w}^*,u_0) = n(\tilde{w}^*;g_{\alpha\beta})\{1 + \frac{1}{6\sqrt{N}} K_{\alpha\beta\gamma}h^{\alpha\beta\gamma} + \frac{1}{N} A_N(\tilde{w}^*) + O(N^{-3/2})\} ,$$

$$A_N(\tilde{w}^*) = \frac{1}{4} C^2_{\alpha\beta}h^{\alpha\beta} + \frac{1}{24} K_{\alpha\beta\gamma\delta}h^{\alpha\beta\gamma\delta} + \frac{1}{72} K_{\alpha\beta\gamma}K_{\delta\varepsilon\phi}h^{\alpha\beta\gamma\delta\varepsilon\phi} , \tag{3.9}$$

where $n(\tilde{w}^*;g_{\alpha\beta})$ is the multivariate normal density with mean 0 and covariance $g^{\alpha\beta} = (g_{\alpha\beta})^{-1}$, $h^{\alpha\beta\gamma}$ etc. are the tensorial Hermite polynomials in $\tilde{w}^*$ and

$$K_{\alpha\beta\gamma} = -3\Gamma^{(-1/3)}_{\alpha\beta\gamma} ,$$

$$C^2_{\alpha\beta} = \Gamma^{(m)}_{\gamma\delta\alpha}\,\Gamma^{(m)}_{\varepsilon\sigma\beta}\,g^{\gamma\varepsilon}g^{\delta\sigma} , \text{ etc.}$$

The tensorial Hermite polynomials in w with metric $g_{\alpha\beta}$ are defined by

$$h^{\alpha_1\cdots\alpha_k}(w) = (-1)^k\{D^{\alpha_1}\cdots D^{\alpha_k} n(w;g_{\alpha\beta})\}/n(w;g_{\alpha\beta}) ,$$

where $D^\alpha = g^{\alpha\beta}(\partial/\partial w^\beta)$, cf. Amari and Kumon (1983), McCullagh (1984). Hence,

$$h^0 = 1, \quad h^\alpha = w^\alpha, \quad h^{\alpha\beta} = w^\alpha w^\beta - g^{\alpha\beta} ,$$

$$h^{\alpha\beta\gamma} = w^\alpha w^\beta w^\gamma - g^{\alpha\beta}w^\gamma - g^{\alpha\gamma}w^\beta - g^{\beta\gamma}w^\alpha , \text{ etc.}$$

Theorem 3.1 shows the Edgeworth expansion up to order N^{-1} of the joint distribution of $\tilde{u}^*$ and $\tilde{v}^*$, which together carry the full Fisher information. The marginal distribution can easily be obtained by integration.

Theorem 3.2. When the ancillary family is orthogonal, i.e., $g_{a\kappa}(u) = 0$, the distribution $p(\tilde{u}^*,u_0)$ of $\tilde{u}^*$ is given by

$$p(u^*,u_0) = n(u^*;g_{ab})\{1 + \frac{1}{6} N^{-1/2}K_{abc}h^{abc} + N^{-1}A_N(\tilde{u}^*)\} + O(N^{-3/2}) , \tag{3.10}$$

where $K_{abc} = -3\Gamma^{(-1/3)}_{abc}$,

$$A_N(u^*) = \frac{1}{4} C^2_{ab}h^{ab}$$

+ terms common to all the orthogonal ancillary families,

$$C^2_{ab} = (\Gamma^m)^2_{ab} + 2(H^e_M)^2_{ab} + (H^m_A)^2_{ab} , \tag{3.11}$$

$$(\Gamma^m)^2_{ab} = \Gamma^{(m)}_{cda}\Gamma^{(m)}_{efb}\, g^{ce}g^{df} ,$$

$$(H^e_M)^2_{ab} = H^{(e)}_{ac\kappa}\, H^{(e)}_{bd\lambda}\, g^{cd}g^{\kappa\lambda} ,$$

$$(H^m_A)^2_{ab} = H^{(m)}_{\kappa\nu a}\, H^{(m)}_{\lambda\mu b}\, g^{\kappa\lambda}g^{\nu\mu} .$$

3.3 Higher-order efficiency of estimation

Given an estimator $\hat{u} : S\to M$ which maps the observed point $\hat{\eta} = \bar{x}\in S$ to $\hat{u}(\bar{x})\in M$, we can construct the ancillary family $A = \{A(u)\}$ by

$$A(u) = \hat{u}^{-1}(u) = \{\eta\in S \mid \hat{u}(\eta) = u\} .$$

The $A(u)$ includes the point $\eta(u) = \eta(u,0)$, when and only when the estimator is consistent. (We may treat a case when $A(u)$ depends on N, denoting an ancillary

family by $A_N(u)$. In this case, an estimator is consistent if $\lim_{N\to\infty} A_N(u)\to n(u,0)$.) Let us expand the covariance of the estimation error $\tilde{u} = \sqrt{N}(\hat{u} - u_0)$ as

$$\mathrm{cov}[\tilde{u}^a,\tilde{u}^b] = g_1^{ab} + g_2^{ab}N^{-1/2} + g_3^{ab}N^{-1} + O(N^{-3/2}) \ .$$

A consistent estimator is said to be first-order efficient or simply efficient, when its first-order term $g_1^{ab}(u)$ is minimal among all the consistent estimators at any u, where the minimality is in the sense of positive semidefiniteness of matrices. The second- and third-order efficiency is defined similarly.

Since the first-order term g_1^{ab} is given from (3.9) by

$$g_1^{ab} = (g_{ab} - g_{a\kappa}g_{b\lambda}g^{\kappa\lambda})^{-1} \ ,$$

the minimality is attained, when and only when $g_{a\kappa} = 0$, i.e., the associated ancillary family is orthogonal. From this and Theorem 3.2, we have the following results.

Theorem 3.3. A consistent estimator is first-order efficient, iff the associated ancillary family is orthogonal. An efficient estimator is always second-order efficient, because of $g_2^{ab} = 0$.

There exist no third-order efficient estimators in the sense that $g_3^{ab}(u)$ is minimal at all u. This can be checked from the fact that g_3^{ab} includes a term linear in the derivative of the mixture curvature of A(u), see Amari (1985). However, if we calculate the covariance of the bias-corrected version $\tilde{u}^* = \tilde{u} - E_{\hat{u}}[\tilde{u}]$ of an efficient estimator $\hat{u}$, we see that there exists the third-order efficient estimator among the class of all the bias-corrected efficient estimators. To state the result, let $g_{3ab} = g_3^{cd}g_{ca}g_{bd}$ be the lower index version of g_3^{ab}.

Theorem 3.4. The third-order term g_{3ab} of the covariance of a bias-corrected efficient estimator $\tilde{u}^*$ is given by the sum of the three non-negative geometric quantities

$$g_{3ab} = \frac{1}{2}(\Gamma^m)^2_{ab} + (H^e_M)^2_{ab} + \frac{1}{2}(H^m_A)^2_{ab} \ . \qquad (3.12)$$

The first is the square of mixture connection components of M, and depends on the parametrization of M but is common to all the estimators. The second is the square of the exponential curvature of M, which does not depend on the estimator. The third is the square of the mixture curvature of the ancillary submanifold A(u) at $\eta(u)$, which depends on the estimator. An efficient estimator is third-order efficient, when and only when the associated ancillary family is mixture-flat at $\eta(u)$. The m.l.e. is third-order efficient, because it is given by the mixture-projection of $\hat{\eta}$ to M.

The Edgeworth expansion (3.10) tells more about the characteristics of an efficient estimator $\tilde{u}^*$. When $H^{(m)}_{\kappa\lambda a}$ vanishes, an estimator is shown to be mostly concentrated around the true parameter u and is third-order optimal under a symmetric unimodal loss function. The effect of the manner of parametrizing M is also clear from (3.10). The α-normal coordinate system (parameter) in which the components of the α-connection become zero at a fixed point is very important (cf. Hougaard, 1983; Kass, 1984).

3.4 Higher-order efficiency of tests

Let us consider a test T of a null hypothesis $H_0 : u \in D$ against the alternative $H_1 : u \notin D$ in an (n,m)-curved exponential family, where D is a region or a submanifold in M. Let R be a critical region of test T such that the hypothesis H_0 is rejected when and only when the observed point $\hat{\eta} = \bar{x}$ belongs to R. When T has a test statistic $\lambda(\bar{x})$, the equation $\lambda(\eta) = \text{const.}$ gives the boundary of the critical region R. The power function $P_T(u)$ of the test T at point u is given by

$$P_T(u) = \int_R \bar{p}(\bar{x};u)\, d\bar{x} ,$$

where $\bar{p}(\bar{x};u)$ is the density function of $\bar{x}$ when the true parameter is u.

Given a test T, we can compose an ancillary family $A = \{A(u)\}$ such that the critical region R is given by the union of some of A(u)'s, i.e., it can be written as

$$R = \underline{U}_{u \in R_M} A(u) ,$$

where R_M is a subset of M. Then, when we decompose the observed statistic $\hat{\eta} = \bar{x}$ into $(\hat{u},\hat{v})$ by $\bar{x} = \eta(\hat{u},\hat{v})$ in terms of the related w-coordinates, the hypothesis H_0 is rejected when and only when $\hat{u}\varepsilon R_M$. Hence, the test statistics $\lambda(\bar{x})$ is a function of only $\hat{u}$. Since we have already obtained the Edgeworth expansion of the joint distribution of $(\hat{u},\hat{v})$ or of $(\hat{u}^*,\hat{v}^*)$, we can analyze the characteristics of a test in terms of geometry of associated A(u)'s.

We first consider the case where $M = \{q(x,u)\}$ is one-dimensional, so that $u = (u^a)$ is a scalar parameter, indices a, b, etc becoming equal to 1. We test the null hypothesis $H_0 : u = u_0$ against the alternative $H_1 : u \neq u_0$. Let u_t be a point which approaches u_0 as N tends to infinity by

$$u_t = u_0 + t(Ng)^{-1/2} , \tag{3.13}$$

i.e., the point whose Riemannian distance from u_0 is approximately $tN^{-1/2}$, where $g = g_{ab}(u_0)$. The power $P_T(u_t,N)$ of a test T at u_t is expanded as

$$P_T(u_t,N) = P_{T1}(t) + P_{T2}(t)N^{-1/2} + P_{T3}(t)N^{-1} + O(N^{-3/2}) .$$

A test T is said to be first-order uniformly efficient or, simply, efficient, if the first-order term $P_{T1}(t)$ satisfies $P_{T1}(t) \geq P_{T'1}(t)$ at all t, compared with any other test T' of the same level. The second- and third-order uniform efficiency is defined similarly. Let $P(u_t,N)$ be the envelope power function of $P_T(u_t,N)$'s defined by

$$P(u_t,N) = \sup_T P_T(u_t,N) . \tag{3.14}$$

Let us expand it as

$$P(u_t,N) = P_1(t) + P_2(t)N^{-1/2} + P_3(t)N^{-1} + O(N^{-3/2}) .$$

It is clear that a test T is i-th order uniformly efficient, iff

$$P_{Tk}(t) = P_k(t)$$

holds at any t for $k = 1,\ldots,i$.

An ancillary family $A = \{A(u)\}$ in this case consists of (n-1)-dimensional submanifolds A(u) attached to each u or $\eta(u)\varepsilon M$. The critical region R is bounded by one of the ancillary submanifolds, say $A(u_+)$, in the

one-sided case, and by two submanifolds $A(u_+)$ and $A(u_-)$ in the two-sided unbiased case. The asymptotic behavior of a test T is determined by the geometric features of the boundary ∂R, i.e., $A(u_+)$[and $A(u_-)$]. In particular, the angle between M and A(u) is important. The angle is given by the inner product $g_{a\kappa}(u) = \langle \partial_a, \partial_\kappa \rangle$ of the tangent ∂_a of M and tangents ∂_κ of A(u). When $g_{a\kappa}(u) = 0$ for all u, A is orthogonal. In the case of a test, the critical region and hence the associated ancillary A and $g_{a\kappa}(u)$ depend on N. An ancillary family is said to be asymptotically orthogonal, when $g_{a\kappa}(u)$ is of order $N^{-1/2}$. We can assume $g_{a\kappa}(u_0) = 0$, and $g_{a\kappa}(u_t)$ can be expanded as

$$g_{a\kappa}(u_t) = t\, Q_{ab\kappa}(Ng)^{-1/2} , \tag{3.15}$$

where $Q_{ab} = \partial_a g_{b\kappa}(u_0)$. The quantity $Q_{ab\kappa}$ represents the direction and the magnitude of inclination of A(u) from being exactly orthogonal to M. We can now state the asymptotic properties of a test in geometrical terms (Kumon and Amari (1983), (1985)).

Theorem 3.5. A test T is first-order uniformly efficient, iff the associated ancillary family A is asymptotically orthogonal. A first-order uniformly efficient test is second-order uniformly efficient.

Unfortunately, there exist no third-order uniformly efficient test (unless the model M is exponential family). An efficient test T is said to be third-order t_0-efficient, when its third-order power $P_{T3}(t)$ is minimal among all the other efficient tests at t_0, i.e., when $P_{T3}(t_0) = P_3(t_0)$, and when there exist no tests T' satisfying $P_{T'3}(t) \geq P_{T3}(t)$ for all t. An efficient test is third-order admissible, when it is t_0 - efficient at some t_0. We define the third-order power loss function (deficiency function) $\Delta P_T(t)$ of an efficient test T by

$$P_T(t) = \lim_{N\to\infty} N\{P(u_t,N) - P_T(u_t,N)\} = P_3(t) - P_{T3}(t) . \tag{3.16}$$

It characterizes the behaviors of an efficient test T. The power loss function can be explicitly given in geometrical terms of the associated ancillary A (Kumon and Amari (1983), Amari (1983a)).

Theorem 3.6. An efficient test T is third-order admissible, only when the mixture curvature of A(u) vanishes as $N\to\infty$ and the A(u) is not exactly orthogonal to M but asymptotically orthogonal to compensate the exponential curvature $H^{(e)}_{ab}$ of model M such that

$$Q_{ab\kappa} = cH^{(e)}_{ab\kappa} \tag{3.17}$$

holds for some constant c. The third-order power loss function is then given by

$$\Delta P_T(t) = a_i(t,\alpha)\{c - J_i(t,\alpha)\}^2\gamma^2 , \tag{3.18}$$

where $a_i(t,\alpha)$ is some fixed function of t and α, α being the level of the test,

$$\gamma^2 = g^{\kappa\lambda}H^{(e)}_{ab\kappa}\, H^{(e)}_{cd\lambda}\, g^{ac}g^{bd} \tag{3.19}$$

is the square of the exponential curvature (Efron's curvature) of M, and

$$J_1(t,\alpha) = 1 - t/\{2u_1(\alpha)\},$$

$$J_2(t,\alpha) = 1 - t/[2u_2(\alpha)\tanh\{tu_2(\alpha)\}],$$

i = 1 for the one-sided case and i = 2 for two-sided case, n being the standard normal density function, and $u_1(\alpha)$ and $u_2(\alpha)$ being the one-sided and two-sided $100\alpha\%$ points of the normal density, respectively.

The theorem shows that a third-order admissible test is characterized by its c value. It is interesting that the third-order power loss function (3.18) depends on the model M only through the statistical curvature γ^2, so that $\Delta P_T(t)/\gamma^2$ gives a universal power loss curve common to all the statistical models. It depends only on the value of c. Various widely used tests will next be shown to be third-order admissible, so that they are characterized by c values as follows.

Theorem 3.7. The test based on the maximum likelihood estimator (e.g. Wald test) is characterized by c = 0. The likelihood ratio test is characterized by c = 1/2. The locally most powerful test is characterized by c = 1 in the one-sided case and $c = 1 - 1/\{2u_2^2(\alpha)\}$ in the two-sided case. The conditional test conditioned on the approximate ancillary statistic $a = H^{(e)}_{ab\kappa}\hat{v}^{\kappa}$ is characterized also by c = 1/2. The efficient-score test is characterized by

c = 1, and is inadmissible in the two-sided case.

We show the universal third-order power loss functions of various tests in Fig. 6 in the two-sided case and in Fig. 7 in the one-sided case, where α = 0.05 (from Amari (1983a)). It is shown that the likelihood ratio test has fairly good performances throughout a wide range of t, while the locally most powerful test behaves badly when $t \geq 2$. The m.l.e. test is good at around $t = 3 \sim 4$.

We can generalize the present theory to the multi-parameter cases with and without nuisance parameters. It is interesting that none of the above tests are third-order admissible in the multi-parameter case. However, it is easy to modify a test to get a third-order t_0-efficient test by the use of the asymptotic ancillary statistic a (Kumon and Amari, 1985). We can also design the third-order t_0-most-powerful confidence region estimators and the third-order minimal size confidence region estimators.

It is also possible to extend the present results of estimation and testing in a statistical model with nuisance parameter ξ. In this case, a set $M(u_0)$ of distributions in which the parameter of interest takes a fixed value u_0, but ξ takes arbitrary values, forms a submanifold. The mixture curvature and the exponential twister curvature of $M(u_0)$ are responsible for the higher-order characteristics of statistical inference. The third-order admissibility of the likelihood ratio test and others is again proved. See Amari (1985).

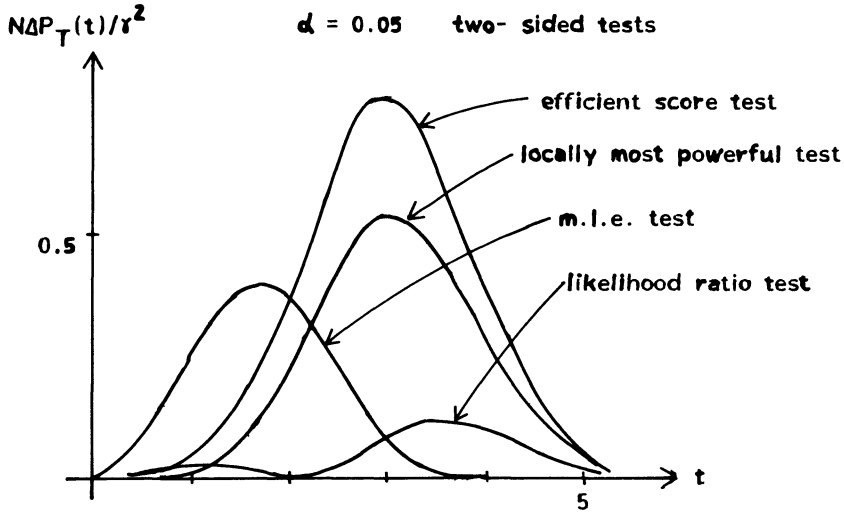

Figure 6

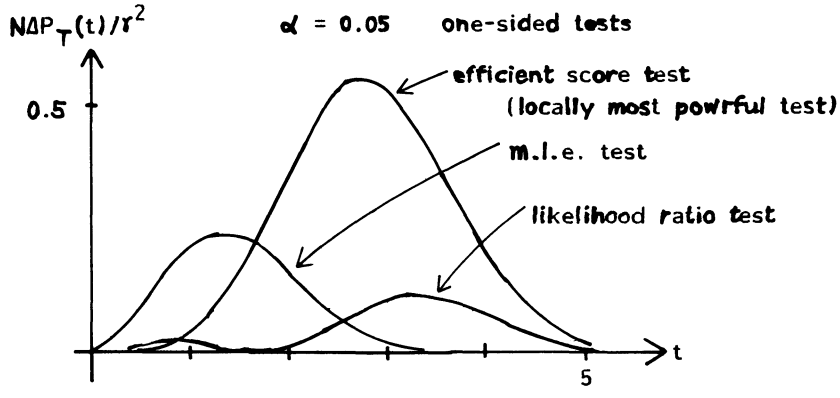

Figure 7

4. INFORMATION, SUFFICIENCY AND ANCILLARITY

HIGHER ORDER THEORY

Information and conditional information

Given a statistical model $M = \{p(x,u)\}$, $u = (u^a)$, we can follow Fisher and define the amount $g_{ab}(T)$ of information included in a statistic $T = t(x)$ by

$$g_{ab}(T) = E[\partial_a \ell(t,u)\partial_b \ell(t,u)] \ , \tag{4.1}$$

where $\ell(t,u)$ is the logarithm of the density function of t when the true parameter is u. The information $g_{ab}(T)$ is a positive-semidefinite matrix depending on u. Obviously, for the statistic X, $g_{ab}(X)$ is the Fisher information matrix. Let T(X) and S(X) be two statistics. We similarly define, by using the joint distribution of T and S, the amount $g_{ab}(T,S)$ of information which T and S together carry. The additivity

$$g_{ab}(T,S) = g_{ab}(T) + g_{ab}(S)$$

does not hold except when T and S are independent. We define the amount of conditional information carried by T when S is known by

$$g_{ab}(T|S) = E_S E_{T|S}[\partial_a \ell(t|s,u)\partial_b \ell(t|s,u)] \ , \tag{4.2}$$

where $\ell(t|s,u)$ is the logarithm of the conditional density function of T conditioned on S. Then, the following relation holds,

$$g_{ab}(T,S) = g_{ab}(T) + g_{ab}(S|T) = g_{ab}(S) + g_{ab}(T|S) \ .$$

From $g_{ab}(S|T) = g_{ab}(T,S) - g_{ab}(T)$, we see that the conditional information denotes the amount of loss of information when we discard s from a pair of statistics s and t, keeping only t. Especially,

$$\Delta g_{ab}(T) = g_{ab}(X) - g_{ab}(T) = g_{ab}(X|T) \tag{4.3}$$

is the amount of loss of information when we keep only t(x) instead of keeping the original x. The following relation is useful for calculation,

$$\Delta g_{ab}(T) = E_T \mathrm{Cov}[\partial_a \ell(x,u), \partial_b \ell(x,u)|t] , \tag{4.4}$$

$$g_{ab}(S|T) = g_{ab}(T) - g_{ab}(T,S) , \tag{4.5}$$

where Cov[.|t] is the conditional covariance.

A statistic S is sufficient, when $g_{ab}(S) = g_{ab}(X)$ or $\Delta g_{ab}(S) = 0$. When S is sufficient, $g_{ab}(T|S) = 0$ holds for any statistic T. A statistic a is ancillary, when $g_{ab}(A) = 0$. When A is ancillary, $g_{ab}(T,A) = g_{ab}(T|A)$ for any T. It is interesting that, although A itself has no information, A together with another statistic T recovers the amount

$$g_{ab}(A|T) = g_{ab}(T,A) - g_{ab}(T)$$

of information. An ancillary statistic carries some information in this sense, and this is the reason why an ancillarity is important in statistical inference. We call $g_{ab}(A|T)$ the amount of information of ancillary A relative to statistic T.

When N independent observations $x_1,\ldots,x_N$ are available, the Fisher information $g_{ab}(X^N)$ is $Ng_{ab}(X)$, N times that of one observation. When M is a curved exponential family, $\bar{x} = \Sigma x_i/N$ is a sufficient statistic, keeping the whole information, $g_{ab}(\bar{X}) = Ng_{ab}(X)$. Let $t(\bar{x})$ be a statistic which is a function of $\bar{x}$. It is said to be asymptotically sufficient of order q, when

$$\Delta g_{ab}(T) = g_{ab}(\bar{X}) - g_{ab}(T) = O(N^{-q+1}) . \tag{4.6}$$

Similarly, a statistic $t(\bar{x})$ is said to be asymptotically ancillary of order q, when

$$g_{ab}(T) = O(N^{-q}) \tag{4.7}$$

holds. (The definition of the order in the present article is different from that by Cox (1980) etc.)

4.2 Asymptotic efficiency and ancillarity

Given a consistent estimator $\hat{u}(\bar{x})$ in an (n,m)-curved exponential family M, we can construct the associated ancillary family A. By introducing an adequate coordinate system v in each A(u), the sufficient statistic $\bar{x}$ is decomposed into two statistics $(\hat{u},\hat{v})$ by $\bar{x} = \eta(\hat{u},\hat{v})$. The amount $\Delta g_{ab}(\hat{U})$ of information loss of estimator $\hat{u}$ is calculated from (4.4) by using the stochastic expansion of $\partial_a \ell(\bar{x},u)$ as

$$\Delta g_{ab}(\hat{U}) = N g_{a\kappa} g_{b\kappa} g^{\kappa\lambda} + O(1) .$$

Hence, when and only when A is orthogonal, i.e., $g_{a\kappa}(u) = 0$, $\hat{u}$ is first-order sufficient. In this case, $\hat{u}$ is (first-order) efficient. The loss of information of an efficient estimator $\hat{u}$ is calculated as

$$\Delta g_{ab}(\hat{U}) = (H^e_M)^2_{ab} + (1/2)(H^m_A)^2_{ab} + O(N^{-1}) , \tag{4.8}$$

where $(H^e_M)^2$ is the square of the exponential curvature of the model M and $(H^m_A)^2$ is the square of the mixture curvature of the associated ancillary family A at v = 0. Hence, the loss of information is minimized uniformly in u, iff the mixture curvature of the associated ancillary family A(u) vanishes at v = 0 for all u. In this case, the estimator $\hat{u}$ is third-order efficient in the sense of the covariance in §3. The m.l.e. is such a higher-order efficient estimator.

Among all third-order efficient estimators, does there exist one whose loss of information is minimal at all u up to the term of order N^{-1}? Is the m.l.e. such a one? This problem is related to the asymptotic efficiency of estimators of order higher than three. By using the Edgeworth expansion (3.9) and the stochastic expansion of $\partial_a \ell(\bar{x},u)$, we can calculate the terms, which depend on the estimator, of the information loss of order N^{-1} in geometrical terms of the related ancillary family. The loss of order N^{-1} includes a term related to the derivatives of the mixture curvature $H^{(m)}_{\kappa\lambda a}$ of A in the direction of ∂_μ and ∂_a (unpublished note). From this formula, one can conclude that there exist no estimators whose loss $\Delta g_{ab}(\hat{U})$ of information is minimal up to the term of order N^{-1} at all u among all other estimators. Hence, the loss of information of the m.l.e. is not uniformly minimal at all u, when the loss is

evaluated up to the term of order N^{-1}.

We have already obtained the Edgeworth expansion up to order N^{-1} of the joint distribution of $(\hat{u},\hat{v})$, or equivalently $(\tilde{u}^*,\tilde{v}^*)$ in (3.9). By integration, we have the distribution of $\tilde{v}^*$,

$$p(\tilde{v}^*;u) = n(\tilde{v}^*;g_{\kappa\lambda})\{1 + \frac{1}{6}\sqrt{N}\, K_{\kappa\lambda\mu}h^{\kappa\lambda\mu} + O(N^{-1})\}, \tag{4.9}$$

where $g_{\kappa\lambda}(u)$ and $K_{\kappa\lambda\mu}(u)$ depend on the coordinate system v introduced to each A(u). The information $g_{ab}(\hat{V}^*)$ of $\tilde{v}^*$ can be calculated from this. It depends on the coordinate system v, too. It is always possible to choose a coordinate system v in each A(u) such that $\{\partial_\kappa\}$ is an orthonormal system at v = 0, i.e., $g_{\kappa\lambda}(u) = \delta_{\kappa\lambda}$. Then, $\tilde{v}^*$ is first-order ancillary. It is always possible to choose such a coordinate system that $K_{\kappa\lambda\mu}(u) = 0$ further holds at v = 0 in every A(u). This coordinate system is indeed given by the ($\alpha = -1/3$)-normal coordinate system at v = 0. The $\tilde{v}^*$ is second-order ancillary in this coordinate system. By evaluating the term of order N^{-1} in (4.9), we can prove that there exists in general no third-order ancillary $\hat{v}$.

However, Skovgaard (1985), by using the method of Chernoff (1949), showed that one can always construct an ancillary $\hat{v}_q$ of order q for any q by modifying $\hat{v}$ successively. The q-th order ancillary $\hat{v}_q$ is a function of $\bar{x}$ depending on N. Hence, our previous result implies only that one cannot in general construct the third-order ancillary by using a function of $\bar{x}$ not depending on N, or by relying on an ancillary family A = {A(u)} not depending on N. There is no reason to stick to an ancillary family not depending on N, as Skovgaard argued.

4.3 Decomposition of information

Since $(\hat{u},\hat{v})$ together are sufficient, the information lost by summarizing $\bar{x}$ into $\hat{u}$ is recovered by knowing the ancillary $\hat{v}$. The amount of recovered information $g_{ab}(\hat{V}|\hat{U})$ is equal to $\Delta g_{ab}(\hat{U})$. Obviously, the amount of information of $\hat{v}$ relative to $\hat{u}$ does not depend on the coordinate system of A(u). In order to recover the information of order 1 in $\Delta g_{ab}(\hat{U})$, not all the components of $\hat{v}$ are necessary. Some functions of $\hat{v}$ can recover the full information

of order 1. Some other functions of $\hat{v}$ will recover the information of order N^{-1} and some others further will recover the information of order N^{-2}. We can decompose the whole ancillary $\hat{v}$ into parts according to the order of the magnitude of the amount of relative information.

The tangent space $T_u(A)$ of the ancillary subspace $A(u)$ associated with an efficient estimator $\hat{u}$ is spanned by $n - m$ vectors ∂_κ. The ancillary $\hat{v}$ can be regarded as a vector $\hat{v} = \hat{v}^\kappa \partial_\kappa$ belonging to $T_u(A)$. Now we decompose $T_u(A)$ as follows. Let us define

$$K_{a_1 \cdots a_p}{}^i = (\nabla^{(e)}_{a_1} \cdots \nabla^{(e)}_{a_{p-1}} B^i_{a_p}), \qquad p \geq 2 \tag{4.10}$$

which is a tensor representing the higher-order exponential curvature of the model. When $p = 2$, it is nothing but the exponential curvature $H^{(e)i}_{ab}$, and when $p = 3$, $K_{abc}{}^i$ represents the rate of change in the curvature $H^{(e)i}_{ab}$, and so on. For fixed indices $a_1, \ldots, a_p$, $K_{a_1 \cdots a_p}{}^i$ is a vector in $T_u(S)$, and its projection to $T_u(A)$ is given by

$$K_{a_1 \cdots a_p \kappa} = K_{a_1 \cdots a_p}{}^i B_{\kappa i} .$$

Let $T_u(A)_p$ $(p \geq 2)$ be the subspace of $T_u(A)$ spanned by vectors $K_{a_1 a_2 \kappa}$, $K_{a_1 a_2 a_3 \kappa}, \ldots, K_{a_1 \cdots a_p \kappa}$, and let $P^\lambda_{\kappa, p}$ be the orthogonal projection from $T_u(A)$ to $T_u(A)_p$. We call

$$H^{(e)}_{a_1 \cdots a_p \kappa} = (I^\lambda_\kappa - P^\lambda_{\kappa, p-1}) K_{a_1 \cdots a_p \kappa} \tag{4.11}$$

the p-th order exponential curvature tensor of the model M, where $I = (I^\lambda_\kappa)$ is the identity operator. The square of the p-th order curvature is defined by

$$(H^2_M)^p_{ab} = H^{(e)}_{a\, a_1 \cdots a_{p-1} \kappa} H^{(e)}_{b\, b_1 \cdots b_{p-1} \lambda} g^{\kappa\lambda} g^{a_1 b_1} \cdots g^{a_{p-1} b_{p-1}} . \tag{4.12}$$

There exists a finite p_0 such that $H^{(e)}_{a_1 \cdots a_p}$ vanishes for $p \geq p_0$.

Now let us consider the following sequence of statistics,

$$T_1 = \{\hat{u}\}, \qquad T_2 = H^{(e)}_{a_1 a_2 \kappa}(\hat{u}) \hat{v}^\kappa, \ldots \quad .$$

Moreover, let $t_a = \partial_a \ell(\bar{x}, \hat{u})$, which vanishes if $\hat{u}$ is the m.l.e. Obviously, the sequence $T_2, T_3, \ldots$ gives a decomposition of the ancillary statistic $\hat{v} = (\hat{v}^\kappa)$

into the higher-order curvature directions of M. Let

$$\tau_1 = T_1, \quad \tau_2 = \{t,T_1,T_2\},\ldots,\tau_p = \{\tau_{p-1},T_p\} .$$

Then, we have the following theorems (see Amari (1985)).

Theorem 4.1. The set of statistics τ_p is asymptotically sufficient of order p. The statistic T_p carries information of order p relative to τ_{p-1},

$$g_{ab}(T_p|\tau_{p-1}) = N^{-p+2}(H_M^2)^p_{ab} . \tag{4.13}$$

Theorem 4.2. The Fisher information $g_{ab}(\bar{X}) = Ng_{ab}(X)$ is decomposed into

$$g_{ab}(\bar{X}) = \sum_{p=1} g_{ab}(T_p|\tau_{p-1}) = g_{ab}(\hat{U}) + \sum_{p=2} N^{-p+2}(H_M^2)^p_{ab} . \tag{4.14}$$

The theorems imply the following. An efficient estimator $\hat{u}$ carries all the information of order N. The ancillary $\hat{v}$, which together with $\hat{u}$ carries the remaining smaller-order information, is decomposed into the sum of p-th order curvature-direction components $a_{a_1\ldots a_p} = H^{(e)}_{a_1\ldots a_p\kappa}\hat{v}^\kappa$, which carries all the missing information of order N^{-p+2} relative to τ_{p-1}. The proof is obtained by expanding $\partial_a \ell(\bar{x},u)$, where $\bar{u} = u - \hat{u}$, as

$$\partial_a \ell(\bar{x},u) = \partial_a \ell(\bar{x},\hat{u}) + \sum_{p=1} \frac{(-1)}{p!}\partial_a\partial_{a_1}\ldots\partial_{a_p}\ell(\bar{x},\hat{u})\bar{u}^{a_1}\ldots\bar{u}^{a_p}$$

and by calculating $g_{ab}(T_p|\tau_{p-1})$. The information carried by $\partial_a\partial_{a_1}\ldots\partial_{a_p}\ell(\bar{x},\hat{u})$ is equivalent to $(\partial_a B_{a_1\ldots a_p}{}^i)B_{\kappa i}\hat{v}^\kappa$ or $H^{(e)}_{a\,a_1\ldots a_p\kappa}\hat{v}^\kappa$ relative to τ_{p-1} up to the necessary order.

4.4. Conditional inference

When there exists an exact ancillary statistic a, the conditionality principle requires that statistical inference should be done by conditioning on a. However, there exist no non-trivial ancillary statistics in many problems. Instead, there exists an asymptotically ancillary statistic $\hat{v}$, which can be refined to be higher-order ancillary. The asymptotic ancillary statistic carries information of order 1, and is very useful in improving higher-order characteristics of statistical inference. For example, the conditional covariance of an efficient estimator is evaluated by

$$N\,\mathrm{Cov}[\hat{u}^a,\hat{u}^b|\hat{v}] = (g_{ab} + H^{(e)}_{ab\kappa}\bar{v}^\kappa)^{-1} + \text{higher order terms}\ ,$$

where $g_{ab} + H^{(e)}_{ab\kappa}\hat{v}^\kappa = -\partial_a\partial_b\ell(\bar{x},\hat{u})$ is the observed Fisher information. When two groups of independent observations are obtained, we cannot get a third-order efficient estimator for the entire set of observations by combining only the two third-order efficient estimators $\hat{u}_1$ and $\hat{u}_2$ for the respective samples. If we can use the asymptotic ancillaries $H^{(e)}_{ab\kappa}\hat{v}^\kappa_1$ and $H^{(e)}_{ab\kappa}\hat{v}^\kappa_2$, we can calculate the third-order efficient estimator (see Chap. 5). Moreover, the ancillary $H^{(e)}_{ab\kappa}\hat{v}^\kappa$ can be used to change the characteristics of an efficient test and of an efficient interval estimator. We can obtain the third-order t_0-efficient test or interval estimator by using the ancillary for any given t_0. It is interesting that the conditional test conditioned on the asymptotic ancillary $\hat{v}$ is third-order admissible and its characteristic (deficiency curve) is the same as that of the likelihood-ratio test (Kumon and Amari (1983)).

In the above discussions, it is not necessary to refine $\hat{v}$ to be a higher-order asymptotic ancillary. The curvature-direction components $H^{(e)}_{ab\kappa}\hat{v}^\kappa$ are important, and the other components play no role. Hence, we may say that $H^{(e)}_{ab\kappa}\hat{v}^\kappa$ is useful not because it is (higher-order) ancillary but because it recovers necessary information. It seems that we need a more fundamental study on the invariant structures of a model to elucidate the conditionality principle and ancillarity (see Kariya (1983), Barndorff-Nielsen, (1987).) There are many interesting discussions in Efron and Hinkely (1978), Hinkley (1980), Cox (1980), Barndorff-Nielsen (1980). See also Amari (1985).

5. FIBRE-BUNDLE THEORY OF STATISTICAL MODELS

Hilbert bundle of a statistical model

In order to treat general statistical models other then curved exponential families, we need the notion of fibre bundle of a statistical model. Let $M = \{q(x,u)\}$ be a general regular m-dimensional statistical model parametrized by $u = (u^a)$. To each point $u \in M$, we associate a linear space H_u consisting of functions $r(x)$ in x defined by

$$H_u = \{r(x) | E_u[r(x)] = 0,\ E_u[r^2(x)] < \infty\}, \tag{5.1}$$

where E_u denotes the expectation with respect to the distribution $q(x,u)$. Intuitively, each element $r(x) \in H_u$ denotes a direction of deviation of the distribution $q(x,u)$ as follows. Let $\varepsilon q(x)$ be a small disturbance of $q(x,u)$, where ε is a small constant, yielding another distribution $q(x,u) + \varepsilon q(x)$, which does not necessarily belong to M. Here, $\int q(x)dP = 0$ should be satisfied. The logarithm is written as

$$\log\{q(x,u) + \varepsilon q(x)\} \doteqdot \ell(x,u) + \varepsilon \frac{q(x)}{q(x,u)},$$

where $\ell(x,u) = \log q(x,u)$. If we put

$$r(x) = \frac{q(x)}{q(x,u)},$$

it satisfies $E_u[r(x)] = 0$. Hence, $r(x) \in H_u$ denotes the deviation of $q(x,u)$ in the direction $q(x) = r(x)q(x,u)$. The condition $E_u[r^2] < \infty$ implies that we consider only deviations having a second moment. (Note that given $r(x) \in H_u$, the function

$$q(x,u) + \varepsilon r(x) q(x,u)$$

does not necessarily represent a probability density function, because the positivity condition

$$q(x,u) + \varepsilon r(x)q(x,u) > 0$$

might be broken for $\pm\varepsilon$ even when ε is an infinitesimally small constant.)

We can introduce an inner product in the linear space H_u by

$$\langle r(x), s(x)\rangle = E_u[r(x)s(x)]$$

for $r(x)$, $s(x)\in H_u$. Thus, H_u is a Hilbert space. Since the tangent vectors $\partial_a \ell(x,u)$, which span $T_u(M)$, satisfy $E[\partial_a \ell] = 0$, $E[(\partial_a \ell)^2] = g_{aa}(u) < \infty$, they belong to H_u. Indeed, the tangent space $T_u(M)$ of M at u is a linear subspace of H_u, and the inner product defined in T_u is compatible with that in H_u. Let N_u be the orthogonal complement of T_u in H_u. Then, H_u is decomposed into the direct sum

$$H_u = T_u + N_u \ .$$

The aggregate of all H_u's attached to every $u\in M$ with a suitable topology,

$$\underline{H}(M) = \bigcup_{u\in M} H_u \ , \tag{5.2}$$

is called the fibre bundle with base space M and fibre space H. Since the fibre space is a Hilbert space, it is called a Hilbert bundle of M. It should be noted that H_u and $H_{u'}$ are different Hilbert spaces when $u \neq u'$. Hence, it is convenient to establish a one-to-one correspondence between H_u and $H_{u'}$, when u and u' are neighboring points in M. When the correspondence is affine, it is called an affine connection. Let us assume that a vector $r(x)\in H_u$ at u corresponds to $r(x) + dr(x)\in H_{u+du}$ at a neighboring point u + du, where d denotes infinitesimally small change. From

$$E_{u+du}[r(x) + dr(x)] = \int \{q(x,u) + dq(x,u)\}\{r(x) + dr(x)\}dP$$

$$= E_u[r] + E_u[dr(x) + \partial_a \ell(x,u)r(x)du^a] = 0$$

and $E_u[r] = 0$, we see that $dr(x)$ must satisfy

$$E_u[dr] = - E[\partial_a \ell r]\, du^a \ ,$$

where we neglected higher-order terms. This leads us to the following defini-

tion of the α-connection: When $dr(x)$ is given by

$$dr(x) = -\frac{1+\alpha}{2} E[\partial_a \ell r]\, du^a - \frac{1-\alpha}{2} \partial_a \ell r du^a , \tag{5.3}$$

the correspondence is called the α-connection. More formally, the α-connection is given by the following α-covariant derivative $\nabla^{(\alpha)}$. Let $r(x,u)$ be a vector field, which attaches a vector $r(x,u)$ to every point $u \in M$. Then, the rate of the intrinsic change of the vector $r(x,u)$ as u changes in the direction ∂_a is given by the α-covariant derivative,

$$\nabla^{(\alpha)}_{\partial_a} r = \partial_a r(x,u) - \frac{1+\alpha}{2} E_u[\partial_a r] + \frac{1-\alpha}{2} r \partial_a \ell, \tag{5.4}$$

where $E[\partial_a \ell r] = - E[\partial_a r]$ is used. The α-covariant derivative in the direction $A = A^a \partial_a \in T_u(M)$ is given by

$$\nabla^{(\alpha)}_A r = A^a \nabla^{(\alpha)}_{\partial_a} r .$$

The 1-connection is called the exponential connection, and the -1-connection is called the mixture connection.

When we attach the tangent space $T_u(M)$ to each point $u \in M$ instead of attaching the Hilbert space H_u, we have a smaller aggregate

$$\underline{T}(M) = \bigcup_{u \in M} T_u(M) ,$$

which is a subset of $\underline{H}$ called the tangent bundle of M. We can define an affine connection in $\underline{T}(M)$ by introducing an affine correspondence between neighboring T_u and $T_{u'}$. When an affine connection is given in $\underline{H}(M)$ such that $r \in H_u$ corresponds to $r + dr \in H_{u+du}$, it naturally induces an affine connection in $\underline{T}(M)$ such that $r \in T_u(M) \subset H_u$ corresponds to the orthogonal projection of $r + dr \in H_{u+du}$ to $T_{u+du}(M)$. It can easily be shown that the geometry of M is indeed that of $\underline{T}(M)$, so that the α-connection of $\underline{T}(M)$ or M, which we have defined in Chapter 2, is exactly the one which the present α-connection of $\underline{H}(M)$ naturally induces. Hence, the α-geometry of $\underline{H}(M)$ is a natural extension of that of M.

Let $u = u(t)$ be a curve in M. A vector field $r(x,t) \in H_{u(t)}$ defined along the curve is said to be α-parallel, when

$$\nabla^{(\alpha)}_{\dot{u}} r = \dot{r} - \frac{1+\alpha}{2} E_u[\dot{r}] + \frac{1-\alpha}{2} r \dot{\ell} = 0 \tag{5.5}$$

is satisfied, where $\dot{r}$ denotes $\partial r/\partial t$, etc. A vector $r_1(x) \in H_u$ is the α-parallel shift of $r_0(x) \in H_{u_0}$ along a curve $u(t)$ connecting $u_0 = u(t_0)$ and $u_1 = u(t_1)$, when $r_0(x) = r(x,t_0)$ and $r_1(x) = r(x,t_1)$ in the solution $r(x,t)$ of (5.5).

The parallel shift of a vector $r(x)$ from u to u' in general depends on the curve $u(t)$ along which the parallel shift takes place. When and only when the curvature of the connection vanishes, the shift is defined independently of the curve connecting u and u'. We can prove that the curvature of $\underline{H}(M)$ always vanishes for $\alpha = \pm 1$ connections, so that the e-parallel shift ($\alpha = 1$) and the m-parallel shift ($\alpha = -1$) can be performed from a point u to another point u' independently of the curve. Let ${}^{(e)}\pi_u^{u'}$ and ${}^{(m)}\pi_u^{u'}$ be the e- and m-parallel shift operators from u to u'. Then, we can prove the following important theorem.

Theorem 5.1. The exponential and mixture connections of $\underline{H}(M)$ are curvature-free. Their parallel shift operators are given, respectively, by

$$ {}^{(e)}\pi_u^{u'} r(x) = r(x) - E_{u'}[r(x)] , \tag{5.6} $$

$$ {}^{(m)}\pi_u^{u'} r(x) = \frac{q(x,u)}{q(x,u')} r(x) . \tag{5.7} $$

The e- and m-connections are dual in the sense of

$$ \langle r,s \rangle_u = \langle {}^{(e)}\pi_u^{u'} r , {}^{(m)}\pi_u^{u'} s \rangle_{u'} , $$

where $\langle .,. \rangle_u$ is the inner product at u.

Proof. Let c: $u(t)$ be a curve connecting two points $u = u(0)$ and $u' = u(1)$. Let $r^{(\alpha)}(x,t)$ be an α-parallel vector defined along the curve c. Then, it satisfies (5.5). When $\alpha = 1$, it reduces to

$$ \dot{r}^{(e)}(x,t) = E_{u(t)}[\dot{r}^{(e)}(x,t)]. $$

Since the right-hand side does not depend x, the solution of this equation with the initial condition $r(x) = r^{(e)}(x,0)$ is given by

$$ r^{(e)}(x,t) = r(x) + a(t) . $$

where $a(t)$ is determined from

$$ E_{u(t)}[r^{(e)}(x,t)] = 0 $$

as

$$a(t) = - E_{u(t)}[r(x)] .$$

This yields (5.6), where we put $u(t) = u'$. Since $E_{u'}[r(x)]$ does not depend on the path connecting u and u', the exponential connection is curvature free. Similarly, when $\alpha = -1$, (5.5) reduces to

$$\dot{r}^{(m)}(x,t) + r^{(m)}(x,t)\dot{\ell}(x,u(t)) = 0 .$$

The solution is

$$r^{(m)}(x,t)q(x,u(t)) = a(x) ,$$

which yields (5.7). This shows that the mixture connection is also curvature free. The duality relation is directly checked from (5.6) and (5.7).

We have defined the imbedding α-curvature $H^{(\alpha)}_{ab\kappa}$ of a curved exponential family. The concept of the imbedding curvature (which sometimes is called the relative or Euler-Schouten curvature) can be defined for a general M as follows. Let P^N_u be the projection operator of H_u to N_u which is the orthogonal subspace of $T_u(M)$ in H_u. Then, the imbedding α-curvature of M is a function in x defined by

$$H^{(\alpha)}_{ab}(x) = P^N_u \nabla^{(\alpha)}_{\partial_a} \partial_b \ell(x,u) ,$$

which is an element of $N_u \subset H_u$. The square of the α-curvature is given by

$$(H^{(\alpha)}_M)^2_{ab} = \langle H^{(\alpha)}_{ac}(x), H^{(\alpha)}_{bd}(x) \rangle g^{cd} . \tag{5.8}$$

The scalar $\gamma^2 = g^{ab}(H^{(e)}_M)^2_{ab}$ is the statistical curvature defined by Efron in the one-dimensional case.

5.2. Exponential bundle

Given a statistical model $M = \{q(x,u)\}$, we define the following elements in H_u,

$$X_{1a} = \partial_a \ell(x,u) ,$$

$$X_{2ab} = \nabla^{(\alpha)}_{\partial_a} X_{1b} ,$$

$$X_{ka_1 \ldots a_k} = \nabla^{(\alpha)}_{\partial a_i} X_{ka_2 \ldots a_k} ,$$

and attach to each point $u \in M$ the vector space $T^{(\alpha,k)}_u$ spanned by these vectors,

where we assume that they are linearly independent. The aggregate

$$\underline{T}^{(\alpha,k)}(M) = \underset{u\in M}{U} T_u^{(\alpha,k)} \tag{5.9}$$

with suitable topology is then called the α-tangent bundle of degree k of M. All the α-tangent bundles of degree 1 are the same, and are merely the tangent bundle $\underline{T}(M)$ of M. In the present paper, we treat only the exponential (i.e., $\alpha = 1$) tangent bundle of degree 2, which we call the local exponential bundle of degree 2, although it is immediate to generalize our results to the general α-bundle of degree k. Note that when we replace the covariant derivative $\nabla^{(\alpha)}$ by the partial derivative ∂, we have the so-called jet bundle. Its structures are the same as the exponential bundle, because $\nabla^{(e)}$ reduces to ∂ in the logarithm expression $\partial_a \ell(x,u)$ of tangent vectors.

The space $T_u^{(1,2)}$, which we will also more briefly denote by $T_u^{(2)}$, is spanned by vectors X_1 and X_2, where X_1 consists of m vectors

$$X_a(x,u) = \partial_a \ell(x,u),\ a = 1,\ldots,m$$

and X_2 consists of $m(m+1)/2$ vectors

$$X_{ab}(x,u) = \nabla^{(e)}_{\partial_a}\partial_b = \partial_a\partial_b\ell(x,u) + g_{ab}(u),\ a,\ b = 1,\ldots,m\ .$$

(See Fig. 8.) We often omit the indices a or a, b in the notation X_a or X_{ab}, briefly showing them as X_1 or X_2. Since the space $T_u^{(2)}$ consists of all the linear combinations of X_1 and X_2, it is written as

$$T_u^{(2)} = \{\theta^i X_i(x,u)\}$$

where the coefficients $\theta = (\theta^1,\theta^2)$ consist of $\theta^1 = (\theta^a)$, $\theta^2 = (\theta^{ab})$, and

$$\theta^i X_i = \theta^1 X_1 + \theta^2 X_2 = \theta^a X_a + \theta^{ab} X_{ab}\ .$$

The set X_i forms a basis of the linear space $T_u^{(2)}$. The metric tensor of $T_u^{(2)}$ is then given by

$$g_{ij} = \langle X_i, X_j\rangle = E_u[X_i(x,u)X_j(x,u)]\ .$$

Here, g_{11} denotes an $m \times m$ matrix

$$g_{11} = \langle X_a, X_b\rangle = E[\partial_a\ell\partial_b\ell] = g_{ab}$$

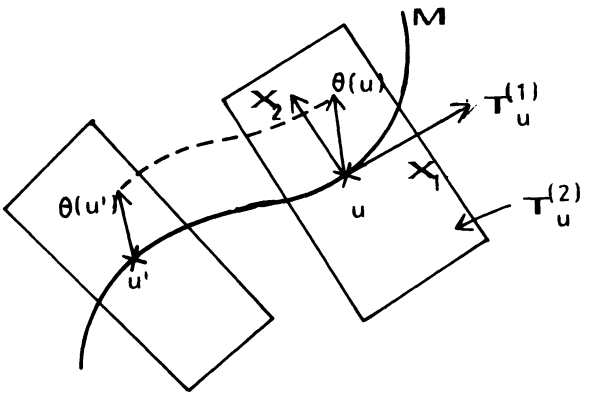

Figure 8

which is the metric tensor of the tangent space $T_u(M)$ of M. The component $g_{21} = g_{12}$ represents

$$g_{21} = g_{abc} = \langle X_{ab}, X_c \rangle = \Gamma^{(e)}_{abc} .$$

Similarly, g_{22} is a quantity having four indices

$$g_{22} = \langle X_{ab}, X_{cd} \rangle .$$

The exponential connection can be introduced naturally in the local exponential fibre bundle $\underline{T}^{(2)}(M)$ of degree 2 by the following principle:

1) The origin of $T^{(2)}_{u+du}$ corresponds to the point

$$X_1 du = X_a(x,u) du^a \varepsilon T^{(2)}_u$$

2) The basis vector $X_i(x,u + du) \varepsilon T^{(2)}_{u+du}$ is mapped to $T^{(2)}_u$ by 1-parallelly shifting it in the Hilbert bundle $\underline{H}$ and then projecting it to $T^{(2)}_u$. We thus have the affine correspondence of elements in $T^{(2)}_{u+du}$ and $T^{(2)}_u$,

$$X_i(u + du) \leftrightarrow X_i(u) + dX_i = X_i(u) + \Gamma^j_{ai} X_j(u) du^a ,$$

where Γ^j_{aj} are the coefficients of the exponential affine connection in $\underline{T}^{(2)}(M)$. The coefficients are given from the above principle (2) by

$$\Gamma^1_{a1} = 0, \quad \Gamma^2_{a1} = \delta^c_a \delta^d_d, \quad \Gamma^i_{a2} = g^{ij} E[X_j \partial_a \partial_b \partial_c \ell(x,u)] . \tag{5.10}$$

We remark again that the index i = 1 stands for a single index b, for example, and i = 2 stands for a pair of indices, for example b, c.

Let $\theta(u) = \theta^i(u)X_i(x,u) \in T_u^{(2)}$ be a point in $T_u^{(2)}$. We can shift the point $\theta(u) \in T_u^{(2)}$ to point $\theta(u') \in T_{u'}^{(2)}$ belonging to another point u' along a curve u = u(t). Since the point $\theta^i(u)X_i(u) \in T_u^{(2)}$ corresponds to the point $\theta^i(u + du)(X_i + dX_i) + X_1 du \in T_{u+du}^{(2)}$, where dX_i is determined from the affine connection and the last term $X_1 du$ corresponds to the change in the origin, we have the following equation

$$\dot{\theta}^i + \Gamma^i_{aj}\theta^j\dot{u}^a + \dot{u}^a\delta^i_a = 0 \ . \tag{5.11}$$

whose solution $\theta(t)$ represents the corresponding point in $T_{u(t)}^{(2)}$, where $\dot{\theta}^i = \dot{u}^a\partial_a\theta^i(u)$. Note that we are here talking about the parallel shift of a point in affine spaces, and not about the parallel shift of a vector in linear spaces where the origin is always fixed in the latter case.

Let u' be a point close to u. Let $\theta(u';u)$ be the point in $T_u^{(2)}$ corresponding to the origin $\theta(u') = 0$ of the affine space $T_{u'}^{(2)}$. The map depends in general on the curve connecting u and u'. However, when |u' - u| is small, the point $\theta(u';u)$ is given by

$$\theta^i(u';u) = \delta^i_1(u'-u) + \frac{1}{2}\delta^i_2(u'-u)^2 + O(|u'-u|^3) \ .$$

Hence, if we neglect the term of order $|u'-u|^3$, the map does not depend on the route. In the component form,

$$\theta^1(u';u) = \theta^a(u';u) = u'^a - u^a \ ,$$

$$\theta^2(u';u) = \theta^{bc}(u';u) = \frac{1}{2}(u'^b - u^b)(u'^c - u^c) \ , \tag{5.12}$$

where we neglected the term of order $|u'-u|^3$. Since the origin $\theta(u') = 0$ of $T_{u'}^{(2)}$ can be identified with the point u' (the distribution q(x,u')) in the model M, this shows that, in the neighborhood of u, the model M is approximately represented in $T_u^{(2)}$ as a paraboloid given by (5.12).

Let us consider the exponential family $E_u = \{p(x,\theta;u)\}$ depending on u, whose density function is given by

$$p(x,\theta;u) = q(x,u)\exp\{\theta^i X_i(x,u) - \psi_u(\theta)\} \ , \tag{5.13}$$

where θ is the natural parameter. We can identify the affine space $T_u^{(2)}$ with the exponential family E_u, by letting the point $\theta = \theta^i X_i \in T_u^{(2)}$ represent the

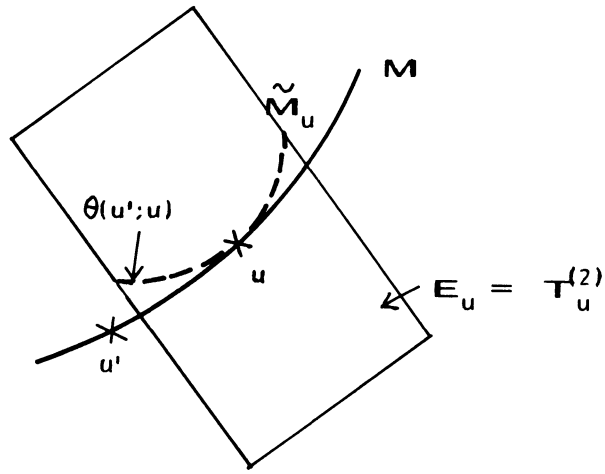

Figure 9

distribution $p(x,\theta;u)\in E_u$ specified by θ. We call E_u the local exponential family approximating M at u. The aggregate

$$\underline{E}(M) = \bigcup_{u\in M} E_u$$

with suitable topology is called the fibre bundle of local exponential family of degree 2 of M. The metric and connection may be defined from the resulting identification of $\underline{E}(M)$ with $\underline{T}^{(2)}(M)$. The distribution q(x,u) exactly corresponds to the distribution p(x,0;u) in E_u, i.e., the origin $\theta = 0$ of E_u or $T_u^{(2)}$. Hence, the point $\theta = \theta(u';u)$ which is the parallel shift of $\theta(u') = 0$ at $E_{u'}$, is the counterpart in E_u of the $q(x,u')\in M$, i.e., the distribution $p\{x,\theta(u',u);\ u\}\in E_u$ is an approximation in E_u of $q(x,u')\in M$. For a fixed u, the distributions

$$\tilde{M}_u = \{\tilde{q}(x,u';u)\}\ ,$$

$$\tilde{q}(x,u';u) = p\{x,\theta(u';u);\ u\}$$

form an m-dimensional curved exponential family imbedded in E_u (Fig. 9). The point of this construction is that M is approximated by a curved exponential family $\tilde{M}_u$ in the neighborhood of u. The tangent spaces $T_u(M)$ of M and $T_u(\tilde{M}_u)$ of $\tilde{M}_u$ exactly correspond at u, so that their metric structures are the same at u. Moreover, the squares of the imbedding curvatures are the same for both M and $\tilde{M}_u$ at u, because the curvature is obtained from the second covariant

derivative of $X_1 = \partial_a \ell$. This suggests that we can solve statistical inference problems in the curved exponential family $\tilde{M}_u$ instead of in M, provided u is sufficiently close to the true parameter u_0.

5.3. Statistical inference in a local exponential family

Given N independent observations $x_{(1)},\ldots,x_{(N)}$, we can define the observed point $\hat{\eta}(u)\in E_u$, for each u, by

$$\hat{\eta}_i(u) = \bar{X}_i(u) = \frac{1}{N} \sum_{j=1}^{N} X_i(x_{(j)},u) \ . \tag{5.14}$$

We consider estimators based on the statistics $\hat{\eta}(u)$. We temporarily fix a point u, and approximate model M by $\tilde{M}_u$, which is a curved exponential family imbedded in E_u. Let e be a mapping from E_u to $\tilde{M}_u$ that maps the observed $\bar{X}(u)\in E_u$ to the estimated value e(u) in $\tilde{M}_u$ when u is fixed, by denoting it as

$$e(u) = e\{\bar{X}(u);u\} \ .$$

The estimated value depends on the point u at which M is approximated by $\tilde{M}_u$. The estimator e defines the associated ancillary family $A_u = \{A_u(u'),\ u'\in\tilde{M}_u\}$ for every u, where

$$A_u(u') = e^{-1}(u';u) = \{\eta\in E_u | e(\eta;u) = u'\} \ .$$

When the fixed u is equal to the true parameter u_0, $\tilde{M}_{u_0}$ approximates M very well in the neighborhood of u_0. However, we do not know u_0. To get an estimator $\hat{u}$ from e, let us consider the equation

$$e\{\bar{X}(u);u\} = u \ .$$

The solution $\hat{u}$ of this equation is a statistic. It implies that, when M is approximated at $\hat{u}$, the value of the estimator e at $E_{\hat{u}}$ is exactly equal to $\hat{u}$. The characteristics of the estimator $\hat{u}$ associated with the estimator e in $\tilde{M}_u$ are given by the following geometrical theorems, which are direct extensions of the theorems in the curved exponential family.

Theorem 5.2. An estimator $\hat{u}$ derived from e is first-order efficient when the associated ancillary family A_u is orthogonal to $\tilde{M}_u$. A first-order efficient estimator is second-order efficient.

Theorem 5.3. The third-order term of the covariance of a bias corrected efficient estimator is given by

$$g_{3ab} = \frac{1}{2}(\Gamma^{(m)})^2_{ab} + (H^{(e)}_M)^2_{ab} + \frac{1}{2}(H^{(m)}_A)^2_{ab} .$$

The bias corrected maximum likelihood estimator is third-order efficient, because the associated ancillary family has vanishing mixture curvature.

The proof is obtained in the way sketched in the following. The true distribution $q(x,u_0)$ is identical with the distribution $q(x,\theta(u_0);u_0)$ at u_0 of the curved exponential family $\tilde{M}_{u_0}$. Moreover, when we expand $q(x,u)$ and $q(x,\theta(u);u_0)$ at u_0 in the Taylor series, they exactly coincide up to the terms of $u-u_0$ and $(u-u_0)^2$, because E_u is composed of X_1 and X_2. Hence, if the estimation is performed in E_{u_0}, we can easily prove that Theorems 5.2 and 5.3 hold, because the Edgeworth expansion of the distribution $\hat{u}$ is determined from the expansion of $\ell(\bar{x},u)$ up to the second order if the bias correction is used. However, we do not know the true u_0, so that the estimation is performed in $E_{\hat{u}}$. In order to evaluate the estimator $\hat{u}$, we can map $E_{\hat{u}}$ (and $\tilde{M}_{\hat{u}}$) to M_{u_0} by the exponential connection. In estimating the true parameter, we first summarize N observations into $\bar{X}(u)$ which is a vector function of u, and then decompose it into the statistics $\bar{X}(\hat{u}) = \{\bar{X}_1(\hat{u}),\bar{X}_2(\hat{u})\}$, where $e(\bar{X}(\hat{u});\hat{u}) = \hat{u}$. The $\bar{X}_2(\hat{u})$ becomes an asymptotic ancillary. When the estimator is the m.l.e., we have $\bar{X}_1(\hat{u}) = 0$ and $\bar{X}_2(\hat{u}) = H^{(e)}_{ab\kappa}\hat{v}^\kappa$ in $\tilde{M}_{\hat{u}}$. The theorems can be proved by calculating the Edgeworth expansion of the joint distribution of $\bar{X}(\hat{u})$ or $(\hat{u},\hat{v})$. The result is the same as before.

We have assumed that our estimator e is based on $\bar{X}(u)$. When a general estimator

$$\hat{u}' = f(x_{(1)},\ldots,x_{(N)})$$

is given, we can construct the related estimator given by the solution of $e_f(\bar{X}_{(u)};u) = u$, where

$$e_f(X;u) = E_u[f(x_{(1)},\ldots,x_{(N)})|\bar{X}(u) = X] .$$

Obviously, $e_f(X;u)$ is the conditional expectation of $\hat{u}'$ given $\bar{X}(u) = X$. By virtue of the asymptotic version of the Rao-Blackwell theorem, the behavior of e_f is equal to or better than $\hat{u}'$ up to the third-order. This guarantees the

validity of the present theory.

The problem of testing the null hypothesis $H_0: u = u_0$ against $H_1 : u \neq u_0$ can be solved immediately in the local exponential family E_u. When H_0 is not simple, we can also construct a similar theory by the use of the statistics $\hat{u}$ and $\bar{X}(\hat{u})$. It is possible to evaluate the behaviors of various third-order efficient tests. The result is again the same as before.

We finally treat the problem of getting a better estimator $\hat{u}$ by gathering asymptotically sufficient statistics $\bar{X}(\hat{u})$'s from a number of independent samples which are subject to the same distribution $q(x,u_0)$ in the same model. To be specific, let $x_{(1)1},\ldots,x_{(1)N}$ and $x_{(2)1},\ldots,x_{(2)N}$ be two independent samples each consisting of N independent observations. Let $\hat{u}_1$ and $\hat{u}_2$ be the m.l.e. based on the respective samples. Let $\bar{X}_{(i)}(\hat{u}_i)$ be the observed point in $E_{\hat{u}_i}$, i = 1, 2. The statistic $\bar{X}_{(i)}$ consists of two components $\bar{X}_{(i)1} = (\bar{X}_{(i)a})$ and $\bar{X}_{(i)2} = (\bar{X}_{(i)ab})$. Since $\hat{u}_i$ is the m.l.e.,

$$\bar{X}_{(i)1}(\hat{u}_i) = 0$$

is satisfied. The statistic $\hat{u}_i$ carries the whole information of order N included in the sample and the statistic $\bar{X}_2(\hat{u}_i)$, which is asymptotically ancillary, carries whole information of order 1 together with $\hat{u}_i$. Obviously $\bar{X}_{(i)2}$ is the curvature-direction component statistic, $\bar{X}_{(i)2} = H^{(e)}_{ab} v^{\kappa}_{(i)}$ in the curved exponential family $E_{\hat{u}_i}$.

Given two sets of statistics $(\hat{u}_i, \bar{X}_{(i)2}(\hat{u}_i))$, i = 1, 2, which summarize the original data, the problem is to obtain an estimator $\hat{u}$, which is third-order efficient for the 2N observations. Since the two statistics $\bar{X}(\hat{u}_i)$ give points $\hat{\eta}_{(i)} = \bar{X}(\hat{u}_i)$ in the different $E_{\hat{u}_i}$, in order to summarize them it is necessary to shift these points in parallel to a common $E_{u'}$. Then, we can average the two observed points in the common $E_{u'}$ and get an estimator $\hat{u}$ in this $E_{u'}$. The parallel affine shift of a point in E_u to a different $E_{u'}$ has already been given by (5.11) in the θ-coordinate system. This can be rewritten in the η-coordinate system. In particular, when $du = u - u'$ is of order $N^{-1/2}$ and $\hat{\eta}(u)$ is also of order $N^{-1/2}$, the parallel affine shift of $\eta(u) \in E_u$ to $E_{u'}$ is

given in the following expanded form for $\hat{\eta} = (\hat{\eta}_1, \hat{\eta}_2)$, $\hat{\eta}_1 = (\hat{\eta}_a)$ and $\hat{\eta}_2 = (\hat{\eta}_{ab})$,

$$\hat{\eta}_a(u') = \hat{\eta}_a(u) + g_{ab}du^b - \hat{\eta}_{ab}(u)du^b + \frac{1}{2}\Gamma^{(m)}_{bca}du^b du^c + O(N^{-3/2}) ,$$

$$\hat{\eta}_{ab}(u') = \hat{\eta}_{ab}(u) + O(N^{-1}) .$$

Now, we shift the two observed points $\bar{X}_{(i)}(\hat{u}_i)$ to a common $E_{u'}$, where u' may be any point between u_1 and u_2, because the same estimator $\hat{u}$ is obtained up to the necessary order by using any $E_{u'}$. Here, we simply put $u' = (u_1 + u_2)/2$, and let δ be

$$\delta = (u_1 - u_2)/2 .$$

Then, the point $\bar{X}_{(i)}(\hat{u}_i)$ is shifted to $\tilde{X}_{(i)}(u')$ of $E_{u'}$ as

$$\tilde{X}_{(1)a} = \bar{X}_{(1)a} + g_{ab}(u')\delta^b - \bar{X}_{(1)ab}\delta^b + \frac{1}{2}\Gamma^{(m)}_{bca}\delta^b\delta^c + O(N^{-3/2}) ,$$

$$\tilde{X}_{(1)ab} = \bar{X}_{(1)ab} + O(N^{-1}) ,$$

and we get similar expressions for $\tilde{X}_{(2)}$ by changing δ to $-\delta$. Since $\hat{u}_i$ is the m.l.e., $\bar{X}_{(i)a} = 0$. The average of $\tilde{X}_{(1)}$ and $\tilde{X}_{(2)}$ in the common $E_{u'}$ gives the estimated observed point $\tilde{X}(u') = (\tilde{X}_1, \tilde{X}_2)$ from the pooled statistics $(\hat{u}_i, \bar{X}_{(i)}(\hat{u}_i))$,

$$\tilde{X}_1 = \frac{1}{2}(\bar{X}_{2ab} - \bar{X}_{1ab})\delta^b + \frac{1}{2}\Gamma^{(m)}_{bca}\delta^b\delta^c ,$$

$$\tilde{X}_2 = \frac{1}{2}(\bar{X}_{2ab} + \bar{X}_{1ab}) .$$

By taking the m.l.e. in $E_{u'}$ based on $(\tilde{X}_1, \tilde{X}_2)$, we have the estimator

$$\hat{u}^a = u'^a - \frac{1}{2}g^{ab}(\bar{X}_{2bc} - \bar{X}_{1bc})\delta^c + \frac{1}{2}g^{ab}\Gamma^{(m)}_{cdb}\delta^c\delta^d ,$$

which indeed coincides with that obtained by the equation $e(\hat{u}) = \hat{u}$ up to the third order. Therefore, the estimator $\hat{u}$ is third-order efficient, so that it coincides with the m.l.e. based on all the 2N observations up to the necessary order.

The above result can be generalized in the situation where k asymptotically sufficient statistics $(\hat{u}_i, \bar{X}_{(i)2})$ are given in $E_{\hat{u}_i}$, $i = 1,\ldots,k$, $\hat{u}_i$ being the m.l.e. from N_i independent observations. Let

$$u' = \Sigma N_i\hat{u}_i/\Sigma N_i .$$

Moreover, we define the following matrices

$$G_{iab} = N_i[g_{ab}(u') + \frac{1}{2}\Gamma^{(m)}_{bca}(\hat{u}_i^c - u'^c) - \bar{X}_{2ab}] ,$$

$$G_{ab} = \sum_{i=1}^{k} G_{iab} , \qquad (G^{ab}) = (G_{ba})^{-1} .$$

Then, we have the following theorem.

Theorem 5.4. The bias corrected version of the estimator defined by

$$\hat{u}^a = G^{ab}[\sum_i G_{ibc}\hat{u}_i^c]$$

is third-order efficient.

This theorem shows that the best estimator is given by the weighted average of the estimators from the partial samples, where the weights are given by G_{iab}. It is interesting that G_{iab} is different from the observed Fisher information matrix

$$J_{iab} = -\Sigma\partial_a\partial_b\ell(x_{(i)},u') .$$

They are related by

$$G_{iab} = J_{iab} + \frac{1}{2}N_i\Gamma^{(m)}_{bca}(\hat{u}_i^c - u'^c) .$$

See Akahira and Takeuchi [1981] and Amari [1985].

6. ESTIMATION OF STRUCTURAL PARAMETER IN THE PRESENCE OF INFINITELY MANY NUISANCE PARAMETERS

Estimating function and asymptotic variance

Let $M = \{p(x;\theta,\xi)\}$ be a family of probability density functions of a (vector) random variable x specified by two scalar parameters θ and ξ. Let $x_1, x_2,\ldots,x_N$ be a sequence of independent observations such that the i-th observation x_i is a realization from the distribution $p(x;\theta,\xi_i)$, where both θ and ξ_i are unknown. In other words, the distributions of x_i are assumed to be specified by the common fixed but unknown parameter θ and also by the unknown parameter ξ_i whose value changes from observation to observation. We call θ the structural parameter and ξ the incidental or nuisance parameter. The problem is to find the asymptotic best estimator $\hat{\theta}_N = \hat{\theta}_N(x_1,x_2,\ldots,x_N)$ of the structural parameter θ, when the number N of observations is large. The asymptotic variance of a consistent estimator is defined by

$$AV(\hat{\theta},\Xi) = \lim_{N\to\infty} V[\sqrt{N}(\hat{\theta}_N - \theta)] \tag{6.1}$$

where V denotes the variance and Ξ denotes an infinite sequence $\Xi = (\xi_1,\xi_2,\ldots)$ of the nuisance parameter. An estimator $\hat{\theta}$ is said to be best in a class C of estimators, when its asymptotic variance satisfies, at any θ,

$$AV[\hat{\theta},\Xi] \leq AV[\hat{\theta}',\Xi]$$

for all allowable Ξ and for any estimator $\hat{\theta}' \in C$. Obviously, there does not necessarily exist a best estimator in a given class C.

Now we restrict our attention to some classes of estimators. An estimator $\hat{\theta}$ is said to belong to class C_0, when it is given by the solution of the equation

$$\sum_{i=1}^{N} y(x_i,\hat{\theta}) = 0 ,$$

where $y(x,\theta)$ is a function of x and θ only, i.e., it does not depend on ξ. The function y is called the estimating function. Let C_1 be a subclass of C_0, consisting of all the consistent estimators in C_0. The following theorem is well known (see, e.g., Kumon and Amari [1984]).

Theorem 6.1. An estimator $\hat{\theta}\in C_0$ is consistent if and only if its estimating function y satisfies

$$E_{\theta,\xi}[y(x,\theta)] = 0 , \qquad E_{\theta,\xi}[\partial_\theta y(x,\theta)] \neq 0 ,$$

where $E_{\theta,\xi}$ denotes the expectation with respect to $p(x;\theta,\xi)$ and $\partial_\theta = \partial/\partial\theta$. The asymptotic variance of an estimator $\hat{\theta}\in C_1$ is given by

$$AV(\hat{\theta},\Xi) = \lim N\, \Sigma V[y(x_i,\theta)] \,/\{(\Sigma\partial_\theta y)\}^2 ,$$

where $\Sigma\partial_\theta y(x_i,\theta)/N$ is assumed to converge to a constant depending on θ and Ξ.

Let $H_{\theta,\xi}(M)$ be the Hilbert space attached to a point $(\theta,\xi)\in M$,

$$H_{\theta,\xi}(M) = \{a(x) \mid E_{\theta,\xi}[a] = 0 , E_{\theta,\xi}[a^2] < \infty\}.$$

The tangent space $T_{\theta,\xi}(M) \subset H_{\theta,\xi}(M)$ is spanned by $u(x;\theta,\xi) = \partial_\theta \ell(x;\theta,\xi)$ and $v(x;\theta,\xi) = \partial_\xi \ell(x;\theta,\xi)$. Let w be

$$w(x;\theta,\xi) = u - \frac{\langle u,v\rangle}{\langle v^2\rangle} v ,$$

where $\langle v^2\rangle = \langle v,v\rangle$. Then, the partial information $\bar{g}_{\theta\theta}$ is given by

$$\bar{g}_{\theta\theta} = g_{\theta\theta} - g_{\theta\xi}^2/g_{\xi\xi} = \langle w^2\rangle ,$$

where $g_{\theta\theta} = \langle u^2\rangle$, $g_{\xi\xi} = \langle v^2\rangle$, $g_{\theta\xi} = \langle u,v\rangle$ are the components of Fisher information matrix. The theorem shows that the estimating function $y(x,\theta)$ of a consistent estimator belongs to $H_{\theta,\xi}$ for any ξ. Hence, it can be decomposed as

$$y(x,\theta) = a(\theta,\xi)u(x;\theta,\xi) + b(\theta,\xi)v(x;\theta,\xi) + n(x;\theta,\xi) ,$$

where n belongs to the orthogonal complement of $T_{\theta,\xi}$ in $H_{\theta,\xi}$, i.e.,

$$\langle u,n\rangle = \langle v,n\rangle = 0 .$$

The class C_1 is often too large to guarantee the existence of the best estimator. A consistent estimator is said to be uniformly informative

(Kumon and Amari, 1984) when its estimating function $y(x,\theta)$ can be decomposed as

$$y(x,\theta) = w(x;\theta,\xi) + n(x;\theta,\xi) .$$

The class of the uniformly informative estimators is denoted by C_{UI}. A uniformly informative estimator satisfies

$$\langle y,w\rangle_{\theta,\xi} = \langle w^2\rangle_{\theta,\xi} = \bar{g}_{\theta\theta}(\theta,\xi) .$$

Let C_{IU} be the class of the information unbiased estimators introduced by Lindsay [1982], which satisfy a similar relation,

$$\langle y,w\rangle_{\theta,\xi} = \langle y^2\rangle_{\theta,\xi} .$$

Note that $\langle y,w\rangle = \langle y,u\rangle$ holds.

Let us define the two quantities

$$g^0(\Xi) = \lim_{N\to\infty} \frac{1}{N} \langle \Sigma n(x;\theta,\xi_i)^2\rangle ,$$

which depends on the estimating function $y(x,\theta)$ and

$$\bar{g}(\Xi) = \lim \frac{1}{N} \Sigma \bar{g}_{\theta\theta}(\theta,\xi_i) ,$$

which latter is common to all the estimators. Then, the following theorem gives a new bound for the asymptotic variance in the class C_{IU} (see Kumon and Amari (1984)).

Theorem 6.2. For an information unbiased estimator $\hat{\theta}$

$$AV[\hat{\theta};\Xi] = \bar{g}^{-1} + \bar{g}^{-2} g^0 .$$

We go further beyond this theory by the use of the Hilbert bundle theory.

6.2. Information, nuisance and orthogonal subspaces

We have already defined the exponential and mixture covariant derivatives $\nabla^{(e)}$ and $\nabla^{(m)}$ in the Hilbert bundle $\underline{H} = \underline{U}_{(\theta,\xi)} H_{\theta,\xi}(M)$. A field $r(x;\theta,\xi) \in H_{\theta,\xi}(M)$ defined at all (θ,ξ) is said to be e-invariant, when $\nabla^{(e)}_{\partial_\xi} r = 0$ holds. A field $r(x;\theta,\xi)$ is said to be strongly e-invariant (se-invariant), when r does not depend on ξ. A se-invariant field is e-invariant. An estimating function $y(x,\theta)$ belonging to C_1 is an se-invariant field, and conversely, an se-invariant $y(x,\theta)$ gives a consistent estimator, provided $\langle u,y\rangle \neq 0$. Hence, the problem of the existence of a consistent estimator in C_0 reduces to

the problem of the existence of an se-invariant field in the Hilbert bundle $\underline{H}(M)$.

We next define the subspace $H^T_{\theta,\xi}$ of $H_{\theta,\xi}$ by

$$H^T_{\theta,\xi} = \underline{U}_{\xi'} \{ {}^{(m)}\pi^{\xi}_{\xi'} a(x) \mid a(x) \in T_{\theta,\xi'} \} ,$$

i.e., the subspace composed of all the m-parallel shifts to (θ,ξ) of the vectors belonging to the tangent space $T_{\theta,\xi'}$ at all (θ,ξ')'s with common θ. Then, $H_{\theta,\xi}$ is decomposed into the direct sum

$$H_{\theta,\xi} = H^T_{\theta,\xi} \oplus H^0_{\theta,\xi} ,$$

where $H^0_{\theta,\xi}$ is the orthogonal complement of $H^T_{\theta,\xi}$. We call $H^0_{\theta,\xi}$ the orthogonal subspace at (θ,ξ). We next define the nuisance subspace $H^N_{\theta,\xi}$ at (θ,ξ) spanned by the m-parallel shifts ${}^{(m)}\pi^{\xi}_{\xi'} v$ from (θ,ξ') to (θ,ξ) of the ξ-score vectors $v(x;\theta,\xi') = \partial_{\xi} \ell$ for all ξ'. It is a subspace of $H^T_{\theta,\xi}$, so that we have the decomposition

$$H^T_{\theta,\xi} = H^N_{\theta,\xi} \oplus H^I_{\theta,\xi} ,$$

where $H^I_{\theta,\xi}$ is the orthogonal complement of $H^N_{\theta,\xi}$ in $H^T_{\theta,\xi}$. It is called the information subspace at (θ,ξ). Hence,

$$H_{\theta,\xi} = H^I_{\theta,\xi} \oplus H^N_{\theta,\xi} \oplus H^0_{\theta,\xi} .$$

Any vector $r(x;\theta,\xi) \in H_{\theta,\xi}$ can uniquely be decomposed into the sum,

$$r(x;\theta,\xi) = r^I(x;\theta,\xi) + r^N(x;\theta,\xi) + r^0(x;\theta,\xi) , \tag{6.2}$$

where $r^I \in H^I_{\theta,\xi}$, $r^N \in H^N_{\theta,\xi}$ and $r^0 \in H^0_{\theta,\xi}$ are called respectively the I-, N- and 0-parts of r.

We now define some important vectors. Let us first decompose the θ-score vector $u = \partial_{\theta} \ell \in T_{\theta,\xi}$ into the three components. Let $u^I(x;\theta,\xi) \in H^I_{\theta,\xi}$ be the I-part of the θ-score $u \in T_{\theta,\xi}$. We next define the vector

$$\bar{u}(x;\theta,\xi;\xi') = {}^{(m)}\pi^{\xi}_{\xi'} u(x;\theta,\xi') \tag{6.3}$$

in $H_{\theta,\xi}$, which is the m-shift of the θ-score vector $u \in T_{\theta,\xi'}$ from (θ,ξ') to (θ,ξ). Let $\bar{u}^I$ be its I-part. The vectors $\bar{u}^I(x;\theta,\xi;\xi')$ in $H^I_{\theta,\xi}$ where (θ,ξ) is fixed, form a curve parametrized by ξ' in the information subspace $H^I_{\theta,\xi}$. When

all of $\bar{g}^{-1}_{\theta\theta}(\xi')\bar{u}^I(x;\theta,\xi;\xi') \in H^I_{\theta,\xi}$ lie in a hyperplane in $H^I_{\theta,\xi}$ for all ξ', we say that $\bar{u}^I$ are coplanar. In this case, there exists a vector $w^I \in H^I_{\theta,\xi}$ for which

$$<w^I,\bar{u}^I(x;\theta,\xi;\xi')> = \bar{g}_{\theta\theta}(\xi') \tag{6.4}$$

holds for any ξ'. The vector $w^I(w;\theta,\xi) \in H^I_{\theta,\xi}$ is called the information vector. When it exists, it is unique.

6.3. Existence theorems and optimality theorems

It is easy to show that a field $r(x;\theta,\xi)$ is se-invariant if its nuisance part r^N vanishes identically. Hence, any estimating function $y(x,\theta) \in C_1$ is decomposed into the sum

$$y(x,\theta) = y^I(x;\theta,\xi) + y^0(x;\theta,\xi) .$$

We can prove the following existence theorems.

Theorem 6.3. The class C_1 of the consistent estimators is nonempty if the information subspace $H^I_{\theta,\xi}$ includes a non-zero vector.

Theorem 6.4. The class C_{UI} of the uniformly informative estimators in C_1 is nonempty, if $\bar{u}^I(x;\theta,\xi;\xi')$ are coplanar. All the uniformly informative estimators have the identical I-part $y^I(x;\theta,\xi)$, which is equal to the information vector $w^I(x;\theta,\xi)$.

Outline of proof of Theorem 6.3. When the class C_1 is nonempty, there exist an estimating function $y(x,\theta)$ in C_1. It is decomposed as

$$y(x,\theta) = y^I(x;\theta,\xi) + y^0(x;\theta,\xi) .$$

Since y^0 is orthogonal to the tangent space $H^T_{\theta,\xi}$, we have

$$<y^0,u> = 0 .$$

By differentiating $<y(x,\theta)> = 0$ with respect to θ, we have

$$\begin{aligned} 0 &= <\partial_\theta y> + <y,u> \\ &= <\partial_\theta y> + <y^I,u> . \end{aligned}$$

Since $<\partial_\theta y> = 0$, we see that $y^I(x;\theta,\xi) \neq 0$, proving that $H^I_{\theta,\xi}$ includes a non-zero vector. Conversely, assume that there exists a non-zero vector $a(x,\theta)$ in $H^I_{\theta,\xi}$ for some ξ. Then, we define a vector

$$y(x;\theta,\xi') = {}^{(e)}\pi^{\xi'}_{\xi}\, a(x,\theta) = a(x,\theta) - E_{\theta,\xi'}[a]$$

in each $H_{\theta,\xi'}$ by shifting $a(x,\theta)$ in parallel in the sense of the exponential connection. By differentiating $\langle a\rangle_{\theta,\xi} = E_{\theta,\xi}[a]$ with respect to ξ, we have

$$\partial_\xi \langle a\rangle = \langle \partial_\xi a\rangle + \langle a, v\rangle = 0 ,$$

because a does not include ξ and a is orthogonal to $H^N_{\theta,\xi}$. This proves

$$E_{\theta,\xi'}[a] = 0 .$$

Hence, the above $y(x;\theta,\xi')$ does not depend on ξ' so that it is an estimating function belonging to C_1. Hence, C_1 is nonempty, proving theorem 6.3.

Outline of proof of Theorem 6.4. Assume that there exists an estimating function $y(x,\theta)$ belonging to C_{UI}. Then, we have

$$\langle y, u(x;\theta,\xi)\rangle_{\theta,\xi} = \bar{g}_{\theta\theta}(\xi) ,$$

because of $\langle y,v\rangle = 0$. Hence, when we shift y in exponential parallel and we shift u in mixture parallel along the ξ-axis, the duality yields

$$\langle {}^{(e)}_{\pi} y,\ {}^{(m)}_{\pi}(\bar{g}^{-1}u)\rangle = 1 ,$$

or

$$\langle y^I(x;\theta,\xi),\ \bar{u}^I(x;\theta,\xi;\xi')\rangle = \bar{g}_{\theta\theta}(\xi') .$$

This shows that $\bar{u}^I$ are coplanar, and the information vector w^I is given by projecting y to $H^I_{\theta,\xi'}$. Conversely, when $\bar{u}^I$ are coplanar, there exists the information vector $w^I \in H^I_{\theta,\xi}$. We can extend it to any ξ' by shifting it in exponential parallel,

$$y(x,\theta) = {}^{(e)}_{\pi_\xi^{\xi'}} w^I ,$$

which yields an estimating function belonging to C_{UI}.

The classes C_1 and C_{UI} are sometimes empty. We will give an example later. Even when they are nonempty, the best estimators do not necessarily exist in C_1 and in C_{IU}. The following are the main theorems concerning best estimators. (See Lindsay (1982) and Begun et al. (1983) for other approaches to this problem.)

Theorem 6.5. A best estimator exists in C_1, iff the vector field $u^I(x;\theta,\xi)$, which is the I-part of the θ-score u, is e-invariant. The best estimating function $y(x,\theta)$ is given by the e-invariant u^I, which in this case is se-invariant.

Theorem 6.6. A best estimator exists in C_{UI}, iff the information vector $w^I(x;\theta,\xi)$ is e-invariant. The best estimating function y is given by the e-invariant w^I, which in this case is se-invariant.

Outline of proofs. Let $\hat{\theta}$ be an estimator in C_1 whose estimating function is $y(x,\theta)$. It is decomposed into the following sum,

$$y(x,\theta) = c(\theta,\xi)\, u^I + a^I(x;\theta,\xi) + y^0(x;\theta,\xi) ,$$

where $u^I(x,\theta)$ is the projection of $u(x;\theta,\xi)$ to $H^I_{\theta,\xi}$, $c(\theta,\xi)$ is a scalar, and $a^I \varepsilon H^I_{\theta,\xi}$ is orthogonal to u^I in $H^I_{\theta,\xi}$. The asymptotic variance of $\hat{\theta}$ is calculated as

$$AV[\hat{\theta};\Xi] = \lim_{N\to\infty} N\{\Sigma(c_i^2 A_i + B_i)\}/\{(\Sigma c_i A_i)^2\} ,$$

where $\Xi = (\xi_1,\xi_2,\ldots)$, $c_i = c(\theta,\xi_i)$, and

$$A_i = \langle u^I, u^I \rangle_{\xi_i} ,$$

$$B_i = \langle (a^I(x))^2 \rangle_{\xi_i} + \langle (y^0)^2 \rangle_{\xi_i} .$$

From this, we can prove that, when and only when $B_i = 0$, the estimator is uniformly best for all sequences Ξ. The best estimating function is $u^I(x;\theta,\xi)$ for $\Xi = (\xi,\xi,\xi, \ldots)$. Hence it is required that u^I is se-invariant. This proves Theorem 6.5. The proof of Theorem 6.6 is obtained in a similar manner by using w^I instead of u^I.

6.4. Some typical examples: nuisance exponential family

The following family of distributions,

$$p(x;\theta,\xi) = \exp\{s(x,\theta)\xi + r(x,\theta) - \psi(\theta,\xi)\} \tag{6.5}$$

is used frequently in the literature treating the present problem. When θ is fixed, it is an exponential family with the natural parameter ξ, admitting a minimal sufficient statistic $s(x,\theta)$ for ξ. We call this an n-exponential family. We can elucidate the geometrical structures of the present theory by applying it to this family. The tangent vectors are given by

$$u = \xi\partial_\theta s + \partial_\theta r - \partial_\theta \psi , \qquad v = s - \partial_\xi \psi .$$

The m-parallel shift of $a(x)$ from (θ,ξ') to (θ,ξ) is

$$(m)_{\pi_\xi^\xi},\ a(x) = a(x)\exp\{(\xi - \xi')s - \psi(\xi) + \psi(\xi')\} .$$

From this follows a useful Lemma.

Lemma. The nuisance subspace $H^N_{\theta,\xi}$ is composed of random variables of the following form,

$$H^N_{\theta,\xi} = \{f[s(x,\theta) - c(\theta,\xi)]\} ,$$

where f is an arbitrary function and $c(\theta,\xi) = E_{\theta,\xi}[f(s)]$. The I-part a^I of a(x) is explicitly given as

$$a^I(x) = a(x) - E_{\theta,\xi}[a(x) \mid s(x,\theta)] , \tag{6.6}$$

by the use of the conditional expectation E[a|s]. The information subspace $H^I_{\theta,\xi}$ is given by

$$H^I_{\theta,\xi} = \{h(s;\theta,\xi)(\partial_\theta s)^I + f(s;\theta,\xi)(\partial_\theta r)^I\}$$

for any f, where $h = \partial_s f + \xi f$.

We first show the existence of consistent estimators in C_1 by applying Theorem 6.3.

Theorem 6.7. The class C_1 of consistent estimators is nonempty in an n-exponential family, unless both s and r are functionally dependent on s, i.e., unless

$$(\partial_\theta s)^I = (\partial_\theta r)^I = 0 .$$

On the other hand, a consistent estimator does not necessarily exist in general. We give a simple example: Let $x = (x_1,x_2)$ be a pair of random variables taking on two values 0 and 1 with probabilities

$$P(x_1 = 0) = 1/(1 + \exp\{\theta + \xi\}) ,$$

$$P(x_2 = 0) = 1/(1 + \exp\{k(\xi)\}) ,$$

where k is a known nonlinear function. The family M is of n-exponential type only when k is a linear function. We can prove that $H^I_{\theta,\xi} = \{0\}$, unless k is linear. This proves that there are no consistent estimators in this problem.

Now we can obtain the best estimator when it exists for n-exponential family. The I-part of the θ-score u is given by

$$u^I(x;\theta,\xi) = (\partial_\theta s)^I + (\partial_\theta r)^I .$$

It is e-invariant, when and only when $(\partial_\theta s)^I = 0$.

Theorem 6.8. The optimal estimator exists in C_1 when and only when $(\partial_\theta s)^I = 0$, i.e., $\partial_\theta s(x,\theta)$ is functionally dependent on s. The optimal estimating function is given in this case by the conditional score $u^I = (\partial_\theta r)^I = \partial_\theta r - E[\partial_\theta r \mid s]$, and moreover the optimal estimator is information unbiased in this case.

According to Theorem 6.4, in order to guarantee the existence of uniformly informative estimators, it is sufficient to show the coplanarity of $\bar{u}^I(x;\theta,\xi;\xi')$, which guarantees the existence of the information vector $w(x;\theta,\xi)\in H^I_{\theta,\xi}$. By putting $w = h(s)(\partial_\theta s)^I + f(s)(\partial_\theta r)^I$, this reduces to the integral-differential equation in f,

$$\langle w,\xi'(\partial_\theta s)^I + (\partial_\theta r)^I\rangle_{\xi'} = \bar{g}_{\theta\theta}(\xi') . \tag{6.7}$$

When the above equation has a solution $f(s;\theta,\xi)$, $\bar{u}^I$ are coplanar and the information vector w^I exists. Moreover, we can prove that when $(\partial_\theta r)^I = 0$, the information vector w^I is e-invariant.

Theorem 6.9. The best uniformly informative estimator exists when $(\partial_\theta r)^I = 0$. The best estimating function is given by solving

$$E_{\theta,\xi'}[h(s)V[\partial_\theta s \mid s]] = \bar{g}_{\theta\theta}(\xi')/\xi' , \tag{6.8}$$

where $h(s;\theta)$ does not depend on ξ' and $V[\partial_\theta s \mid s]$ is the conditional covariance.

We give another example to help understanding. Let $x = (x_1,x_2)$ be a pair of independent normal random variables, $x_1\sim N(\xi,1)$, $x_2\sim N(\theta\xi,1)$. Then, the logarithm of their joint density is

$$\ell(x;\theta,\xi) = -\frac{1}{2}[(x_1 - \xi)^2 + (x_2 - \theta\xi)^2 - \log(2\pi)]$$

$$= \xi s(x,\theta) + r(x,\theta) - \psi(\theta,\xi) ,$$

where $s(x,\theta) = x_1 + \theta x_2$, $r(x,\theta) = -(x_1^2 + x_2^2)/2$, $\psi(\theta,\xi) = \xi^2(1 + \theta^2)/2 + \log(2\pi)$. From $\partial_\theta s = x_2$, $\partial_\theta r = 0$, we have

$$(\partial_\theta s)^I = (x_2 - \theta x_1)/(1 + \theta^2) , \qquad (\partial_\theta r)^I = 0.$$

Hence, from Theorems 6.7 and 6.8, the class C_1 is nonempty, but the best estimator does not exist in C_1. Indeed, we have

$$u^I(x;\theta,\xi) = \xi(x_2 - \theta x_1)/(1 + \theta^2) ,$$

which depends on ξ so that it is not e-invariant. Since any vector w in $H^I_{\theta,\xi}$ can be written as

$$w = h(s)(\partial_\theta s)^I$$

for some $h(s;\theta,\xi)$, the information vector $w^I(x;\theta,\xi)\in H^I_{\theta,\xi}$ can be obtained by solving (6.4) or (6.7), which reduces in the present case to

$$E_{\theta,\xi}[h(s)(x_2 - \theta x_1)] = \xi(1 + \theta^2) .$$

Hence, we have

$$h(s) = s/(1 + \theta^2) ,$$

which does not depend on ξ. Therefore, there exists a best uniformly informative estimator whose estimating function is given by

$$y(x,\theta) = w^I(x,\theta) = h(s)(\partial_\theta s)^I = (x_2 - \theta x_1)(x_1 + \theta x_2)/(1 + \theta^2)^2$$

or equivalently by $(x_2 - \theta x_1)(x_1 + \theta x_2)$. This is the m.l.e. estimator. This is not information unbiased.

7. PARAMETRIC MODELS OF STATIONARY GAUSSIAN TIME SERIES

α-representation of spectrum

Let M be the set of all the power spectrum functions $S(\omega)$ of zero-mean discrete-time stationary regular Gaussian time series, $S(\omega)$ satisfying the Paley-Wiener condition,

$$\int \log S(\omega) d\omega > -\infty .$$

Stochastic properties of a stationary Gaussian time series $\{x_t\}$, $t = \ldots, -1, 0, 1, 2, \ldots$, are indeed specified by its power spectrum $S(\omega)$, which is connected with the autocovariance coefficients c_t by

$$c_t = \frac{1}{2\pi} \int_{-\pi}^{\pi} S(\omega) \cos\omega t d\omega , \tag{7.1}$$

$$S(\omega) = c_0 + 2 \sum_{t>0} c_t \cos\omega t , \tag{7.2}$$

where

$$c_t = E[x_r x_{r+t}]$$

for any r. A power spectrum $S(\omega)$ specifies a probability measure on the sample space $X = \{x_t\}$ of the stochastic processes. We study the geometrical structure of the manifold M of the probability measures given by $S(\omega)$. A specific parametric model, such as the AR model M_n^{AR} of order n, is treated as a submanifold imbedded in M.

Let us define the α-representation $\ell^{(\alpha)}(\omega)$ of the power spectrum $S(\omega)$ by

$$\ell^{(\alpha)}(\omega) = \begin{cases} -\frac{1}{\alpha} \{S(\omega)\}^{-\alpha}, & \alpha \neq 0 , \\ \log S(\omega) , & \alpha = 0 . \end{cases} \tag{7.3}$$

(Remark: It is better to define the α-representation by $-(1/\alpha)[S(\omega)^{-\alpha}-1]$. However, calculations are easier in the former definition, although the following discussions are the same for both representations.) We impose the regularity condition on the members of M that $\ell^{(\alpha)}$ can be expanded into the Fourier series for any α as

$$\ell^{(\alpha)}(\omega) = \xi_0^{(\alpha)} + 2\sum_{t>0}\xi_t^{(\alpha)}\cos\omega t \ , \tag{7.4}$$

where

$$\xi_t^{(\alpha)} = \frac{1}{2\pi}\int \ell^{(\alpha)}(\omega)\cos\omega t d\omega \ , \qquad t = 0\ 1,\ 2,\ \ldots .$$

We may denote the $\ell^{(\alpha)}(\omega)$ specified by $\xi^{\alpha} = \{\xi_t^{(\alpha)}\}$ by $\ell^{(\alpha)}(\omega;\xi^{(\alpha)})$. An infinite number of parameters $\{\xi_t^{(\alpha)}\}$ together specify a power function by

$$S(\omega;\xi^{(\alpha)}) = \begin{cases} [-\alpha\ell^{(\alpha)}(\omega;\xi^{(\alpha)})]^{-1/\alpha}, & \alpha \neq 0 \\ \exp\{\ell^{(0)}(\omega;\xi^{(0)})\}, & \alpha = 0 \ . \end{cases} \tag{7.5}$$

Therefore, they are regarded as defining an infinite-dimensional coordinate system in M. We call $\xi_t^{(\alpha)}$ the α-coordinate system of M. Obviously, the -1-coordinates are given by the autocovariances, $\xi_t^{(-1)} = c_t$. The negative of the 1-coordinates $\xi_t^{(1)}$, which are the Fourier coefficients of $S^{-1}(\omega)$, are denoted by $\tilde{c}_t$ and are called the inverse autocovariances, $\xi_t^{(1)} = -\tilde{c}_t$.

7.2. Geometry of parametric and non-parametric time-series models

Let M_n be a set of the power spectra $S(\omega;u)$ which are smoothly specified by an n-dimensional parameter $u = (u^a)$, $a = 1, 2, \ldots, n$, such that M_n becomes a submanifold of M., e.g., M_n could be an autoregressive process. This M_n is called a parametric time-series model. However, any member of M can be specified by an infinite-dimensional parameter u, e.g., by the α-coordinates $\xi^{(\alpha)} = \{\xi_t^{(\alpha)}\}$, $t = 0, 1, \ldots$ in the form $S(\omega,\xi^{(\alpha)})$. The following discussions are hence common to both the parametric and non-parametric models, irrespective of the dimension n of the parameter space.

We can introduce a geometrical structure in M or M_n in the same manner as we introduced before in a family of probability distributions on

sample space X, except that $X = \{x_t\}$ is infinite-dimensional in the present time-series case (see Amari, 1983 c). Let $p_T(x_1,\ldots,x_T; u)$ be the joint probability density of the T consecutive observations $x_1,\ldots,x_T$ of a time series specified by u. Let

$$\ell_T(x_1,\ldots,x_T;u) = \log p(x_1,\ldots,x_T;u) .$$

Then, we can introduce in M or M_n the following geometrical structures as before,

$$g_{ab}(u)=\lim_{T\to\infty} \frac{1}{T} E[\partial_a \ell_T \partial_b \ell_T] ,$$

$$\Gamma^{(\alpha)}_{abc} = \lim_{T\to\infty} \frac{1}{T} E[\{\partial_a\partial_b \ell_T - \frac{1-\alpha}{2} \partial_a \ell_T \partial_b \ell_T\}\partial_c \ell_T] .$$

However, the limiting process is tedious, and we define the geometrical structure in terms of the spectral density $S(\omega)$ in the following.

Let us consider the tangent space T_u at u of M or M_n, which is spanned by a finite or infinite number of basis vectors $\partial_a = \partial/\partial u^a$ associated with the coordinate system u. The α-representation of ∂_a is the following function in ω,

$$\partial_a = (\partial/\partial u^a)\ell^{(\alpha)}(\omega;u) .$$

Hence, in M, the basis $\partial^{(\alpha)}_t$ associated with the α-coordinates $\xi^{(\alpha)}_t$ is

$$\partial^{(\alpha)}_t = \begin{cases} 1 , & t = 0 \\ 2\cos\omega t , & t \neq 0 . \end{cases}$$

Let us introduce the inner product g_{ab} of ∂_a and ∂_b in T_u by

$$g_{ab}(u) = \langle \partial_a, \partial_b \rangle = E_\alpha[\partial_a \ell^{(\alpha)}(\omega;u)\partial_b \ell^{(\alpha)}(\omega;u)] ,$$

where E_α is the operator defined at u by

$$E_\alpha[a(\omega)] = \int \{S(w;u)\}^{2\alpha} a(\omega) d\omega .$$

The above inner product does not depend on α, and is written as

$$\langle \partial_a, \partial_b \rangle = \int \partial_a[\log S(\omega,u)]\partial_b[\log S(\omega,u)]d\omega . \qquad (7.6)$$

We next define the α-covariant derivative $\nabla^{(\alpha)}_{\partial_a}\partial_b$ of ∂_b in the

direction of ∂_a by the projection of $\partial_a\partial_b \ell^{(\alpha)}$ to T_u. Then, the components of the α-connection are given by

$$\Gamma^{(\alpha)}_{abc}(u) = \langle \nabla^{(\alpha)}_{\partial_a} \partial_b, \partial_c \rangle = \int S^{2\alpha} \partial_a\partial_b \ell^{(\alpha)} \partial_c \ell^{(\alpha)} d\omega \ . \tag{7.7}$$

If we use 0-representation, it is given by

$$\Gamma^{(\alpha)}_{abc}(u) = \int (\partial_a\partial_b \log S - \partial_a \log S \partial_b \log S)\partial_c \log S \, d\omega \ .$$

From (7.4) and (7.7), we easily see that the α-connection vanishes in M identically, if the α-coordinate system $\xi^{(\alpha)}$ is used. Hence, we have

Theorem 7.1. The non-parametric M is α-flat for any α. The α-affine coordinate system is given by $\xi^{(\alpha)}$. The two-coordinate systems $\xi^{(\alpha)}$ and $\xi^{(-\alpha)}$ are mutually dual.

Since M is α-flat, we can define the α-divergence from $S_1(\omega)$ to $S_2(\omega)$ in M. It is calculated as follows.

Theorem 7.2. The α-divergence from S_1 to S_2 is given by

$$D_\alpha(S_1,S_2) = \begin{cases} (1/\alpha^2) \int \{[S_2(\omega)/S_1(\omega)]^\alpha - 1 - \alpha \log[S_2/S_1]\} d\omega \ , & \alpha \neq 0 \\ (1/2) \int [\log S_1(\omega) - \log S_2(\omega)]^2 d\omega \ , & \alpha = 0 \ . \end{cases}$$

7.3. α-flat models

An α-model M^α_n of order n is a parametric model such that the α-representation of the power spectrum of a member in M^α_n is specified by n + 1 parameters $u = (u^k)$, $k = 0, 1, \ldots, n$, as

$$\ell^{(\alpha)}(\omega;u) = u_0 + 2 \sum_{k=1}^{n} u_k \cos k\omega \ .$$

Obviously, M^α_n is α-flat (and hence $-\alpha$-flat), and u is its α-affine coordinate system.

The AR-model M^{AR}_n of order n consists of the stochastic processes defined recursively by

$$\sum_{k=0}^{n} a_k x_{t-k} = \varepsilon_t$$

where $\{\varepsilon_t\}$ is a white noise Gaussian process with unit variance and $a = (a_0, a_1, \ldots, a_n)$ is the (n+1)-dimensional parameter specifying the members of M^{AR}_n.

Hence, it is an (n+1)-dimensional submanifold of M. The power spectrum $S(\omega;a)$ of the process specified by a is given by

$$S(\omega;a) = |\sum_{k=0}^{n} a_k e^{ik\omega}|^{-2} \quad .$$

We can calculate the geometric quantities of M_n^{AR} in terms of the AR-coordinate system a the above expression.

Similarly, the MA-model M_n^{MA} of order n is defined by the processes

$$x_t = \sum_{k=0}^{n} b_k \varepsilon_{t-k}$$

where $b = (b_0, b_1, \ldots, b_n)$ is the MA-parameter. The power spectrum $S(\omega;b)$ of the process specified by b is

$$S(\omega;b) = |\sum_{k} b_k e^{ik\omega}|^2 \quad .$$

The exponential model M_n^{EXP} of order n introduced by Bloomfield (1973) is composed of the following power spectra $S(\omega;e)$ parameterized by $e = (e_0, e_1, \ldots, e_n)$,

$$S(\omega;e) = \exp\{e_0 + 2\sum_{k=0}^{n} e_k \cos k\omega\} \quad .$$

It is easy to show that the 1-representation of $S(\omega;a)$ in M_n^{AR} is given by

$$\tilde{c}_k = \sum_{t=k}^{n} a_t a_{t-k} \quad , \qquad k = 0, 1, \ldots, n$$

$$\tilde{c}_k = 0 \quad , \qquad k > n$$

where

$$\ell^{(1)}(\omega;a) = -S^{-1}(\omega;a) = \sum_{k} \tilde{c}_k e^{i\omega k} \quad .$$

This shows that M_n^{AR} is a submanifold specified by $\tilde{c}_k = 0$, $(k > n)$ in M. Hence, it coincides exactly with a one-model $M_n^{(1)}$, although the coordinate system a is not 1-affine but curved. Similar discussions hold for M_n^{MA}.

Theorem 7.3. The AR-model M_n^{AR} coincides with $M_n^{(1)}$, and hence is ± 1-flat. The MA-model M_n^{MA} coincides with $M_n^{(-1)}$, and hence is also ± 1-flat. The exponential model M_n^{EXP} coincides with $M_n^{(0)}$, and is 0-flat. Since it is self-dual, it is an (n+1)-dimensional Euclidean space with an orthogonal Cartesian coordinate system e.

7.4. α-approximation and α-projection

Given a parametric model $M_n = \{S(\omega;u)\}$, it is sometimes necessary to approximate a spectrum $S(\omega)$ by one belonging to M_n. For example, given a finite observations $x_1, \ldots, x_T$ of $\{x_t\}$, one tries to estimate u in the parametric model M_n by obtaining first a non-parametric estimate $\hat{S}(\omega)$ based on $x_1, \ldots, x_T$ and then approximating it by $S(\omega;u)\in M_n$. The α-approximation of $\hat{S}$ is the one that minimizes the α-divergence $D_\alpha[\hat{S}(\omega), S(\omega,u)]$, $u\in M_n$. It is well known that the -1-approximation is related to the maximum likelihood principle. As we have shown in §2, the α-approximation is given by the α-projection of $\hat{S}(\omega)$ to M_n. We now discuss the accuracy of the α-approximation. To this end, we consider a family of nested models $\{M_n\}$ such that $M_0 \supset M_1 \supset M_2 \supset \ldots M_\infty = M$. The $\{M_n^{AR}\}, \{M_n^{MA}\}$ and $\{M_n^{EXP}\}$ are nested models, in which M_0 is composed of the white noises of various powers.

Let $\{M_n^\alpha\}$ be a family of the α-flat nested models, and let $\hat{S}_n(\omega;\hat{u}_n)\in M_n$ be the -α-approximation of $\hat{S}(\omega)$, where $\hat{u}_n$ is the (n+1)-dimensional parameter given by

$$\min_{S_n\in M_n^\alpha} D_{-\alpha}[\hat{S},S_n(\omega)] = D_{-\alpha}[\hat{S},\hat{S}_n(\omega;\hat{u}_n)] .$$

The error of the approximation by $\hat{S}_n\in M_n$ is measured by the -α-divergence $D_{-\alpha}(\hat{S},\hat{S}_n)$. We define

$$E_n(\hat{S}) = \min_{S_n\in M_n^\alpha} D_{-\alpha}(\hat{S},S_n) = D_{-\alpha}(\hat{S},\hat{S}_n) . \tag{7.8}$$

It is an interesting problem to find out how $E_n(\hat{S})$ decreases as n increases. We can prove the following Pythagorean relation (Fig. 10).

$$D_{-\alpha}(\hat{S},\hat{S}_n) = D_{-\alpha}(\hat{S},\hat{S}_{n+1}) + D_{-\alpha}(\hat{S}_{n+1},\hat{S}_n) .$$

The following theorem is a direct consequence of this relation.

Theorem 7.4. The approximation error $E_n(S)$ of S is decomposed as

$$E_n(S) = \sum_{k=n}^{\infty} D_{-\alpha}(\hat{S}_{k+1},\hat{S}_k) . \tag{7.9}$$

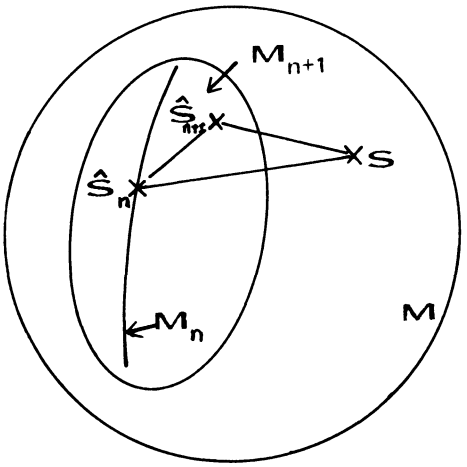

Figure 10

Hence,

$$D_{-\alpha}(S,\hat{S}_0) = \sum_{n=0}^{\infty} D_{-\alpha}(\hat{S}_{n+1},\hat{S}_n) .$$

The theorem is proved by the Pythagorean relation for the right triangle $\Delta S\hat{S}_n\hat{S}_0$ composed of the α-geodesic $\hat{S}_n\hat{S}_0$ included in M_n^{α} and $-\alpha$-geodesic $S\hat{S}_n$ intersecting at $\hat{S}_n$ perpendicularly. The theorem shows that the approximation error $E_n(S)$ is decomposed into the sum of the $-\alpha$-divergences of the successive approximations $\hat{S}_k$, $k = n+1, \ldots,\infty$, where $\hat{S}_{\infty} = S$ is assumed. Moreover, we can prove that the $-\alpha$-approximation of $\hat{S}_k$ in M_n^{α} $(n < k)$ is $\hat{S}_n$. In other words, the sequence $\{\hat{S}_n\}$ of the approximations of S has the following property that $\hat{S}_n$ is the best approximation of $\hat{S}_k$ $(k > n)$ and that the approximation error $E_n(S)$ is decomposed into the sum of the $-\alpha$-divergences between the further successive approximations. This is proved from the fact that the α-geodesic in M connecting two points S and S' belonging to M_n^{α} is completely included in M_n^{α} for an α-model M_n^{α}.

Let us consider the family $\{M_n^{AR}\}$ of the AR-models. It coincides with M_n^{1}. Let $\hat{S}_n$ be the -1-approximation of S. Let $c_t(S)$ and $\tilde{c}_t(S)$ be, respectively, the autocovariances and inverse autocovariances. Since c_t and $\tilde{c}_t$ are the mutually dual -1-affine and 1-affine coordinate systems, the -1-approx-

imation $\hat{S}_n$ of S is determined by the following relations

$$1)\quad c_t(\hat{S}_n) = c_t(S), \qquad t = 0, 1, \ldots, n$$

$$2)\quad \tilde{c}_t(\hat{S}_n) = 0, \qquad t = n+1, n+2, \ldots.$$

This implies that the autocovariances of $\hat{S}_n$ are the same as those of S up to $t = n$, and that the inverse autocovariances $\tilde{c}_t$ of $\hat{S}_n$ vanish for $t > n$. Similar relations hold for any other α-flat nested models, where c_t and $\tilde{c}_t$ are replaced by the dual pair of α- and $-\alpha$-affine coordinates. Especially, since $\{M_n^{EXP}\}$ are the nested Euclidean submanifolds with the self-dual coordinates $\xi^{(0)}$, their properties are extremely simple.

We have derived some fundamental properties of α-flat nested parametric models. These properties seem to be useful for constructing the theory of estimation and approximation of time series. Although we have not discussed about them here, the ARMA-modes, which are not α-flat for any α, also have interesting global and local geometrical properties.

Acknowledgements

The author would like to express his sincere gratitude to Dr. M. Kumon and Mr. H. Nagaoka for their collaboration in developing differential geometrical theory. Some results of the present paper are due to joint work with them. The author would like to thank Professor K. Takeuchi for his encouragement. He also appreciates valuable suggestions and comments from the referees of the paper.

REFERENCES

Akahira, M. and Takeuchi, K. (1981). On asymptotic deficiency of estimators in pooled samples. Tech. Rep. Limburgs Univ. Centr. Belgium.

Amari, S. (1968). Theory of information spaces --- a geometrical foundation of the analysis of communication systems. RAAG Memoirs 4, 373-418.

Amari, S. (1980). Theory of information spaces --- a differential geometrical foundation of statistics. POST RAAG Report, No. 106.

Amari, S. (1982a). Differential geometry of curved exponential families --- curvatures and information loss. Ann. Statist. 10, 357-387.

Amari, S. (1982b). Geometrical theory of asymptotic ancillarity and conditional inference. Biometrika 69, 1-17.

Amari, S. (1983a). Comparisons of asymptotically efficient tests in terms of geometry of statistical structures. Bull. Int. Statist. Inst., Proc. 44th Session, Book 2, 1190-1206.

Amari, S. (1983b). Differential geometry of statistical inference, Probability Theory and Mathematical Statistics (ed. Ito, K. and Prokhorov, J. V.), Springer Lecture Notes in Math 1021, 26-40.

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

Amari, S. (1985). Differential-Geometrical Methods in Statistics. Springer Lecture Notes in Statistics, 28, Springer.

Amari, S. and Kumon, M. (1983). Differential geometry of Edgeworth expansions in curved exponential family, Ann. Inst. Statist. Math. 35A, 1-24.

Atkinson, C. and Mitchell, A. F. (1981). Rao's distance measure, Sankya A43, 345-365.

Barndorff-Nielsen, O. E. (1980). Conditionality resolutions. Biometrika 67, 293-310.

Barndorff-Nielsen, O. E. (1987). Differential and integral geometry in statistical inference. IMS Monograph, this volume.

Bates, D. M. and Watts, D. G. (1980). Relative curvature measures of non-linearity, J. Roy. Statist. Soc. B40, 1-25.

Beale, E. M. L. (1960). Confidence regions in non-linear estimation. J. Roy. Statist. Soc. B22, 41-88.

Begun, J. M., Hall, W. J., Huang, W.-M. and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. Ann. Statist. 11, 432-452.

Bhattacharya, R. N. and Ghosh, J. K. (1978). On the validity of the formal Edgeworth expansion. Ann. Statist. 6, 434-451.

Bloomfield, P. (1973). An exponential model for the spectrum of a scalar time series. Biometrika 60, 217-226.

Burbea, J. and Rao. C. R. (1982). Entropy differential metric, distance and divergence measures in probability spaces: A unified approach. J. Multi. Var. Analys. 12, 575-596.

Chentsov, N. N. (1972). Statistical Decision Rules and Optimal Inference (in Russian). Nauka, Moscow, translated in English (1982), AMS, Rhode Island.

Chernoff, H. (1949). Asymptotic studentization in testing of hypotheses, Ann. Math. Stat. 20, 268-278.

Cox, D. R. (1980). Local ancillarity. Biometrika 67, 279-286.

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

Dawid, A. P. (1975). Discussions to Efron's paper. Ann. Statist. 3, 1231-1234.

Dawid, A. P. (1977). Further comments on a paper by Bradley Efron. Ann. Statist. 5, 1249.

Efron, B. (1975). Defining the curvature of a statistical problem (with application to second order efficiency) (with Discussion). Ann. Statist. 3, 1189-1242.

Efron, B. (1978). The geometry of exponential families. Ann. Statist. 6, 362-376.

Efron, B. and Hinkely, D. B. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information (with Discussion). Biometrika 65, 457-487.

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

Hinkely, D. V. (1980). Likelihood as approximate pivotal distribution. Biometrika 67, 287-292.

Hougaard, P. (1983). Parametrization of non-linear models. J. R. Statist. Soc. B44, 244-252.

James, A. T. (1973). The variance information manifold and the function on it. Multivariate Analysis (ed. Krishnaiah, P. K.), Academic Press, 157-169.

Kariya, T. (1983). An invariance approach in a curved model. Discussion paper Ser. 88, Hitotsubashi Univ.

Kass, R. E. (1980). The Riemannian structure of model spaces: A geometrical approach to inference. Ph.D. Thesis, Univ. of Chicago.

Kass, R. E. (1984). Canonical parametrization and zero parameter effects curvature. J. Roy. Statist. Soc. B46, 86-92.

Kumon, M. and Amari, S. (1983). Geometrical theory of higher-order asymptotics of test, interval estimator and conditional inference, Proc. Roy. Soc. London A387, 429-458.

Kumon, M. and Amari, S. (1984). Estimation of structural parameter in the presence of a large number of nuisance parameters. Biometrika 71, 445-459.

Kumon, M. and Amari, S. (1985). Differential geometry of testing hypothesis: a higher order asymptotic theory in multiparameter curved exponential family, METR 85-2, Univ. Tokyo.

Lauritzen, S. L. (1987). Some differential geometrical notions and their use in statistical theory. IMS Monograph, this volume.

Lindsay, B. G. (1982). Conditional score functions: Some optimality results. Biometrika 69, 503-512.

McCullagh, P. (1984). Tensor notation and cumulants of polynomials. Biometrika 71, 461-476.

Madsen, L. T. (1979). The geometry of statistical model --- a generalization of curvature. Research Report, 79-1, Statist. Res. Unit., Danish Medical Res. Council.

Nagaoka, H. and Amari, S. (1982). Differential geometry of smooth families of probability distributions, METR 82-7, Univ. Tokyo.

Pfanzagl, J. (1982). Contributions to General Asymptotic Statistical Theory. Lecture Notes in Statistics 13, Springer.

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

Reeds, J. (1975). Discussions to Efron's paper. Ann. Statist. 3, 1234-1238.

Skovgaard, Ib. (1985). A second-order investigation of asymptotic ancillarity, Ann. Statist. 13, 534-551.

Skovgaard, L. T. (1984). A Riemannian geometry of the multivariate normal model, Scand. J. Statist. 11, 211-223.

Yoshizawa, T. (1971). A geometrical interpretation of location and scale parameters. Memo TYH-2, Harvard Univ.