The Annals of Statistics
1988, Vol. 16, No. 3, 1044–1068

ESTIMATION IN THE PRESENCE OF INFINITELY MANY NUISANCE PARAMETERS—GEOMETRY OF ESTIMATING FUNCTIONS

BY SHUN-ICHI AMARI AND MASAYUKI KUMON

University of Tokyo

When there exist nuisance parameters whose number increases in proportion to that of independent observations, it is in general difficult to get a consistent or efficient estimator of a common structural parameter. The present paper proposes a new theory based on a vector bundle consisting of certain random variables over the statistical model. Structures and properties of estimating functions are elucidated in the class of consistent estimators. A necessary and sufficient condition is obtained for the existence of a consistent estimator given by an estimating function. A necessary and sufficient condition is then given for the existence of the optimal estimator in this class, which is further obtained when it exists. In their derivations, the concept of dual connections and parallel transports plays an essential role. The results are applied to a special type of exponential family, and the optimal estimators are explicitly obtained in some examples. This explains the reason why the conditional score plays an important role.

1. Introduction. The present paper treats the problem of estimating a structural parameter when there exist nuisance parameters whose number increases in proportion to the number of independent observations. Let $x_1, \dots, x_n$ be n independent observations, where each x_i is a vector random variable whose density function is given by $p(x_i; \theta, \xi_i)$. The scalar parameter θ is to be estimated and is called a structural parameter. The ξ_i, $i = 1, \dots, n$, are scalar nuisance parameters, which are unknown, taking on arbitrary values, and they are not necessarily an i.i.d. sample from an unknown distribution $p(\xi)$ of ξ. The problem is considered asymptotically in n, so that estimators of the structural parameter are evaluated by their asymptotic behaviors for large n.

This problem was treated by Neyman and Scott (1948), where it was remarked that the maximum likelihood estimator does not necessarily enjoy the consistency or the asymptotic efficiency in the sense of attaining the Cramér–Rao bound. Since then, many researchers have tackled this problem. Andersen (1970) showed that the conditional maximum likelihood estimator is consistent in a special type of model. See also Cox (1975). Concerning efficiency, Godambe (1976) introduced the conditional score function and showed its optimality in a finite but special case. Lindsay (1982) extended this idea in an asymptotic but more general situation.

Received December 1985; revised November 1987.

AMS 1980 *subject classifications*. Primary 62F10; secondary 62F12.

Key words and phrases. Asymptotic theory, bound for asymptotic variance, differential geometry, estimating function, exponential and mixture parallel transports, Hilbert bundle, nuisance parameter, structural parameter.

We treated this problem in Kumon and Amari (1984) from the geometrical point of view. Three classes of estimators $C_0 \supset C_1 \supset C_2$ were introduced: A C_0-estimator is given as the solution of the equation

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

where $y(x, \theta)$ is called an estimating function; see Godambe (1960, 1976) and Godambe and Thompson (1978). The class C_1 consists of all the consistent estimators in C_0, and the class C_2 consists of all the uniformly informative estimators in C_1. The main result in Kumon and Amari (1984) is that a new lower bound, other than the Cramér–Rao bound, is given for the asymptotic variance of estimators in C_2 and the optimal estimators in C_2 are explicitly given for some examples.

The aim of the present paper is to study the structure and the asymptotic behavior of all the C_1-estimators or estimating functions, providing a geometric theory of estimating functions; cf. Godambe (1960, 1976) and Godambe and Thompson (1978). We solve the following three problems for C_1-estimators by constructing a differential geometrical theory. They are (i) to give a necessary and sufficient condition for the existence of C_1-estimators, (ii) to elucidate structures of all the C_1-estimators and (iii) to give a necessary and sufficient condition for the existence of the optimal C_1-estimator and to obtain it when it exists. The theory is applied to statistical models in which there exists a sufficient statistic $s(x, \theta)$ for the nuisance parameter ξ when θ is fixed, where abstract results in this paper become very transparent. The present theory also explains why the conditional score plays an important role [Andersen (1970) and Godambe (1976)]. We can also treat the case when the number of dimensions of x_i is different for each i. This case occurs when a different number q_i of independent observations are made from a distribution $p(x; \theta, \xi_i)$ for each i. It is possible to construct a geometrical theory of C_2-estimators, although limitation of space does not permit us to describe it.

It should be remarked that there are many other approaches to this type of problem. Our approach is to obtain the estimator in C_1 which is universally optimal for any sequence of nuisance parameters $\xi_1, \xi_2, \dots$. This criterion of optimality is very strong so that the best estimator exists only in a very restricted class of distributions. Indeed, it is the aim of the present paper to obtain the class of such distributions and to obtain the optimal estimator in this class. Another approach is to obtain the optimal estimator depending on the realized sequence $\xi_1, \xi_2, \dots$ when it can be regarded as an i.i.d. sample from an unknown but fixed distribution $p(\xi)$. Such a problem is said to be semiparametric [Begun, Hall, Huang and Wellner (1983)], and the optimal estimator has already been obtained by using the adaptive estimator technique [Stein (1956), Bickel (1982) and Pfanzagl (1982); see also Bickel and Klaassen (1986)]. Lindsay (1983, 1985) also treated this problem. The third one is the minimax approach, aimed at obtaining the optimal estimator in the minimax sense, that is, to minimize the estimating loss provided the worst sequence $\xi_1, \xi_2, \dots$ is chosen from some prescribed class; see Hasminskii and Ibragimov (1983) and Nussbaum (1984). Obviously, the optimal solutions are different when the problems are set up in different ways. However, it is expected that the geometrical approach is useful and helpful for other approaches.

The present theory is constructed on the basis of differential geometry [see Amari (1982), Amari (1985), Amari, Barndorff-Nielsen, Kass, Lauritzen and Rao (1987) and Barndorff-Nielsen, Cox and Reid (1986)], and the concept of Hilbert fiber bundle (see Appendix D) and a pair of dual affine connections or parallel transports play a fundamental role. This suggests a new direction of development as well as its usefulness of the differential geometrical approach in statistics. However, we cannot be able to spare so much space for rigorous differential geometrical treatment. Refer to Kobayashi and Nomizu (1963) for detailed accounts on fiber bundles.

2. Explanation of the problem and classes of estimators. We begin with describing the problem, the classes of estimators treated here and the regularity conditions. Let $M = \{p(x; \theta, \xi)\}$ be a parametric statistical model, where $p(x; \theta, \xi)$ is the probability density function of a q-dimensional vector random variable x specified by two scalar parameters θ and ξ. Let $x_1, x_2, \dots, x_n, \dots$ be a sequence of independent observations, where the ith observation x_i is assumed to be subject to $p(x; \theta, \xi_i)$, $i = 1, 2, \dots$. In other words, the parameter θ, which is called the structural parameter, is common to all the observations but the value of the parameter ξ, which is called the nuisance or incidental parameter, changes observation by observation. Let $\bar{x} = (x_1, x_2, \dots)$ be an infinite sequence of observations and let $\bar{\xi} = (\xi_1, \xi_2, \dots)$ be a corresponding infinite sequence of the values of ξ. We put $\bar{x}_n = (x_1, \dots, x_n)$ and $\bar{\xi}_n = (\xi_1, \dots, \xi_n)$ for their n sections. The problem is to estimate the structural parameter θ based on the n observations $\bar{x}_n$ without any knowledge of the true $\bar{\xi}_n$. We evaluate the asymptotic behavior of an estimator $\hat{\theta}(\bar{x}_n)$ for large n and search for the best estimator in a class of estimators given in the following discussion. In Section 5 we treat the case where the dimension number q_i of the ith observation x_i is not necessarily the same for all i. This is the case when an unequal number q_i of independent observations is repeated for the distribution $p(x; \theta, \xi_i)$ in the ith trial.

We introduce the following classes of estimators [Kumon and Amari (1984)].

(i) Class C_0: An estimator $\hat{\theta}$ belongs to C_0, when it is given by the solution of the estimating equation for some $y(x, \theta)$,

$$(2.1) \qquad \sum_{i=1}^{n} y(x_i, \hat{\theta}) = 0.$$

The function $y(x, \theta)$, which depends neither on ξ nor on n, is called the estimating function of the estimator. Notice that two estimating functions $y(x, \theta)$ and $a(\theta)y(x, \theta)$ give the same estimator $\hat{\theta}$, provided $a(\theta) \neq 0$ for all θ.

(ii) Class C_1: An estimator $\hat{\theta}$ in class C_0 belongs to C_1, when it is consistent and is asymptotically normally distributed with a variance of order n^{-1}.

When there is no fear of confusion, we write as $y(x, \theta) \in C_1$ when the associated $\hat{\theta}$ belongs to C_1. The present paper studies characteristics of estimators in the class C_1.

We assume several regularity conditions in the following. Since we have interest in geometrical structures of the present problem rather than the rigorous regularity conditions, they are stronger than needed.

(1) The statistical model M forms a two-dimensional differentiable manifold in which the parameters $(\theta, \xi) \in \Theta \times \Xi$ give a coordinate system of M. Thus, the parameter space $\Theta \times \Xi$ is an open subset of $\mathbb{R}^2$.

(2) Let $u(x; \theta, \xi)$ and $v(x; \theta, \xi)$ be the θ- and ξ-score functions for the likelihood $p(x; \theta, \xi)$,

$$u(x; \theta, \xi) = \partial_\theta \log p(x; \theta, \xi), \tag{2.2}$$

$$v(x; \theta, \xi) = \partial_\xi \log p(x; \theta, \xi), \tag{2.3}$$

where $\partial_\theta = \partial/\partial\theta$, $\partial_\xi = \partial/\partial\xi$. It is assumed that

$$E_{\theta, \xi'}\left[u(x; \theta, \xi)^2\right] < \infty, \tag{2.4}$$

$$E_{\theta, \xi'}\left[v(x; \theta, \xi)^2\right] < \infty, \tag{2.5}$$

for any θ, ξ, ξ', where $E_{\theta, \xi'}$ denotes the expectation with respect to $p(x; \theta, \xi')$. Moreover, it is assumed that the Fisher information matrix is nondegenerate.

(3) Estimating functions are smooth in θ, and $y(x, \theta)$ and $\partial_\theta y(x, \theta)$, where $\partial_\theta = \partial/\partial\theta$, have finite third-order moments on the entire M.

The asymptotic variance of an estimator $\hat{\theta}$ is calculated as follows. By expanding the estimating equation (2.1) around the true θ, we have

$$\sum \left\{ y(x_i, \theta) + \partial_\theta y(x_i, \theta)(\hat{\theta} - \theta) \right\} = O_p(|\hat{\theta} - \theta|^2).$$

From the regularity condition (3), we see that, for any sequence $\bar{\xi}$ and any θ, $n^{-1}\Sigma y(x_i, \theta)$ and $n^{-1}\Sigma\, \partial_\theta y(x_i, \theta)$ converge to their expectations in probability. Hence, we have

$$n^{1/2}(\hat{\theta} - \theta) = -\left\{ n^{-1/2} \sum y(x_i, \theta) \right\} \Big/ \left\{ n^{-1} \sum \partial_\theta y(x_i, \theta) \right\} + O_p(n^{-1/2}).$$

Moreover, the central limit theorem holds for the random variable $n^{-1/2}\Sigma y(x_i, \theta)$. This shows that $\hat{\theta}$ is asymptotically normally distributed and is consistent for any $\bar{\xi}$ and any θ, when and only when its estimating function $y(x, \theta)$ satisfies

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

for all (θ, ξ), provided $E_{\theta, \xi}[\partial_\theta y(x, \theta)] \neq 0$. The asymptotic variance of $\hat{\theta}$ defined by

$$\text{a.v.}[\hat{\theta}; \theta, \bar{\xi}] = \lim_{n \to \infty} E_{\theta, \bar{\xi}_n}\left[n(\hat{\theta} - \theta)^2\right]$$

is given by

$$\text{a.v.}[\hat{\theta}; \theta, \bar{\xi}] = \langle\langle y(x, \theta)^2 \rangle\rangle_{\bar{\xi}} / \langle\langle \partial_\theta y(x, \theta) \rangle\rangle_{\bar{\xi}}^2, \tag{2.6}$$

where $\langle\langle\ \rangle\rangle_{\bar{\xi}}$ denotes

$$\langle\langle a(x)\rangle\rangle_{\bar{\xi}} = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} E_{\theta,\xi_i}[a(x_i)].$$

We thus have

THEOREM 1. *An estimator $\hat{\theta}$ belongs to C_1, when and only when its estimating function $y(x,\theta)$ satisfies*

$$E_{\theta,\xi}[y(x,\theta)] = 0, \qquad E_{\theta,\xi}[y^2] < \infty, \qquad E_{\theta,\xi}[\partial_\theta y]^2 > \varepsilon,$$

for some $\varepsilon > 0$.

An estimator $\hat{\theta} \in C_1$ is said to be asymptotically universally optimal or shortly optimal when for all $\bar{\xi}$ and θ its asymptotic variance is not larger than that of any other $\hat{\theta}' \in C_1$, i.e.,

$$\text{a.v.}[\hat{\theta};\theta,\bar{\xi}] \le \text{a.v.}[\hat{\theta}';\theta,\bar{\xi}]$$

holds for all $\bar{\xi}$ and θ.

One may think that the estimating equation (2.1) is very special, because it is a sum of independent random variables $y(x_1,\theta),\dots,y(x_n,\theta)$. A more general form of estimating equation is $Y(\bar{x}_n,\hat{\theta}) = 0$, which cannot in general be decomposed into the sum $\Sigma y(x_i,\theta)$. Therefore, we need to treat an extended class C_1' consisting of general estimating functions $Y(\bar{x}_n,\theta)$'s satisfying $E_{\theta,\bar{\xi}_n}[Y(\bar{x}_n,\theta)] = 0$, $E_{\theta,\bar{\xi}_n}[Y^2] < \infty$ and $E_{\theta,\bar{\xi}_n}[\partial_\theta Y(\bar{x}_n,\theta)]^2 > \varepsilon$ for any θ and $\bar{\xi}_n$, where $E_{\theta,\bar{\xi}_n}$ denotes the expectation with respect to $p(\bar{x}_n;\theta,\bar{\xi}_n) = \Pi_{i=1}^n p(x_i;\theta,\xi_i)$. However, it can be shown that if the optimal estimator exists in C_1' and is symmetric with respect to n variables $\bar{x}_n = (x_1,x_2,\dots,x_n)$, then the corresponding estimating function $Y^*(\bar{x}_n,\theta)$ can be expressed as the sum $Y^*(\bar{x}_n,\theta) = \Sigma y^*(x_i,\theta)$ (see Appendix A).

3. Hilbert bundle on statistical model.

3.1. *Hilbert bundle.* This subsection introduces the notion of a vector bundle on the statistical model M. This bundle will provide a framework for analyzing the class C_1 of estimators. With each point (θ,ξ) of the two-dimensional manifold $M = \{p(x;\theta,\xi)\}$, we associate a linear space $R_{\theta,\xi}$ consisting of all the random variables $r(x)$ which have zero expectations with respect to $p(x,\theta,\xi)$ and finite second-order moments,

$$R_{\theta,\xi} = \left\{r(x) | E_{\theta,\xi}[r(x)] = 0,\ E_{\theta,\xi'}\left[r(x)^2\right] < \infty\right\}, \tag{3.1}$$

for any ξ'. It includes the θ- and ξ-score functions, $u(x;\theta,\xi) \in R_{\theta,\xi}$, $v(x;\theta,\xi) \in R_{\theta,\xi}$. Intuitively, any $r(x) \in R_{\theta,\xi}$ represents a direction of small relative deviation of the probability distribution $p(x;\theta,\xi)$, because

$$p(x;\theta,\xi)\{1 + \varepsilon r(x)\}$$

is a probability distribution which does not necessarily belong to M, but is close to $p(x; \theta, \xi)$, when ε is infinitesimally small. The score functions u and v represent the deviations of $p(x; \theta, \xi)$ caused by a small change in θ and ξ, respectively. The linear space spanned by these scores

$$T_{\theta,\xi} = \{au(x; \theta, \xi) + bv(x; \theta, \xi)\}, \tag{3.2}$$

where a and b are scalar coefficients, is the tangent space at (θ, ξ) of M, where the natural basis ∂_θ and ∂_ξ are represented, respectively, by the random variables u and v [see Amari (1985)]. It is a linear subspace of $R_{\theta,\xi}$.

An aggregate of $R_{\theta,\xi}$'s at all $(\theta, \xi) \in M$ is denoted by

$$R = \bigcup_{\theta,\xi} R_{\theta,\xi}. \tag{3.3}$$

Such an aggregate is called a fiber bundle over M, when all the $R_{\theta,\xi}$'s are homeomorphic to a topological space F and, for any point denoted by (θ, ξ), there exists a neighborhood $N(\theta, \xi)$ of (θ, ξ) such that $\bigcup_{(\theta',\xi') \in N(\theta,\xi)} R_{\theta',\xi'}$ is homeomorphic to the direct product $N(\theta, \xi) \times F$ (see Appendix D). The space $R_{\theta,\xi}$ is called a fiber over a point (θ, ξ). The present R is a Hilbert bundle over M, because each fiber $R_{\theta,\xi}$ is a natural Hilbert space of random variables whose inner product is defined by

$$\langle r, s\rangle_{\theta,\xi} = E_{\theta,\xi}[rs], \quad \text{for } r, s \in R_{\theta,\xi}. \tag{3.4}$$

Similarly, we can define a vector bundle by the aggregate of $T_{\theta,\xi}$'s,

$$T = \cup_{\theta,\xi} T_{\theta,\xi}. \tag{3.5}$$

This is called the tangent bundle of M. It is a subbundle of R. Let us denote the inner products of the two score functions $u, v \in T_{\theta,\xi}$ by

$$g_{\theta\theta} = \langle u, u\rangle, \qquad g_{\theta\xi} = \langle u, v\rangle, \qquad g_{\xi\xi} = \langle v, v\rangle.$$

Then the resulting 2×2 matrix

$$g = \begin{bmatrix} g_{\theta\theta} & g_{\theta\xi} \\ g_{\theta\xi} & g_{\xi\xi} \end{bmatrix}$$

is the Fisher information matrix of M. This is a Riemannian metric of M [Rao (1945) and Amari (1985)].

We next define the concept of a section of the bundle R. A function $r(x; \theta, \xi)$ of x, θ and ξ is called a (smooth) section, when it defines an element $r(x; \theta, \xi) \in R_{\theta,\xi}$ for each point $(\theta, \xi) \in M$, i.e., it satisfies

$$E_{\theta,\xi}[r(x; \theta, \xi)] = 0, \qquad E_{\theta,\xi'}[r^2] < \infty,$$

for all (θ, ξ) and ξ'. The set of all the sections of R is denoted by $S(R)$,

$$S(R) = \{r(x; \theta, \xi) | r(x; \theta, \xi) \in R_{\theta,\xi}\}, \tag{3.6}$$

where r is smooth in θ and ξ. It is again a vector space. An estimating function $y(x, \theta) \in C_1$ is an example of a section, because it satisfies $E_{\theta,\xi}[y] = 0$, $E_{\theta,\xi'}[y^2] < \infty$ for all (θ, ξ) and ξ'. However, it is a special section in the sense that $y(x, \theta)$ is free of ξ.

A section t of the tangent bundle T is similarly defined. It is a random variable $t(x;\, \theta, \xi) \in T_{\theta, \xi}$, smoothly depending on θ and ξ. It is also called a vector field on M. The θ-score $u(x;\, \theta, \xi)$ and the ξ-score $v(x;\, \theta, \xi)$ are examples of the vector fields, and a section $t(x;\, \theta, \xi)$ is uniquely decomposed into the sum

$$t(x;\, \theta, \xi) = a(\theta, \xi)u(x;\, \theta, \xi) + b(\theta, \xi)v(x;\, \theta, \xi).$$

The vector space consisting of all the sections of T is denoted by $S(T)$. It is a subspace of $S(R)$.

In order to give an intuitive understanding of our general theory, a special type of statistical model, which we call the ξ-exponential type, will be used as a general example. When the density function is written as

$$p(x;\, \theta, \xi) = \exp\{\xi s(x, \theta) + r(x, \theta) - \psi(\theta, \xi)\}, \tag{3.7}$$

with respect to some dominating measure $P(x)$, a statistical model M is called a ξ-exponential family. For each fixed θ, it is an exponential family for the nuisance parameter ξ with sufficient statistic $s(x, \theta)$. This type of statistical model is widely used for examining the present problem, e.g., Andersen (1970), Godambe (1976), Lindsay (1982), etc.

Example 1. Let $x_1 \sim N(\xi, 1)$ and $x_2 \sim N(\theta\xi, 1)$ be two independent normal random variables with unit variance. Then the joint density function of $x = (x_1, x_2)$ is given by

$$\begin{aligned} p(x;\, \theta, \xi) &= \exp\left\{-\tfrac{1}{2}(x_1 - \xi)^2 - \tfrac{1}{2}(x_2 - \theta\xi)^2 - \log(2\pi)\right\} \\ &= \exp\left\{\xi(x_1 + \theta x_2) - \tfrac{1}{2}\left(x_1^2 + x_2^2\right) - \tfrac{1}{2}\xi^2(\theta^2 + 1) - \log(2\pi)\right\}. \end{aligned}$$

This is a ξ-exponential family with

$$s(x, \theta) = x_1 + \theta x_2, \qquad r(x, \theta) = -\tfrac{1}{2}\left(x_1^2 + x_2^2\right).$$

The scores are given by

$$\begin{aligned} u(x;\, \theta, \xi) &= \xi\left(x_2 - \theta\xi^2\right), \\ v(x;\, \theta, \xi) &= x_1 + \theta x_2 - \xi(1 + \theta^2), \end{aligned}$$

and are examples of the vector field. An example of a section is

$$r(x;\, \theta, \xi) = (x_1 - \xi)^3 + \theta\xi(x_2 - \theta\xi)^3.$$

The quantity

$$y(x;\, \theta, \xi) = (x_1 + \theta x_2)(x_2 - \theta x_1)$$

is another example of a section. Since it is free of ξ, it is an estimating function of a consistent estimator belonging to C_1.

3.2. *A dual pair of parallel transports in R.* A dual pair of covariant derivatives or parallel transports are introduced here. They play a central role in the present geometrical theory. Let us first define a family of differential operators $\nabla_\xi^{(\alpha)}$ depending on a real parameter α. It operates on a section

$r(x;\, \theta, \xi)$ as

$$\nabla_\xi^{(\alpha)} r = \partial_\xi r - \frac{1+\alpha}{2} E_{\theta,\xi}\left[\partial_\xi r\right] + \frac{1-\alpha}{2} r v, \tag{3.8}$$

where v is the ξ-score defined by (2.3).

We call $\nabla_\xi^{(\alpha)}$ the α-covariant derivative in the direction of the ξ-coordinate. This is a natural generalization of the α-covariant derivative or the α-connection [see Amari (1985)]. It is easy to prove

$$E_{\theta,\xi}\left[\nabla_\xi^{(\alpha)} r\right] = 0, \tag{3.9}$$

because of the identity

$$\partial_\xi E_{\theta,\xi}[r] = E_{\theta,\xi}\left[\partial_\xi r\right] + E_{\theta,\xi}[r v].$$

Hence, $\nabla_\xi^{(\alpha)}$ is a mapping from $S(R)$ to itself. The cases with $\alpha = \pm 1$ are especially important. We denote the $\alpha = 1$ and -1 covariant derivatives by $\nabla_\xi^{(e)}$ and $\nabla_\xi^{(m)}$, and call them the exponential (e-) and the mixture (m-) covariant derivatives, respectively. It is easy to define the α-covariant derivative $\nabla_\theta^{(\alpha)}$ in the θ-direction in a similar manner and then to extend it in arbitrary directions. It is also not difficult to prove that the properties of covariant differentiation are satisfied by the preceding definition, although we do not mention them.

A section $r(x;\, \theta, \xi) \in S(R)$ satisfying

$$\nabla_\xi^{(\alpha)} r = 0, \tag{3.10}$$

is called an α-parallel section along the ξ-coordinate. Since (3.10) is a first-order linear ordinary differential equation, we obtain an α-parallel section r by solving it. In particular, when $r_0(x, \theta) \in R_{\theta, \xi_0}$ is specified at one ξ_0, we can uniquely extend it to the α-parallel section $r(x;\, \theta, \xi)$ which satisfies

$$r(x;\, \theta, \xi_0) = r_0(x, \theta).$$

This extension defines the mapping

$${}^{(\alpha)}\pi_{\xi_0}^{\xi}\colon R_{\theta,\xi_0} \to R_{\theta,\xi},$$

which maps $r_0(x, \theta) \in R_{\theta,\xi_0}$ to ${}^{(\alpha)}\pi_{\xi_0}^{\xi} r_0 = r(x;\, \theta, \xi) \in R_{\theta,\xi}$. It is an isomorphism from R_{θ,ξ_0} to $R_{\theta,\xi}$, and we call it the α-parallel transport from R_{θ,ξ_0} to $R_{\theta,\xi}$ along the ξ-coordinate. Using the α-parallel transport ${}^{(\alpha)}\pi_{\xi'}^{\xi}$, the α-covariant derivative can be expressed as

$$\nabla_\xi^{(\alpha)} r = \lim_{\xi' \to \xi} \frac{1}{\xi' - \xi}\left\{{}^{(\alpha)}\pi_{\xi'}^{\xi} r(x;\, \theta, \xi') - r(x;\, \theta, \xi)\right\}.$$

In order to obtain the parallel transport of $a = a(x, \theta) \in R_{\theta,\xi_0}$, we need to solve the differential equation $\nabla_\xi^{(\alpha)} r = 0$ with the initial condition $r(x;\, \theta, \xi_0) = a$ at $\xi = \xi_0$. The e- and m-parallel transports corresponding to $\alpha = 1$ and $\alpha = -1$, which are of our concern, are explicitly given as

$${}^{(e)}\pi_\xi^{\xi'} a = a - E_{\theta,\xi'}[a], \tag{3.11}$$

$${}^{(m)}\pi_\xi^{\xi'} a = \{p(x;\, \theta, \xi)/p(x;\, \theta, \xi')\} a. \tag{3.12}$$

These are derived by solving (3.10) for $\alpha = 1$ and -1, respectively.

A remarkable feature of (3.11) and (3.12) is that they depend only on the end points ξ and ξ'. This is trivial in the present case where ξ is a scalar. However, in a general case when ξ is a vector parameter, there are many routes connecting two points ξ and ξ'. It is easy to define the α-covariant derivative in this general case. The problem is whether the α-parallel transport ${}^{(\alpha)}\pi_\xi^{\xi'}$ depends on the route connecting ξ and ξ' or not. This in turn depends on whether the α-curvature of R vanishes or not. The answer is that R is curvature-free for $\alpha = \pm 1$, and ${}^{(e)}\pi_\xi^{\xi'}$ and ${}^{(m)}\pi_\xi^{\xi'}$ do not depend on the route. The proof is given in Appendix B. Because of this property, the present theory can directly be extended to the case with vector θ and vector ξ without essential change.

As a final account, we note the relationship between the metric and the parallel transport. A parallel transport $\pi_\xi^{\xi'}$ is said to be metric, when it preserves the metric structure of R in the sense that, for $r, s \in R_{\theta, \xi}$,

$$\langle r, s\rangle_\xi = \langle \pi_\xi^{\xi'} r, \pi_\xi^{\xi'} s\rangle_{\xi'},$$

where $\langle r, s\rangle_\xi$ is an abbreviation of $\langle r, s\rangle_{\theta, \xi}$. Two parallel transports $\pi_\xi^{\xi'}$ and $\tilde{\pi}_\xi^{\xi'}$ are said to be mutually dual, when they together preserve the metric in the sense that

$$\langle r, s\rangle_\xi = \langle \pi_\xi^{\xi'} r, \tilde{\pi}_\xi^{\xi'} s\rangle_{\xi'}$$

[see Nagaoka and Amari (1982) and Amari (1985)]. It is easy to prove

THEOREM 2. *The α- and $-\alpha$-parallel transports are mutually dual. In particular, the* 0*-parallel transport is metric.*

For $\alpha = \pm 1$, the e- and m-parallel transports are mutually dual,

$$\langle r, s\rangle_\xi = \langle {}^{(e)}\pi_\xi^{\xi'} r, {}^{(m)}\pi_\xi^{\xi'} s\rangle_{\xi'}. \tag{3.13}$$

Its differential expression is

$$\partial_\xi \langle r, s\rangle_\xi = \langle {}^{(e)}\nabla_\xi r, s\rangle_\xi + \langle r, {}^{(m)}\nabla_\xi s\rangle_\xi,$$

for $r, s \in S(R)$. These formulas will be frequently used in due course.

EXAMPLE 1 (continued). The m- and e-parallel transports of $u(\xi) = u(x; \theta, \xi)$ and $v(\xi) = v(x; \theta, \xi)$ from ξ to ξ' are given as

$${}^{(m)}\pi_\xi^{\xi'} u(\xi) = Au(\xi), \quad {}^{(m)}\pi_\xi^{\xi'} v(\xi) = Av(\xi),$$

where

$$A = \exp\{(\xi - \xi')(x_1 + \theta x_2) - \tfrac{1}{2}(1 + \theta^2)(\xi^2 - \xi'^2)\},$$

$${}^{(e)}\pi_\xi^{\xi'} u(\xi) = \xi(x_2 - \theta\xi') = (\xi/\xi')u(\xi'),$$

$${}^{(e)}\pi_\xi^{\xi'} v(\xi) = x_1 + \theta x_2 - \xi'(1 + \theta^2) = v(\xi').$$

3.3. *Direct sum decomposition of R.* By using the dual e- and m-parallel transports, we study the direct sum decomposition of the Hilbert space $R_{\theta, \xi}$ and the Hilbert bundle R. Let

$$ {}^{(m)}\pi_{\xi'}^{\xi}T_{\theta,\xi'} = \left\{ {}^{(m)}\pi_{\xi'}^{\xi}t \,|\, t \in T_{\theta,\xi'} \right\} $$

be the m-parallel transport frcm ξ' to ξ of the tangent space $T_{\theta,\xi'}$ at (θ, ξ'). It is a linear subspace of $R_{\theta,\xi}$. Let us consider the union of such ${}^{(m)}\pi_{\xi'}^{\xi}T_{\theta,\xi'}$ for all ξ',

$$ \bigcup_{\xi'} {}^{(m)}\pi_{\xi'}^{\xi}T_{\theta,\xi'}, $$

and let $R_{\theta,\xi}^{T}$ be the smallest closed linear subspace of $R_{\theta,\xi}$ including the preceding union. In other words, $R_{\theta,\xi}^{T}$ is the closure of the subspace spanned by the vectors for all ξ' of the m-parallel transports from ξ' to ξ of the tangent vectors $t \in T_{\theta,\xi'}$. Then $R_{\theta,\xi}$ is decomposed into the following direct sum

$$ R_{\theta,\xi} = R_{\theta,\xi}^{T} \oplus R_{\theta,\xi}^{A}, \tag{3.14} $$

where $R_{\theta,\xi}^{A}$ is the orthogonal complement of $R_{\theta,\xi}^{T}$ in $R_{\theta,\xi}$ with respect to the inner product (3.4). We call $R_{\theta,\xi}^{T}$ the tangential subspace of $R_{\theta,\xi}$ and call $R_{\theta,\xi}^{A}$ the ancillary subspace of $R_{\theta,\xi}$.

We next decompose the $R_{\theta,\xi}^{T}$. Let $R_{\theta,\xi}^{N}$ be the closure of the linear subspace spanned by the m-parallel transports from ξ' to ξ of the ξ-scores $v \in T_{\theta,\xi'}$ for all ξ', i.e., the minimal closed subspace including ${}^{(m)}\pi_{\xi'}^{\xi}v(x;\, \theta, \xi')$ for any ξ'. Obviously, $R_{\theta,\xi}^{N}$ which is called the nuisance subspace, is a closed subspace of $R_{\theta,\xi}^{T}$. Hence, $R_{\theta,\xi}^{T}$ is decomposed into the following direct sum

$$ R_{\theta,\xi}^{T} = R_{\theta,\xi}^{N} \oplus R_{\theta,\xi}^{I}, \tag{3.15} $$

where the orthogonal complement $R_{\theta,\xi}^{I}$ of $R_{\theta,\xi}^{N}$ in $R_{\theta,\xi}^{T}$ is called the information subspace of $R_{\theta,\xi}$. Combining (3.14) and (3.15), we obtain the final orthogonal decomposition of $R_{\theta,\xi}$ into three subspaces,

$$ R_{\theta,\xi} = R_{\theta,\xi}^{I} \oplus R_{\theta,\xi}^{N} \oplus R_{\theta,\xi}^{A}. \tag{3.16} $$

Such a decomposition is obtained at every point (θ, ξ) of M. The aggregates

$$ R^{I} = \bigcup_{\theta,\xi} R_{\theta,\xi}^{I}, \tag{3.17} $$

$$ R^{N} = \bigcup_{\theta,\xi} R_{\theta,\xi}^{N}, \tag{3.18} $$

$$ R^{T} = \bigcup_{\theta,\xi} R_{\theta,\xi}^{T}, \qquad R^{A} = \bigcup_{\theta,\xi} R_{\theta,\xi}^{A} $$

define subbundles of R. In particular, R^{I} and R^{N} are called the information and nuisance subbundles, respectively. Corresponding to (3.16), R can also be expressed as

$$ R = R^{I} \oplus R^{A} \oplus R^{N}, \tag{3.19} $$

which is called the Whitney sum of the subbundles R^{I}, R^{A} and R^{N}.

We define the concept of e- and m-closedness of a subbundle, which characterizes the preceding subbundles. A subbundle R^{1} of R is said to be e- (or m-)

closed, when ${}^{(e)}\pi_{\xi'}^{\xi} r \in R^1_{\theta,\xi}$ (or ${}^{(m)}\pi_{\xi'}^{\xi} r \in R^1_{\theta,\xi}$) holds for any $r \in R^1_{\theta,\xi'}$ and for any ξ and ξ'. When R is decomposed into the orthogonal Whitney sum $R = R^1 \oplus R^2$, the following duality holds for the e- and m-closedness.

LEMMA 1. *A subbundle R^1 is e- (or m-) closed, if and only if R^2 is m- (or e-) closed.*

PROOF. For any $r \in R^1_{\theta,\xi}$ and $s \in R^2_{\theta,\xi}$, we have from (3.13)

$$0 = \langle r, s\rangle_\xi = \langle {}^{(m)}\pi_{\xi}^{\xi'} r, {}^{(e)}\pi_{\xi}^{\xi'} s\rangle_{\xi'},$$

at any ξ and ξ'. If R^1 is m-closed, then ${}^{(m)}\pi_{\xi}^{\xi'} r \in R^1_{\theta,\xi'}$. Since r is arbitrary, ${}^{(m)}\pi_{\xi}^{\xi'} r$ cover the entire $R^1_{\theta,\xi'}$. Hence, we have ${}^{(e)}\pi_{\xi}^{\xi'} s \in R^2_{\theta,\xi'}$, so that R^2 is e-closed. The converse is quite the same. □

Among our subbundles, R^N and R^T are m-closed by definition. Hence, it follows from the lemma that their complements $R^I \oplus R^A$ and R^A are e-closed. Note that the information subbundle R^I, although it is the complement of the m-closed R^N, is not in general e-closed, since the complement is taken not in R but in R^T. We summarize the results in the following theorem.

THEOREM 3. *The nuisance and tangential subbundles R^N and R^T are m-closed, and the ancillary subbundle R^A and its Whitney sum $R^A \oplus R^I$ with the information subbundle R^I are e-closed.*

In a ξ-exponential family, the m-parallel transport of $a(x) \in R_{\theta,\xi'}$ from (θ, ξ') to (θ, ξ) is given from (3.12) by

$$\begin{aligned}{}^{(m)}\pi_{\xi'}^{\xi} a &= \{p(x;\theta,\xi')/p(x;\theta,\xi)\}a \\ &= \exp\{(\xi' - \xi)s - (\psi' - \psi)\}a,\end{aligned}$$

where $\psi' = \psi(\theta, \xi')$. Since the ξ-score is $v = s - \partial_\xi \psi$, the nuisance subspace $R^N_{\theta,\xi}$ is spanned by

$$\begin{aligned}{}^{(m)}\pi_{\xi'}^{\xi} v(\xi') &= (s - \partial_\xi \psi')\exp\{(\xi' - \xi)s - (\psi' - \psi)\} \\ &= \exp[-\xi s + \psi]\, \partial_{\xi'}\exp[\xi' s - \psi'].\end{aligned}$$

The linear combination of the preceding functions with a weighting function $b(\xi')$ yields

$$\int b(\xi')^{(m)}\pi_{\xi'}^{\xi} v(\xi')\, d\xi' = \exp[-s\xi + \psi]\int a(\xi')\exp[s\xi' - \psi']\, d\xi',$$

where $\partial_\xi b(\xi) = -a(\xi)$, in which the second factor is regarded as the Laplace transform of an $\exp(-\psi)$. This is a function in s whose expectation vanishes at (θ, ξ). Hence, we have

$$R^N_{\theta,\xi} = \{f(s, \theta, \xi) | \langle f\rangle_{\theta,\xi} = 0\}, \tag{3.20}$$

where f is an arbitrary function expressed by the Laplace transform of an arbitrary function $a(\xi)$ in the previous form. This is the space of zero mean random variables generated by s.

We next study the tangential subspace $R^T_{\theta,\xi}$. Since the θ-score is $u = \xi\, \partial_\theta s + \partial_\theta r - \partial_\theta \psi$, by calculating the linear combination

$$\int a(\xi')^{(m)}\pi^{\xi}_{\xi'} u(\xi')\, d\xi',$$

we see that $R^T_{\theta,\xi}$ is of the form

$$R^T_{\theta,\xi} = \left\{ h(s;\theta,\xi) | \langle h \rangle_{\theta,\xi} = 0 \right\}, \tag{3.21}$$

where

$$h(s;\theta,\xi) = \partial_s k(s;\theta,\xi)\, \partial_\theta s + k(s;\theta,\xi)(\partial_\theta r + \xi\, \partial_\theta s) + f(s;\theta,\xi),$$

with arbitrary functions k and f.

Since the orthogonal projection of the random variable $z(x)$ to $R^N_{\theta,\xi}$ is given by the conditional expectation

$$P^N z = E[z|s],$$

the projection P^I of $z(x) \in R^T_{\theta,\xi}$ to the information subspace $R^I_{\theta,\xi}$ is given by

$$P^I z = z - E[z|s].$$

Hence, the information subspace $R^I_{\theta,\xi} = P^I R^T_{\theta,\xi}$ is written as

$$R^I_{\theta,\xi} = \left\{ (\partial_s k + \xi k(s)) P^I \partial_\theta s + k(s) P^I \partial_\theta r \right\}. \tag{3.22}$$

It is remarked that

$$P^I z = 0,$$

when and only when z is written as a function of s.

Example 1 (continued). The nuisance subspace $R^N_{\theta,\xi}$ of Example 1 consists of the random variables of the form $f(s) - c$, where $s = x_1 + \theta x_2$, $c = E_{\theta,\xi}[f(s)]$ and f is an arbitrary function. The tangential subspace $R^T_{\theta,\xi}$ consists of the random variables of the form

$$f(s) + x_2 k(s) - c(\theta,\xi),$$

where f and k are arbitrary. The ancillary subspace $R^A_{\theta,\xi}$ consists of random variables orthogonal to the preceding. For example, $k(s)\{(x_2 - \theta x_1)^2 - (1 + \theta^2)\}$ belongs to $R^A_{\theta,\xi}$. From

$$x_2 - E[x_2|s] = (x_2 - \theta x_1)/(1 + \theta^2),$$

we see that $R^I_{\theta,\xi}$ is composed of random variables of the form $k(s)(x_2 - \theta x_1)$. This shows that

$$y(x,\theta) = (x_2 - \theta x_1)(x_1 + \theta x_2)$$

belongs to $R^I_{\theta,\xi}$. Hence, it is a section belonging to R^I. An example of sections belonging to R^A is

$$k(s)\left\{(x_2 - \theta x_1)^2 - (1 + \theta^2)\right\}.$$

4. Geometrical structures of estimating functions.

4.1. *Condition for existence of C_1-estimators.* Having defined necessary geometrical concepts, we are now ready to investigate structures of estimating functions and to derive all the C_1-estimators. In some cases, there are no consistent estimators in C_0 and the class C_1 is void. Before obtaining the optimal estimator in C_1, we thus need a necessary and sufficient condition for the existence of C_1-estimators.

An estimating function $y(x, \theta)$ belonging to C_1 is a section because $E_{\theta, \xi}[y(x, \theta)] = 0$ holds at all $(\theta, \xi) \in M$. It moreover satisfies

$$\nabla_{\xi}^{(e)} y(x, \theta) = \partial_{\xi} y - E_{\theta, \xi}\left[\partial_{\xi} y\right] = 0,$$

so that it is an e-parallel section. As can be easily shown, an e-parallel section $r(x;\, \theta, \xi)$ is in general of the form

$$r(x;\, \theta, \xi) = r(x, \theta) - c(\theta, \xi),$$

where $c(\theta, \xi) = E_{\theta, \xi}[r(x, \theta)]$. An estimating function $y(x, \theta) \in C_1$ is special in the sense that it does not depend on ξ at all. Such an e-parallel section is called an e-invariant section or shortly an invariant section. Conversely, an invariant section $y(x, \theta)$ yields a C_1-estimator, provided it satisfies $E_{\theta, \xi}[\partial_{\theta} y]^2 > \varepsilon$. Let I be the set of all the invariant sections. It forms a vector space. Any element $y(x, \theta) \in I$ belongs to $R_{\theta, \xi}$ at any ξ. Hence, I is regarded as a subspace of $R_{\theta, \xi}$ at any (θ, ξ). Now we study the vector space I which yields all the C_1-estimators.

THEOREM 4. *The vector space of invariant sections is decomposed in $R_{\theta, \xi}$ at any ξ as*

$$I = R_{\theta, \xi}^{I} \oplus R_{\theta, \xi}^{A}.$$

PROOF. Given $y(x, \theta) \in I$, differentiating the identity $\langle y \rangle_{\xi'} = 0$ with respect to ξ', we have $\langle y, v(\xi') \rangle_{\xi'} = 0$, where $\langle\ \rangle_{\xi'}$ is the abbreviation of $E_{\theta, \xi'}[\ ]$. Then, using (3.13), we have at any ξ,

$$0 = \langle {}^{(e)}\pi_{\xi'}^{\xi} y, {}^{(m)}\pi_{\xi'}^{\xi} v(\xi') \rangle_{\xi} = \langle y, {}^{(m)}\pi_{\xi'}^{\xi} v(\xi') \rangle_{\xi}.$$

Since ξ' is arbitrary, this implies that y belongs to the complement of $R_{\theta, \xi}^{N}$ or to $R_{\theta, \xi}^{I} \oplus R_{\theta, \xi}^{A}$. Let us next take any $a(x, \theta) \in R_{\theta, \xi'}^{I}$ and $b(x, \theta) \in R_{\theta, \xi'}^{A}$ at one ξ' and put $y_0(x, \theta) = a(x, \theta) + b(x, \theta)$. This $y_0 \in R_{\theta, \xi'}$ can be extended to a section $y(x;\, \theta, \xi)$ by the e-parallel transport along the ξ-coordinate,

$$y(x;\, \theta, \xi) = {}^{(e)}\pi_{\xi'}^{\xi} y_0(x, \theta) = y_0(x, \theta) - \langle y_0 \rangle_{\xi},$$

which obviously is e-parallel. Since $R^I \oplus R^A$ is e-closed, it also belongs to $R^I \oplus R^A$. Hence, what is to be proved is that $y(x;\, \theta, \xi)$ does not depend on ξ or that

$$f(\xi) = \langle y_0 \rangle_{\xi}$$

vanishes at all ξ. By differentiating this, we have

$$\begin{aligned}\partial_\xi f(\xi) &= \langle y_0, v(\xi)\rangle_\xi = \langle y_0 - \langle y_0\rangle_\xi, v(\xi)\rangle_\xi \\ &= \langle {}^{(e)}\pi_\xi^{\xi'} y, {}^{(m)}\pi_\xi^{\xi'} v(\xi)\rangle_{\xi'} = \langle y_0, {}^{(m)}\pi_\xi^{\xi'} v\rangle_{\xi'} = 0,\end{aligned}$$

because of $y_0 \in R^I_{\theta,\xi'} \oplus R^A_{\theta,\xi'}$. Moreover, $f(\xi) = 0$ at $\xi = \xi'$. Hence, $f(\xi) = 0$ for all ξ, proving that

$$y(x;\theta,\xi) = y_0(x;\theta)$$

belongs to I. □

The theorem implies that any $y(x,\theta) \in I$ is uniquely decomposed into

$$y(x,\theta) = y^I(x;\theta,\xi) + y^A(x;\theta,\xi) \tag{4.1}$$

at any ξ, where $y^I \in R^I_{\theta,\xi}$ and $y^A \in R^A_{\theta,\xi}$. Conversely, for any $a(x,\theta) \in R^I_{\theta,\xi_0}$ and $b(x,\theta) \in R^A_{\theta,\xi_0}$ given at an arbitrary ξ_0, the sum $y(x,\theta) = a(x,\theta) + b(x,\theta)$ itself gives an invariant section. Its information part $y^I(x;\theta,\xi)$ is equal to $a(x,\theta)$ at $\xi = \xi_0$ but it in general depends on ξ. It is this information part that is important in estimation.

Lemma 2. *Any estimating function $y(x,\theta)$ in C_1 has nonvanishing information part $y^I(x;\theta,\xi)$ in the decomposition* (4.1) *at all* (θ,ξ).

Proof. Suppose contrary to the lemma that there exists a $y \in C_1$ such that $y^I(x;\theta,\xi_0) = 0$ at some ξ_0. Then $y(x,\theta) = y^A(x;\theta,\xi_0) \in R^A_{\theta,\xi_0}$ at this ξ_0 and $y(x,\theta)$ is the e-parallel extension of this y^A from ξ_0 to all ξ. Since R^A is e-closed, the e-parallel transport of $y(x,\theta) = y^A(x;\theta,\xi_0)$ from ξ_0 to any ξ belongs to $R^A_{\theta,\xi}$, so that $y^I(x;\theta,\xi) = 0$ for all ξ. Since $y \in R^A_{\theta,\xi}$ at any ξ, we have by differentiating $\langle y\rangle_\xi = 0$ with respect to θ,

$$\partial_\theta\langle y\rangle_\xi = -\langle y, u(\xi)\rangle_\xi = 0,$$

because $u(\xi) \in R^T_{\theta,\xi}$. But this contradicts the third condition given in Theorem 1 defining the class C_1. □

Combining the previous theorem and lemma, we have the theorem concerning the existence for consistent estimators in C_0.

Theorem 5. *A necessary and sufficient condition for the existence of an estimator belonging to C_1 is $R^I_{\theta,\xi} \neq \{0\}$ at some (θ,ξ).*

From (3.22), we immediately have

Corollary 1. *There always exist C_1-estimators in a ξ-exponential family, except for the case when both $\partial_\theta s$ and $\partial_\theta r$ are functions of s.*

As was noted in the last section, C_1-estimators exist in Example 1. We next give an example in which C_1-estimators do not exist.

EXAMPLE 2. Let $x = (x_1, x_2)$ be a pair of mutually independent random variables which assume two values 0 and 1 and let their probabilities be given by

$$\Pr(x_1 = 0) = 1/\{1 + \exp(\theta + \xi)\},$$
$$\Pr(x_2 = 0) = 1/\{1 + \exp f(\xi)\},$$

where $f(\xi)$ is a known function. By using the function $\delta_1(z)$ which is equal to 1 when $z = 1$ and otherwise equal to 0, the log likelihood of x can be written as

$$\begin{aligned} l(x; \theta, \xi) = {} & (\theta + \xi)\,\delta_1(x_1) + f(\xi)\,\delta_1(x_2) \\ & -\log\{1 + \exp(\theta + \xi)\}\{1 + \exp f(\xi)\}. \end{aligned}$$

Thus, this is of ξ-exponential type if and only if $f(\xi)$ is a linear function. It is easy to show that $R_{\theta,\xi}$ is spanned by three random variables

$$\delta_1(x_1)\,\delta_1(x_2) - c_{11}, \quad \delta_1(x_1)\{1 - \delta_1(x_2)\} - c_{10}, \quad \{1 - \delta_1(x_1)\}\,\delta_1(x_2) - c_{01},$$

where the constants c_{ij} are added such that the expectations vanish. When $f(\xi)$ is nonlinear, the nuisance subspace $R^N_{\theta,\xi}$ is proved to be three dimensional, too. Hence, $R^I_{\theta,\xi} = \{0\}$, so that from Theorem 5 there exists no C_1-estimator.

4.2. *Optimal estimating function in C_1.* We have shown that every estimating function in C_1 has a nonvanishing information component. We now define a special section u^I belonging to the information subbundle by the projection of the θ-score $u(x; \theta, \xi)$ to the information subspace $R^I_{\theta,\xi}$,

$$u^I(x; \theta, \xi) = P^I u(x; \theta, \xi). \tag{4.2}$$

We call u^I the projected score. The following theorem shows the important role played by the projected score u^I.

THEOREM 6. *A necessary and sufficient condition for the existence of the optimal estimator in C_1 is that one of the following cases holds.*

(i) *The projected score $u^I(x; \theta, \xi)$ is invariant, i.e., it does not depend on ξ,*

$$u^I(x; \theta, \xi) = u^I(x, \theta).$$

(ii) *The information subspace $R^I_{\theta,\xi}$ is one dimensional and the projected score u^I is written as*

$$u^I(x; \theta, \xi) = c(\theta, \xi) u^I_0(x, \theta),$$

i.e., R^I is e-closed. In the previous two cases, the optimal estimating function is given by the ξ-free $u^I(x, \theta)$ or $u^I_0(x, \theta)$.

PROOF. Suppose that there exists the optimal estimating function $y^*(x, \theta) \in C_1$. Since it includes no nuisance component, we decompose it into the sum

$$y^* = c^*(\theta, \xi) u^I + y_2^{*I}(x; \theta, \xi) + y^{*A}(x; \theta, \xi),$$

where $y_2^{*I} \in R^I_{\theta,\xi}$ is orthogonal to u^I and $y^{*A} \in R^A_{\theta,\xi}$. Putting $A(\xi) = \langle (u^I)^2 \rangle_\xi$

and $B(\xi) = \langle (y_2^{*I})^2 \rangle_\xi + \langle (y^{*A})^2 \rangle_\xi$, we have

$$\langle y^{*2} \rangle_\xi = c^{*2}A + B, \qquad -\langle \partial_\theta y^* \rangle_\xi = \langle u, y^* \rangle_\xi = c^* A.$$

Hence, the asymptotic variance of y^* is given from (2.6) by the limit as $n \to \infty$ of $nv(y^*, \bar{\xi}_n)$,

$$v(y^*, \bar{\xi}_n) = \left\{ \sum (c_i^{*2} A_i + B_i) \right\} \Big/ \left(\sum c_i^* A_i \right)^2,$$

where

$$c_i^* = c^*(\theta, \xi_i), \qquad A_i = A(\xi_i), \qquad B_i = B(\xi_i).$$

We now show that $y_2^{*I}(\xi) = y^{*A}(\xi) = 0$ must hold identically in ξ. Suppose on the contrary that there exists a ξ_0 such that $y_2^{*I}(\xi_0)$ or $y^{*A}(\xi_0)$ is nonzero. Then $B(\xi_0) > 0$, so that, for the special sequence $\bar{\xi}_0 = (\xi_0, \xi_0, \xi_0, \dots)$,

$$v(y^*, \bar{\xi}_0) > 1/\{nA(\xi_0)\} = v(u^I(\xi_0), \bar{\xi}_0).$$

This shows that the estimating function $y(x, \theta) = u^I(x, \theta, \xi_0)$ gives a smaller asymptotic variance than y^* at least for the sequence $\bar{\xi}_0$, contradicting the optimality of y^*. Consequently, we have

$$y^*(x, \theta) = c^*(\theta, \xi) u^I(x; \theta, \xi),$$

which means that the optimal y^* must be proportional to the projected score u^I at every (θ, ξ).

We hereafter consider the two cases separately.

(i) $c^*(\theta, \xi)$ is free of ξ, $c^*(\theta, \xi) = c^*(\theta)$. In this case, the projected score u^I should be also free of ξ, and the optimal y^* is equal to $u^I(x, \theta)$.

Conversely, suppose that the projected score u^I is free of ξ. Then the optimality of the estimating function given by u^I is proved in the following way. Let us take any $y \in C_1$ and decompose it as

$$y = c(\theta, \xi) u^I + y_2^I + y^A.$$

Then its asymptotic variance satisfies

$$v(y; \bar{\xi}_n) \geq \left(\sum c_i^2 A_i \right) \Big/ \left\{ \left(\sum c_i A_i \right)^2 \right\} \geq 1 \Big/ \left(\sum A_i \right) = v(u^I, \bar{\xi}_n).$$

This proves that u^I is optimal.

(ii) $c^*(\theta, \xi)$ depends on ξ. The proof is given in Appendix C. □

We have thus obtained the two cases: (i) The projected score u^I is invariant. (ii) The information subspace $R^I_{\theta, \xi}$ is one-dimensional and R^I is e-closed. But the second case seems rather exceptional. Godambe (1976) treated the following case. Suppose that there exists a statistic t with frequency function $h(t; \theta, \xi)$ such that the conditional frequency function of x given t depends only on θ, that is,

$$p(x; \theta, \xi) = f_t(x, \theta) h(t; \theta, \xi).$$

He showed that the optimal estimating function is given by

$$u^*(x, \theta) = \partial_\theta \log f_t(x, \theta).$$

In this case, it is easy to show that the nuisance subspace is given by

$$R^N_{\theta,\xi} = \{a(h;\theta,\xi) | E[a] = 0\},$$

and the information subspace $R^I_{\theta,\xi}$ is one dimensional, which is spanned by $u^*(x,\theta)$. Hence, by projecting the θ-score

$$u(x;\theta,\xi) = u^*(x,\theta) + \partial_\theta \log h(t;\theta,\xi)$$

onto $R^I_{\theta,\xi}$, we have

$$u^I(x,\theta) = u - E[u|h] = u^*(x,\theta).$$

The following theorem shows the relation of the projected score to the information unbiased estimators (Lindsay, 1982).

THEOREM 7. *The optimal estimator in C_1 is information unbiased in the case when the projected score u^I is invariant.*

PROOF. An estimating function $y(x,\theta)$ is said to be information unbiased when it satisfies

$$\langle y^2 \rangle = -\langle \partial_\theta y \rangle = \langle u, y \rangle.$$

By definition, we have

$$\langle (u^I)^2 \rangle_{\theta,\xi} = \langle u, u^I \rangle_{\theta,\xi},$$

so that the optimal estimating function $y = u^I$ is information unbiased. □

COROLLARY 2. *When and only when $P^I \partial_\theta s = 0$ and $P^I \partial_\theta r \neq 0$, there exists the optimal C_1-estimator in a ξ-exponential family. The optimal estimating function is $P^I \partial_\theta r = \partial_\theta r - E[\partial_\theta r | s]$ and is information unbiased.*

PROOF. Since the projected score u^I is given by

$$u^I = \xi P^I \partial_\theta s + P^I \partial_\theta r,$$

we can easily check conditions (i) and (ii) in Theorem 6. Condition (i) means $P^I \partial_\theta s = 0$, and the optimal estimating function is given by the ξ-free $u^I = P^I \partial_\theta r$, if it is not zero. On the other hand, the condition (ii) that $R^I_{\theta,\xi}$ is one dimensional does not hold in a ξ-exponential family, as is seen from (3.22), except for the case $s(x,\theta) = 0$, in which case condition (i) also holds. This proves the corollary. □

EXAMPLE 1 (continued). It is easy to show

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

so that from Corollary 2, there exist no optimal estimators in C_1. We can show that the optimal estimator exists in class C_2.

EXAMPLE 2 (continued). In Example 2, when f is linear, e.g., $f(\xi) = \xi$, the distributions are of ξ-exponential type with $s(x,\theta) = \delta_1(x_1) + \delta_1(x_2)$, $r(x,\theta) =$

$\theta\,\delta_1(x_1)$. Hence, $P^I\partial_\theta s = 0$, $P^I\partial_\theta r \neq 0$, so that from Corollaries 1 and 2, the optimal estimator exists in C_1, which is given by $P^I\partial_\theta r$ or

$$y(x,\theta) = \begin{cases} 0, & \text{when } x_1 = x_2, \\ 1 - \exp\theta, & \text{when } x_1 = 1 \text{ and } x_2 = 0, \\ \exp\theta, & \text{when } x_1 = 0 \text{ and } x_2 = 1. \end{cases}$$

Example 3. For the normal statistical model $N(\xi,\theta)$, let $x = (x_1,\dots,x_q)$ be q, $q \geq 3$, independent realizations from the same $N(\xi,\theta)$. The density function is

$$\begin{aligned} p(x,\theta,\xi) &= (2\pi\theta)^{-q}\exp\left\{-(1/2\theta)\sum_{j=1}^{q}(x_j-\xi)^2\right\} \\ &= \exp\left[\xi(qx./\theta) - q\left(z^2 + x.^2\right)/(2\theta)\right. \\ &\qquad \left.-q\left\{\left(\xi^2/\theta\right) + \log(2\pi\theta)\right\}/2\right], \end{aligned}$$

where

$$x. = \left(\sum_{j=1}^{q} x_j\right)\Big/q, \qquad z^2 = \sum_{j=1} (x_j - x.)^2/q.$$

This is a ξ-exponential family with

$$s(x,\theta) = qx./\theta, \qquad r(x,\theta) = -(q/2\theta)\left(z^2 + x.^2\right).$$

Then

$$\partial_\theta s = -qx./\theta^2 \quad \text{and} \quad \partial_\theta r = q\left(z^2 + x.^2\right)/(2\theta^2),$$

so that

$$P^I\partial_\theta s = 0 \quad \text{but} \quad P^I\partial_\theta r \neq 0.$$

Hence, from Corollary 1, the class C_1 is not empty, and furthermore from Corollary 2, the optimal estimator in C_1 is given by $P^I\partial_\theta r$. An easy calculation shows

$$P^I\partial_\theta r = \left\{qz^2 - (q-1)\theta\right\}/(2\theta^2).$$

This estimating function can be also written as $\hat{u} - \langle\hat{u}\rangle$, where $\hat{u} = u\{x;\,\theta,\hat{\xi}(x,\theta)\}$ and $\hat{\xi}(x,\theta)$ is the maximum likelihood estimator of ξ. That is, the optimal one is the bias-corrected maximum likelihood estimator of θ. In Kumon and Amari (1984), it was derived as the optimal one in the class C_2, but we found that it is optimal in the wider class C_1.

5. Discussion. (1) We have so far assumed that each x_i has the equal dimensionality q. However, there often occurs the case when the ith random variable $x_i = (x_{i1},\dots,x_{iq})$ consists of q_i independent observations x_{ij}, $j = 1,\dots,q_i$, from the same distribution $p(x;\,\theta,\xi_i)$, where q_i are not necessarily equal. We can generalize our theory to be applicable to the case when the dimension number q_i of the ith observation is not fixed. In this case, the probability density is written as $p(x;\, \theta, \xi, q)$, where q denotes the dimension of x explicitly. When x consists of q independent observations $x = (x_1, \dots, x_q)$, we have

$$p(x;\, \theta, \xi, q) = \prod_{j=1}^{q} p(x_j;\, \theta, \xi, 1). \tag{5.1}$$

An estimating function is also written as $y(x, \theta, q)$ by denoting the dimension number explicitly. For each point $(\theta, \xi) \in M$, we associate the Hilbert spaces $R_{\theta,\xi}(q)$ for all possible q's. The subspaces $R^T_{\theta,\xi}(q)$, $R^I_{\theta,\xi}(q)$, etc., are defined for each q separably. Similarly, the score functions $u(x;\, \theta, \xi, q)$ and $v(x;\, \theta, \xi, q)$ as well as $u^I(x;\, \theta, \xi, q)$ are defined for each possible q. The sequence $\bar{\xi} = (\xi_1, \xi_2, \dots)$ should be replaced by $\bar{\xi} = (\xi_1, q_1;\, \xi_2, q_2;\, \dots)$, where the dimension number in each observation is explicitly denoted. Then the structure theorems and optimality theorems for the C_1-estimators given in Sections 4.1 and 4.2 hold without any change, if we read that the conditions in the theorems hold for each q. In particular, when (5.1) holds, the optimal $y(x;\, \theta, q)$ is given by

$$y(x;\, \theta, q) = \sum_{j=1}^{q} y(x_j;\, \theta, 1),$$

provided the optimal $y(x;\, \theta, 1)$ exists in C_1 for $q = 1$. It should be remarked that the case with $q = 1$ is meaningless in some cases, although the optimal $y(x;\, \theta, q)$ exists.

EXAMPLE 3 (continued). If q_i are different in this example, the optimum y is given by

$$y(x;\, \theta, q) = qz^2 - (q-1)\theta = \sum_{j=1}^{q} (x_j - x.)^2 - (q-1)\theta.$$

(2) Following up the previous work of Kumon and Amari (1984), we have analyzed the structures of the estimating functions and derived optimality results. Based on the Hilbert bundle approach, we have identified the class C_1 as the direct sum of Hilbert spaces $R^I_{\theta,\xi} \oplus R^A_{\theta,\xi}$.

We then discussed the optimality result in the class C_1. It is related to the specific section of the information subbundle R^I, the projected score $u^I = P^I u$. It gives the optimal estimator if and only if it is invariant, i.e., free from ξ.

(3) Although the structure and optimality theorems given in Section 4 hold in general situations, there remain some difficulties. The first one is that it is not easy to examine whether or not the conditions of these theorems hold in general models other than the ξ-exponential family. The second and more serious one comes from the setting of the problem. One criterion of the universal optimality might be too strong to guarantee the existence of the optimal estimator in C_1. Then if there exists no optimal estimator in C_1, what shall we do? A customarily used way out is to reformulate the problem by regarding the nuisance parameter ξ as a random variable. The sequence $\bar{\xi}_n = (\xi_1, \dots, \xi_n)$ is then treated as

independent samples from some distribution specifying ξ; see, e.g., Bickel (1982), Begun, Hall, Huang and Wellner (1983) and Lindsay (1985). Another way is the minimax approach; e.g., see Hasminskii and Ibragimov (1983) and Nussbaum (1984). Our geometrical framework is useful for these approaches.

APPENDIX A

PROPOSITION. *Let C_1' be the set of zero-unbiased estimating functions $Y(\bar{x}_n, \theta)$ which are symmetric with respect to $\bar{x}_n = (x_1, \dots, x_n)$. The asymptotic variance of $Y \in C_1'$ is measured by the limit in $n \to \infty$ of $nv(Y, \bar{\xi}_n)$,*

$$v(Y, \bar{\xi}_n) = \langle Y^2 \rangle_{\bar{\xi}_n} / \langle YU(\bar{\xi}_n) \rangle_{\bar{\xi}_n}^2,$$

where $U(\bar{\xi}_n) = U(\bar{x}_n; \theta, \bar{\xi}_n) = \Sigma u(x_i; \theta, \xi_i)$. If there exists the optimal estimating function $Y^*(\bar{x}_n, \theta) \in C_1'$ satisfying $v(Y^*, \bar{\xi}_n) \le v(Y, \bar{\xi}_n)$ for any $Y \in C_1'$ and for any θ and $\bar{\xi}_n$, then it can be expressed as

$$Y^*(\bar{x}_n, \theta) = \sum_{i=1}^{n} y^*(x_i, \theta), \tag{A.1}$$

by using some $y^*(x, \theta)$.

PROOF. Let us introduce the linear spaces of random variables

$$R = \{ r(\bar{x}_n) | \langle r(\bar{x}_n) \rangle = 0, \langle r^2 \rangle < \infty \},$$
$$R_i = \{ r(x_i) | \langle r(x_i) \rangle = 0, \langle r^2 \rangle < \infty \}, \qquad i = 1, \dots, n.$$

For $r_1, r_2 \in R$, by defining the inner product by $\langle r_1, r_2 \rangle = \langle r_1 r_2 \rangle$, R becomes a Hilbert space and each R_i is a closed subspace of R. Note that for any $r_i \in R_i$ and $r_j \in R_j$, $i \ne j$, $\langle r_i, r_j \rangle = 0$ holds because of the independency of x_i and x_j. Let $Y^*(\bar{x}_n, \theta) \in R$ be the optimal estimating function in C_1'. Then by letting

$$y_i^* = E[Y^* | x_i] \tag{A.2}$$

be the conditional expectation of Y^* with respect to x_i, we have $y_i^* \in R_i$. In fact, y_i^* is the orthogonal projection of $Y^* \in R$ onto R_i. Thus, if we put $F = R_1 \oplus \cdots \oplus R_n$, according as the direct sum decomposition $R = F \oplus G$, Y^* is decomposed into

$$Y^* = Y_F^* + Y_G^*,$$

where $Y_F^* = \Sigma y_i^* \in F$, $Y_G^* \in G$, from which we have

$$v(Y^*, \bar{\xi}_n) = [\langle Y_F^{*2} \rangle + \langle Y_G^{*2} \rangle] / \langle Y_F^* U(\bar{\xi}_n) \rangle^2.$$

Note that $U(\bar{\xi}_n) \in F$. We show $Y_G^* = 0$. To do so, let us fix a sequence $\bar{\xi}_0 = (\xi_{01}, \dots, \xi_{0n})$. Then

$$v(Y^*, \bar{\xi}_0) \ge \langle Y_F^{*2} \rangle / \langle Y_F^* U(\bar{\xi}_0) \rangle^2 = v[Y_F^*(\bar{\xi}_0), \bar{\xi}_0],$$

where the equality holds if and only if $Y_G^*(\bar{\xi}_0) = 0$. Since Y^* is optimal and $\bar{\xi}_0$ is arbitrary, it follows that $Y_G^* = 0$ identically. The optimal $Y^* = Y_F^*$ is therefore expressed as the sum of y_i^*, $i = 1, \dots, n$. By the definition (A.2), the y_i^* depends on θ and $\bar{\xi}_{n,i}$ as $y_i^* = y^*(x_i; \theta, \bar{\xi}_{n,i})$, where $\bar{\xi}_{n,i}$ denotes the sequence obtained by deleting ξ_i from $\bar{\xi}_n$ as $\bar{\xi}_{n,i} = (\xi_1, \dots, \xi_{i-1}, \xi_{i+1}, \dots, \xi_n)$.

We show that each y_i^* is free from $\bar{\xi}_{n,i}$, i.e., $y_i^* = y^*(x_i, \theta)$. Let us put $\bar{\xi}_{n,1,2} = (\xi_3, \xi_4, \dots, \xi_n)$ and $\bar{x}_{n,1,2} = (x_3, x_4, \dots, x_n)$. Then from the symmetry $Y^*(x_1, x_2, \bar{x}_{n,1,2}, \theta) = Y^*(x_2, x_1, \bar{x}_{n,1,2}, \theta)$, we have

$$y^*\left(x_1; \theta, \xi_2, \bar{\xi}_{n,1,2}\right) + y^*\left(x_2; \theta, \xi_1, \bar{\xi}_{n,1,2}\right)$$
$$= y^*\left(x_2; \theta, \xi_2, \bar{\xi}_{n,1,2}\right) + y^*\left(x_1; \theta, \xi_1, \bar{\xi}_{n,1,2}\right).$$

From this it follows that there exist y and c such that

$$y^*\left(x_1; \theta, \bar{\xi}_{n,1}\right) = y\left(x_1; \theta, \xi_1, \bar{\xi}_{n,1,2}\right) + c\left(\bar{\xi}_n\right)$$

and

$$y^*\left(x_2; \theta, \bar{\xi}_{n,2}\right) = y\left(x_2; \theta, \xi_2, \bar{\xi}_{n,1,2}\right) - c\left(\bar{\xi}_n\right)$$

hold. Since y_1^* is free from ξ_1, by denoting a new $y(x_1; \theta, \bar{\xi}_{n,1,2}) = y(x_1; \theta, \xi_1^*, \bar{\xi}_{n,1,2})$ and $c(\bar{\xi}_{n,1}) = c(\xi_1^*, \bar{\xi}_{n,1})$ for any fixed ξ_1^*, we have

$$y^*\left(x_1; \theta, \bar{\xi}_{n,1}\right) = y\left(x_1; \theta, \bar{\xi}_{n,1,2}\right) + c\left(\bar{\xi}_{n,1}\right).$$

Furthermore, from the zero-unbiasedness of y^*, we see that

$$E\left[y\left(x_1; \theta, \bar{\xi}_{n,1,2}\right)\right] = -c\left(\bar{\xi}_{n,1}\right).$$

Clearly, the left-hand side is free from ξ_2, so that $c = c(\bar{\xi}_{n,1,2})$. By summarizing, we have

$$y_1^* = y\left(x_1; \theta, \bar{\xi}_{n,1,2}\right) + c\left(\bar{\xi}_{n,1,2}\right).$$

Let us next take the pair y_1^* and y_3^*. Then the same argument leads to

$$y_1^* = y\left(x_1; \theta, \bar{\xi}_{n,1,3}\right) + c\left(\bar{\xi}_{n,1,3}\right),$$

where $\bar{\xi}_{n,1,3} = (\xi_2, \xi_4, \dots, \xi_n)$, and in general

$$y_1^* = y\left(x_1; \theta, \bar{\xi}_{n,1,j}\right) + c\left(\bar{\xi}_{n,1,j}\right), \quad \text{for } j = 2, 3, \dots, n.$$

These $n - 1$ relations together imply that

$$y = y(x_1, \theta) \quad \text{and} \quad c = c(\theta).$$

Hence, we have

$$y_1^* = y(x_1, \theta) + c(\theta),$$

and in general

$$y_i^* = y(x_i, \theta) + c(\theta), \qquad i = 1, \dots, n,$$

showing that y_i^* is free from $\bar{\xi}_{n,i}$. □

APPENDIX B

For a vector parameter $\xi = (\xi^\mu)$, let ${}^{(\alpha)}\nabla_\mu$ be the α-covariant derivative along the ξ^μ-coordinate. Then the α-curvature is defined by the second-order differential operator in $S(R)$,

$$^{(\alpha)}K_{\mu\nu} = {}^{(\alpha)}\nabla_{\mu}{}^{(\alpha)}\nabla_{\nu} - {}^{(\alpha)}\nabla_{\nu}{}^{(\alpha)}\nabla_{\mu}.$$

As it stands, it measures the commutability between ${}^{(\alpha)}\nabla_{\mu}$ and ${}^{(\alpha)}\nabla_{\nu}$. In fact, it can be shown that a necessary and sufficient condition for the route independency of the α-parallel transport is

$$^{(\alpha)}K_{\mu\nu}r = 0,$$

for any $r \in S(R)$ and any μ, ν. By a direct calculation we have

$$^{(\alpha)}K_{\mu\nu}r = (1 - \alpha^2)\big(\langle \partial_{\mu}r\rangle v_{\nu} - \langle \partial_{\nu}r\rangle v_{\mu}\big)/4,$$

where

$$\partial_{\mu}r = \partial r/\partial\xi^{\mu}, \qquad v_{\nu} = \partial \log p(x; \theta, \xi)/\partial\xi^{\nu}.$$

Hence, ${}^{(\alpha)}K_{\mu\nu}r = 0$ identically if and only if $\alpha = \pm 1$.

APPENDIX C

We first define some notation. The projected score $u^I(x; \theta, \xi)$ at each (θ, ξ) spans a one-dimensional subspace $R^I_{1,\theta,\xi}$ of $R^I_{\theta,\xi}$. The information subspace $R^I_{\theta,\xi}$ can then be decomposed into

$$R^I_{\theta,\xi} = R^I_{1,\theta,\xi} + R^I_{2,\theta,\xi},$$

so that the information subbundle R^I is also decomposed into the Whitney sum

$$R^I = R^I_1 \oplus R^I_2,$$

where

$$R^I_1 = \bigcup_{\theta,\xi} R^I_{1,\theta,\xi}, \qquad R^I_2 = \bigcup_{\theta,\xi} R^I_{2,\theta,\xi}.$$

Now we assume that there exists the optimal estimating function $y^*(x, \theta)$ and that $c^* = c^*(\theta, \xi)$ depends on ξ. Then we first show that the bundle $R^I_2 \oplus R^A$ is e-closed. Contrary to that, suppose that there exist ξ_0, ξ_1 ($\neq \xi_0$) and $y^I_2 \in R^I_{2,\theta,\xi_0}$ such that

$$^{(e)}\pi^{\xi_1}_{\xi_0} y^I_2 = bu^I(\xi_1) + y^{I\prime}_2(\xi_1) + y^{A\prime}(\xi_1),$$

where

$$b \neq 0, \qquad y^{I\prime}_2(\xi_1) \in R^I_{2,\theta,\xi_1}, \quad y^{A\prime}(\xi_1) \in R^A_{\theta,\xi_1}.$$

We consider the following estimating function which depends on a scalar parameter ε,

$$\begin{aligned} y_{\varepsilon}(x, \theta) &= c^*(\theta, \xi_0)u^I(x; \theta, \xi_0) + \varepsilon y^I_2(x; \theta, \xi_0) \\ &= \{c^*(\theta, \xi_1) + \varepsilon b\}u^I(x; \theta, \xi_1) + \varepsilon y'(\xi_1), \end{aligned}$$

where

$$y'(\xi_1) \in R^I_{2,\theta,\xi_1} \oplus R^A_{\theta,\xi_1}.$$

For the sequence $\bar{\xi}_{01} = (\xi_0, \xi_1, \xi_0, \xi_1, \dots)$, by differentiating $v(y_\varepsilon, \bar{\xi}_{01})$ with respect to ε, we have at $\varepsilon = 0$,

$$\frac{d}{d\varepsilon} v(y_\varepsilon, \bar{\xi}_{01})|_{\varepsilon=0} = v(y^*, \bar{\xi}_{01})\big[2bA_0A_1c_0^*(c_1^* - c_0^*) \div \big(c_0^{*2}A_0 + c_1^{*2}A_1\big)\big(c_0^*A_0 + c_1^*A_1\big)\big],$$

which is nonzero. Therefore, there exists an $\varepsilon^* > 0$ such that

$$v(y_\varepsilon^*, \bar{\xi}_{01}) < v(y^*, \bar{\xi}_{01})$$

holds. But for the sequence $\bar{\xi}_0 = (\xi_0, \xi_0, \dots)$, we have of course $v(y^*, \bar{\xi}_0) < v(y_\varepsilon^*, \bar{\xi}_0)$, which is a contradiction. Hence, $R_2^I \oplus R^A$ must be e-closed when the optimal estimator exists in case (ii). This in turn implies that in the Whitney sum

$$R = R^N \oplus R_1^I \oplus R_2^I \oplus R^A,$$

$R^N \oplus R_1^I$ is m-closed. Thus, we have

$$^{(m)}\pi_\xi^{\xi_0} u(\xi) = c(\xi, \xi_0)u(\xi_0) + u^N(\xi_0),$$

where $u^N(\xi_0) \in R^N_{\theta, \xi_0}$. By projecting the preceding onto R^I_{θ, ξ_0}, we see that $\bar{U} = \bigcup_\xi u^I(\xi) = R^I_{\theta, \xi_0}$ is a one-dimensional space spanned by $u^I(\xi_0) = c^*(\theta, \xi_0)y^*(x, \theta)$, and thus R^I is e-closed. This completes the necessary part of (ii). To prove the converse, suppose that $\dim R^I_{\theta, \xi} = 1$ and R^I is e-closed. Then it is immediate to show that u^I can be written as

$$u^I(x; \theta, \xi) = c(\theta, \xi)u_0^I(x, \theta),$$

by some ξ-free $u_0^I(x, \theta)$. Furthermore, any $y \in C_1$ can be decomposed into $y = u_0^I + y^A$, $y^A = y^A(x, \theta) \in R^A_{\theta, \xi}$ at any ξ. Thus, we have

$$v(y, \bar{\xi}_n) \geq \left(\sum c_i^{-2}A_i\right)\Big/\left(\sum c_i^{-1}A_i\right)^2 = v(u_0^I, \bar{\xi}_n),$$

proving the optimality of u_0^I.

APPENDIX D

Let us define a Hilbert space $\tilde{R}_0$ by

$$\tilde{R}_0 = \left\{\tilde{r}(x) | E_{\theta, \xi}[\tilde{r}^2] < \infty \text{ for any } (\theta, \xi) \in M\right\}.$$

This is a Hilbert space with an inner product

$$\langle \tilde{r}, \tilde{s} \rangle_0 = E_{\theta_0, \xi_0}[\tilde{r}\tilde{s}], \quad \text{where we fix a point } (\theta_0, \xi_0) \in M.$$

A closed subspace R_0 of $\tilde{R}_0$ is defined by

$$R_0 = \left\{r(x) \in \tilde{R}_0 | \langle r, 1 \rangle_0 = 0\right\},$$

whose topology is given by the inner product $\langle\ ,\ \rangle_0$. The statistical model $M = \{p(x; \theta, \xi)\}$ has a natural topology of two-dimensional Euclidean space $\mathbb{R}^2$. Then $\tilde{R}_0 \times M$ is a (trivial) fiber bundle, and our

$$R = \bigcup_{\theta, \xi} R_{\theta, \xi}, \qquad R_{\theta, \xi} = \left\{r(x) | \langle r, 1 \rangle_{\theta, \xi} = 0\right\}$$

is its subbundle satisfying local triviality.

More formally, we define a map f from our R to the trivial bundle $R_0 \times M$,

$$f \colon R = \bigcup_{\theta,\xi} R_{\theta,\xi} \mapsto R_0 \times M,$$

by

$$f(r) = r - \langle r, 1 \rangle_0.$$

It is easy to show that f is a bijection.

A topology is then neutrally introduced in R by this mapping, i.e., by defining that a subset $\mathcal{O}$ of R is open iff $f(\mathcal{O})$ is open in $R_0 \times M$. By definition, it is evident that f is a homeomorphism between R and $R_0 \times M$. This proves local triviality. Therefore, the R forms a fiber bundle.

REFERENCES

AMARI, S.-I. (1982). Differential geometry of curved exponential families—curvatures and information loss. *Ann. Statist.* **10** 357–385.

AMARI, S.-I. (1985). *Differential-Geometrical Methods in Statistics. Lecture Notes in Statist.* **28.** Springer, New York.

AMARI, S.-I., BARNDORFF-NIELSEN, O. E., KASS, R. E., LAURITZEN, S. L. and RAO, C. R. (1987). *Differential Geometry in Statistical Inference.* IMS, Hayward, Calif.

ANDERSEN, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *J. Roy. Statist. Soc. Ser. B* **32** 283–301.

BARNDORFF-NIELSEN, O. E., COX, D. R. and REID, N. (1986). The role of differential geometry in statistical theory. *Internat. Statist. Rev.* **54** 83–96.

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.

BICKEL, P. J. (1982). On adaptive estimation. *Ann. Statist.* **10** 647–671.

BICKEL, P. J. and KLAASSEN, C. A. J. (1986). Empirical Bayes estimation in functional and structural models, and uniformly adaptive estimation of location. *Adv. in Appl. Math.* **7** 55–69.

COX, D. R. (1975). Partial likelihood. *Biometrika* **62** 269–276.

GODAMBE, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* **31** 1208–1211.

GODAMBE, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika* **63** 277–284.

GODAMBE, V. P. and THOMPSON, M. E. (1978). Some aspects of the theory of estimating equations. *J. Statist. Plann. Inference* **2** 95–104.

HASMINSKII, R. Z. and IBRAGIMOV, I. A. (1983). Efficient estimation in the presence of infinite dimensional incidental parameters. *Probability Theory and Mathematical Statistics. Lecture Notes in Math.* **1021** 195–229. Springer, New York.

KOBAYASHI, S. and NOMIZU, K. (1963). *Foundations of Differential Geometry* **1.** Wiley, New York.

KUMON, M. and AMARI, S.-I. (1984). Estimation of a structural parameter in the presence of a large number of nuisance parameters. *Biometrika* **71** 445–459.

LINDSAY, B. G. (1982). Conditional score functions: Some optimality results. *Biometrika* **69** 503–512.

LINDSAY, B. G. (1983). Efficiency of the conditional score in a mixture setting. *Ann. Statist.* **11** 486–497.

LINDSAY, B. G. (1985). Using empirical partially Bayes inference for increased efficiency. *Ann. Statist.* **13** 914–931.

NAGAOKA, H. and AMARI, S.-I. (1982). Differential geometry of smooth families of probability distributions. Technical Report 82-7, Univ. Tokyo.

NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32.

NUSSBAUM, M. (1984). An asymptotic minimax risk bound for estimation of a linear functional relationship. *J. Multivariate Anal.* **14** 300–314.

PFANZAGL, J. (1982). *Contributions to a General Asymptotic Statistical Theory. Lecture Notes in Statist.* **13**. Springer, New York.

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

STEIN, C. (1956). Efficient nonparametric testing and estimation. *Proc. Third Berkeley Symp. Math. Statist. Probab.* **1** 187–196. Univ. California Press.

DEPARTMENT OF MATHEMATICAL ENGINEERING
AND INSTRUMENTATION PHYSICS
FACULTY OF ENGINEERING
UNIVERSITY OF TOKYO
BUNKYO-KU
TOKYO
JAPAN