数理解析研究所講究録
第507巻 1983年 75-96

Estimation of structural parameter in the presence of a large number of nuisance parameters

公文雅之, 甘利俊一
By MASAYUKI KUMON AND SHUN-ICHI AMARI
Department of Mathematical Engineering and Instrumentation of Phisics,
University of Tokyo, Tokyo, Japan

SUMMARY

When the number of nuisance parameters increases in proportion to the sample size, the so-called Cramér-Rao bound does not necessarily give an attainable lower bound for the asymptotic varianses of estimators of the structural parameter. The present paper, setting up several classes of estimators, presents a new lower bound under the criterion named information uniformity. It is expressed as the sum of the Cramér-Rao bound (the inverse of partial information) and a certain non-negative term, which is derived by differential-geometrical considerations. The optimal estimating function meeting this lower bound, when it exists, is also obtained in a decomposed form. The first term is the modified score function, and the second term is, roughly speaking, given by the normal component of the mixture covariant derivative of some random variable. Furthermore, special versions of these results are given in concise forms, which are then applied to elucidate the efficiency of some famous examples.

Some key words: Asymptotic theory; Bound for asymptotic variance; Differential geometry; Estimating function; Exponential and mixture connections; Mixture curvature; Nuisance parameter; Structural parameter.

1. INTRODUCTION

The present paper treats the asymptotic theory of estimation of the structural parameter in the presence of nuisance parameters whose number increases in proportion to the number of independent observations. Let $x_1,\ldots, x_n$ be n independent vector observations, where x_i is assumed to be subject to the parametric density function $p(x; \theta, \xi_i)$. The θ is the common scalar parameter of interest and is called the structural parameter. The ξ_i; i=1,...,n, are scalar nuisance or incidental parameters which are assumed to take arbitrary values. The problem is to estimate the structural parameter without any knowledge of the true ξ_i. Here, the efficiency of estimators is evaluated by the asymptotic variance of consistent estimators when n is large.

Neyman & Scott (1948) treated the ploblem in detail and pointed out that the maximum likelihood method does not in general give a consistent estimator. Moreover, it is not in general efficient in the sense that the Cramér-Rao bound is not attained even asymptotically. Anderson (1970) showed a method of constructing a consistent estimator by the use of the conditional maximum likelihood estimator in a special class of models. Godambe (1976) obtained some optimality result in a very special but finite sample case. Lindsay (1982) extended this idea to a more general but asymptotic situation. Takeuchi (private communication) considered the problem from the minimax point of view and obtained some optimality results. See also Ibragimov & Khasminskii (1982). The concept of partial likelihood (Cox, 1975; Lindsay, 1980) is also important.

Inspite of these endeavours and progresses, the problem still remains unsolved. The present paper gives a new lower bound for the asymptotic variance in class C_2 of estimators shown below. The new bound is the sum of the Cramér-Rao bound given by the inverse of partial information and a new term connected with a kind of curvature of the statistical model. To obtain the new bound, we introduce class C_0 of estimators, which has so far been widely used in treating the present problem. We then define class C_1 of consistent estimators in C_0. We finally consider a subclass C_2 of C_1, called the class of uniformly informative estimators. This class is introduced in order to preclude a "super efficient" estimator which is efficient for a specific choice of ξ_i's but is not so for other choices of ξ_i's. This class is comparable to the class of information unbiasedness by Lindsay (1982), and the relation between the two classes is discussed in §7. The new lower bound is given for the estimators in the class C_2.

The new bound is still not necessarily attained even in the asymptotic sense. Moreover, it is rather difficult to calculate it for a given statistical model. We hence give a milder bound, which can be calculated immediately, by specializing the fundamental theorem. If an estimator attains to this bound, it is asymptotically optimal. A procedure is given to judge the attainability of this new bound and to obtain the optimal estimator when it exists. Two examples are shown in which the optimal estimator meets the new bound. We also give another method of obtaining the optimal estimator by applying the theorem to a special type of models. Two examples are then shown in which the optimal estimator is obtained by this method. It is interesting that the optimal estimators are the same as those obtained by the method of Lindsay (1982)

in three of the four examples, but in one examle, the optimal estimator in C_2 is better than the information unbiased one.

Although we do not here discuss a detailed differential-geometrical background, the present theory is constructed along the line of differential-geometrical thoughts in statistics (Amari, 1982 a, b; Amari & Kumon, 1983; Kumon & Amari, 1983). We are studying a more fundamental differential-geometrical theory for the present problem by the use of the concept of fibre bundles, which will appear in a forthcoming paper.

2. ESTIMATING FUNCTIONS AND CLASSES OF ESTIMATORS

We begin with describing classes for estimators $\hat{\theta} = \hat{\theta}(x_1,\ldots,x_n)$, where $x_1,\ldots,x_n$ are n independent observations, x_i from the density function $p(x;\theta, \xi_i)$; i=1,...,n. The following class C_0 of estimators has so far been widely used for the present problem.

(i)Class C_0: An estimator $\hat{\theta}$ is said to belong to C_0, when it is given by the solution of the estimating equation

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

The function $y(x,\theta)$, which does not depend on ξ, is called the estimating function of the estimator.

It is easy to show that the maximum likelihood estimator belongs to this class. Let $u(x;\theta,\xi)$ be the θ-score for the likelihood,

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

where

$$l(x;\theta,\xi) = \log p(x;\theta,\xi); \quad \partial_\theta = \partial/\partial\theta.$$

Assume that there exists a unique maximum likelihood estimator $\hat{\xi}_i = \hat{\xi}_i(x_i,\theta)$ of ξ_i for each fixed θ, which is obtained by solving the ξ-score equation

$$v(x_i; \theta, \xi_i) = \partial_\xi l(x_i; \theta, \xi_i) = 0,$$

where $\partial_\xi = \partial/\partial\xi$. The maximum likelihood estimator $\hat{\theta}$ is then obtained by solving

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

where

$$\hat{u}(x, \theta) = u\{x; \theta, \hat{\xi}(x, \theta)\}. \tag{2.3}$$

Therefore, this $\hat{u}$ is the estimating function of the maximum likelihood estimator, which belongs to C_0.

An estimator belonging to C_0 is not necessarily consistent. A one term Taylor expansion around the true θ yields

$$\Sigma \{y(x_i, \theta) + \partial_\theta y(x_i, \theta)(\hat{\theta} - \theta)\} = 0,$$

so that

$$n^{1/2}(\hat{\theta} - \theta) \cong - \{n^{-1/2} \Sigma y(x_i, \theta)\}/\{n^{-1} \Sigma \partial_\theta y(x_i, \theta)\} \tag{2.4}$$

is derived under mild regularity conditions. When the following convergence

$$\Sigma [y(x_i, \theta) - E_{\theta,\xi_i}\{y(x_i, \theta)\}]/n \to 0$$

is almost surely guaranteed, where $E_{\theta,\xi_i}(\cdot)$ implies the expectation with respect to $p(x; \theta, \xi_i)$, the estimator $\hat{\theta}$ is consistent when, and only when,

$$E_{\theta,\xi_i}\{y(x_i, \theta)\} = 0.$$

See Neyman & Scott (1948). This leads us to the following definition of class C_1 of consistent estimators.

(ii) Class C_1: An estimator $\hat{\theta}$ in C_0 is said to belong to C_1, when the expectation of the estimating function $y(x, \theta)$ vanishes for any θ and ξ,

$$E_{\theta,\xi}\{y(x, \theta)\} = 0. \tag{2.5}$$

When the values of the true ξ_i are known, one can construct the estimator that is optimal for these ξ_i. Obviously, such an estimator shows a bad performance when the values of ξ_i are different from the presumed ones. We are searching for the optimal estimator in the sense that its asymptotic variance is not larger than those of any other estimators for whatever values ξ_i take. For this purpose, class C_1 is sterile, because it is so wide that it may include the estimator which is optimal only for presumed ξ_i but not optimal for other ξ_i. Hence, we are forced to restrict the class of estimators to one whose optimal estimator is obtained without any knowledge of ξ_i, as was also indicated by Lindsay (1982).

In order to define class C_2 of estimators which satisfies the above requirement, some geometrical considerations are necessary. Let $M = \{p(x;\theta, \xi)\}$ be a two-dimensional statistical model parametrized by (θ, ξ). With each point (θ, ξ) of M, let us associate a linear space $R_{\theta,\xi}$ consisting of all the random variables r(x) which have vanishing expectations and finite second moments,

$$R_{\theta,\xi} = \{r(x) | E_{\theta,\xi}\{r(x)\} = 0,\ E_{\theta,\xi}\{r(x)^2\} < \infty\}. \tag{2.6}$$

We treat a random variable $r(x;\theta, \xi)$ depending on θ and ξ. Such a random variable is called a field when $r(x;\theta, \xi) \in R_{\theta,\xi}$, i.e., $E_{\theta,\xi}\{r(x;\theta, \xi)\} = 0$ for all θ and ξ. An estimating function $y(x, \theta)$ belonging to C_1 is such a field, which does not depend on ξ. The θ-score $u(x;\theta, \xi)$ and the ξ-score $v(x;\theta, \xi)$ are also examples of the field. Since $R_{\theta,\xi}$ is a vector space, we can define the inner product of two vectors or random variables r(x) and s(x) in $R_{\theta,\xi}$ by

$$\langle r(x), s(x)\rangle = E_{\theta,\xi}\{r(x)s(x)\} = \langle r(x)s(x)\rangle, \tag{2.7}$$

where $< \ >$ is also used to denote the expectation with respect to $p(x; \theta, \xi)$ when no confusion occurs.

Let $T_{\theta,\xi}$ be the two-dimensional subspace of $R_{\theta,\xi}$ spanned by the two score vectors u and v,

$$T_{\theta,\xi} = \{au(x; \theta, \xi) + bv(x; \theta, \xi)\}. \tag{2.8}$$

Geometrically speaking, $T_{\theta,\xi}$ is the tangent space of the manifold M of the statistical model. Let $N_{\theta,\xi}$ be the orthogonal complement of $T_{\theta,\xi}$ in $R_{\theta,\xi}$,

$$N_{\theta,\xi} = \{n(x) | n(x) \in R_{\theta,\xi}, < n(x), t(x) > = 0 \text{ for } t \in T_{\theta,\xi}\}. \tag{2.9}$$

Thus, $R_{\theta,\xi} = T_{\theta,\xi} \oplus N_{\theta,\xi}$, and any $r \in R_{\theta,\xi}$ can uniquely be decomposed into $r(x) = t(x) + n(x)$, $t \in T_{\theta,\xi}$, $n \in N_{\theta,\xi}$. The t is called the tangential component of r, and the n is called the normal component of r.

Any estimating function $y(x, \theta)$ belonging to C_1 can be decomposed into the following sum at each (θ, ξ).

$$y(x, \theta) = a(\theta, \xi)u + b(\theta, \xi)v + n(x; \theta, \xi), \tag{2.10}$$

where $n \in N_{\theta,\xi}$ is its normal component. Here, the θ-score term $a(\theta, \xi)u(x; \theta, \xi)$ carries information about θ, so that if a large coefficient $a(\theta, \xi)$ is assigned to a specific value of ξ, we have an estimator which is good for that value of ξ at the sacrifice of bad performances for other ξ. Hence, in order to get a uniformly good estimator for all unknown ξ's, we require that $a(\theta, \xi)$ does not depend on ξ. This is called the requirement of information uniformity (cf. Lindsay's information unbiasedness, 1982). Without loss in generality, we can put $a(\theta) = 1$, because for $y(x, \theta)$ with $a(\theta)$, $y'(x, \theta) = y(x, \theta)/a(\theta)$ yields the same estimating equation as $y(x, \theta)$.

(iii) Class C_2: The class C_2 of uniformly informative estimators consists of all the estimating functions with $a = 1$ in C_1.

3. ASYMPTOTIC VARIANCE OF ESTIMATOR IN C_2

We hereafter denote the partial derivative ∂_ξ by $\cdot$, as is shown in $\dot{r} = \partial_\xi r$. However, it should be noted that

$$< \dot{r}(x; \theta, \xi)> = 0$$

does not necessarily hold for a field $r(x; \theta, \xi)$. Hence, the ordinary partial derivative does not generate a field from a field. Instead, the operator D^e defined by

$$D^e r = \dot{r} - <\dot{r}>$$

generates a field. However, the operator D^m or shortly D defined by

$$Dr = \dot{r} + rv \tag{3.1}$$

is used more frequently. Since $\partial_\xi <r> = <Dr>$ holds, Dr is also a field when r is a field. From the geometrical point of view, D^e and $D = D^m$ are the exponential and mixture covariant derivatives, respectively. See Amari (1982 b) for details. For two fields r and s, it is easy to show the following identity

$$\partial_\xi <r, s> = <D^e r, s> + <r, Ds> = <\dot{r}, s> + <r, Ds>. \tag{3.2}$$

For an estimating function $y(x, \theta) = u + cv + n \in C_2$, by differentiating $<y> = 0$ with respect to ξ, $<y, v> = 0$ is obtained from $\dot{y} = 0$. Hence, the coefficient c is uniquely determined as

$$c = -<u, v>/<v, v>$$

in C_2, because of $<n, v> = 0$. Thus, the tangential part of $y \in C_2$ is always given by

$$w(x; \theta, \xi) = u - (<u, v>/<v, v>)v. \tag{3.3}$$

This w belongs to $T_{\theta,\xi}$, and is orthogonal to v. The square of the absolute value of w is

$$\bar{g}_{\theta\theta} = <w, w> = <u, u> - (<u, v>^2/<v, v>). \tag{3.4}$$

Since the Fisher information matrix of the model M is given by

$$\begin{bmatrix} g_{\theta\theta} & g_{\theta\xi} \\ g_{\xi\theta} & g_{\xi\xi} \end{bmatrix} = \begin{bmatrix} \langle u, u\rangle & \langle u, v\rangle \\ \langle v, u\rangle & \langle v, v\rangle \end{bmatrix}$$

the $\bar{g}_{\theta\theta}$ is also written as

$$\bar{g}_{\theta\theta} = g_{\theta\theta} - (g_{\theta\xi}^2/g_{\xi\xi}),$$

which is called the partial information. Obviously, $\bar{g}_{\theta\theta} \leq g_{\theta\theta}$, and the equality holds when, and only when, u and v are orthogonal, i.e., $g_{\theta\xi} = 0$.

We assume the existence of the limit

$$\bar{g} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \bar{g}_{\theta\theta}(\theta, \xi_i), \tag{3.5}$$

where (θ, ξ_i) is the true parameter for the ith observation x_i. For an estimating function $y \in C_2$, the existence of the following limit is also assumed,

$$\bar{g}_n = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} E_{\theta,\xi_i}\{n(x; \theta, \xi_i)^2\}, \tag{3.6}$$

where n is the normal component of the y. We call $\lim E\{n(\hat{\theta} - \theta)^2\}$ the asymptotic variance of an estimator $\hat{\theta}$. Then, under the above assumptions, the following lemma is derived.

LEMMA 1. The asymptotic variance of an estimator in C_2 is decomposed into

$$\lim E\{n(\hat{\theta} - \theta)^2\} = \bar{g}^{-1} + \bar{g}^{-2}\, \bar{g}_n. \tag{3.7}$$

Proof. Since the asymptotic variance is calculated from (2.4), we evaluate the denominator and the numerator of (2.4). By differentiating the identity $\langle y\rangle = 0$ with respect to ξ, we have

$$0 = \langle \partial_\theta y + uy\rangle = \langle \partial_\theta y\rangle + \langle w^2\rangle.$$

Hence, by the law of large numbers,

$$- \Sigma \partial_\theta y(x_i, \theta)/n \to \bar{g}$$

holds in the limit $n \to \infty$. From $< y> = 0$ and

$$< y^2> = \bar{g}_{\theta\theta} + < n^2>,$$

the central limit theorem applied to the right-hand side of (2.4) yields the desired asymptotic variance.

The lemma shows that the asymptotic variance is expressed as the sum of two positive terms. The first, being the inverse of the partial information $\bar{g}$, corresponds to the Cramér-Rao lower bound, and is common to all the estimators in C_2. The second depends on the estimating function y. Hence, the problem is to find the estimating function in C_2 which minimizes the second term or the expectation $<n^2>$.

4. MAIN TEOREMS

Let $s(x; \theta, \xi)$ be a field for which

$$\partial_\xi < s, n> = 0 \tag{4.1}$$

holds for the normal part n of any estimating functions $y = w + n \in C_2$. Note that the $n \in N_{\theta,\xi}$ is characterized by $\dot{n}(x; \theta, \xi) = - \dot{w}(x; \theta, \xi)$, because of $\dot{y} = 0$. Let $S = \{s(x; \theta, \xi)\}$ be the set of all such fields s. The set S is not empty, because when $s(x; \theta, \xi) \in T_{\theta,\xi}$, $< s, n> = 0$ holds so that $s \in S$. For $s \in S$, let $D^n s$ and $D^t s$ be the normal and tangential parts of Ds, respectively, in the decomposition $Ds = D^n s + D^t s$, $D^n s \in N_{\theta,\xi}$, $D^t s \in T_{\theta,\xi}$. Then, we define a function f(s) of $s \in S$ by

$$f(s) = < \dot{w}, s>^2 / <(D^n s)^2>. \tag{4.2}$$

Moreover, for the sequence ξ_i, let

$$\bar{f}(s) = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} f(s)_i, \tag{4.3}$$

where $f(s)_i$ is the value of f(s) evaluated at (θ, ξ_i).

THEOREM 1. The asymptotic variance of any estimator belonging to C_2 cannot be smaller than $\bar{g}^{-1} + \bar{g}^{-2}\bar{f}(s)$ for any $s \in S$. Hence, a lower bound for the asymptotic variance is given by

$$\bar{g}^{-1} + \bar{g}^{-2} \sup_{s \in S} \bar{f}(s). \tag{4.4}$$

Proof. For an estimating function $y = w + n \in C_2$, and $s \in S$, the Cauchy-Schwarz inequality gives

$$\langle n^2 \rangle \langle (D^n s)^2 \rangle \geqq \langle D^n s, n \rangle^2,$$

which is used to evaluate $\langle n^2 \rangle$. The right-hand side can be rewritten as

$$\langle D^n s, n \rangle = \langle Ds, n \rangle.$$

Since s satisfies $\partial_\xi \langle s, n \rangle = 0$, from (3.2) follows

$$\langle Ds, n \rangle = - \langle \dot{n}, s \rangle = \langle \dot{w}, s \rangle.$$

Hence, we have

$$\langle n^2 \rangle \geqq \langle \dot{w}, s \rangle^2 / \langle (D^n s)^2 \rangle$$

for any $s \in S$. This proves the theorem.

We have thus obtained a new lower bound for the asymptotic variance of an estimator. However, it is not sure whether this bound is attainable or not, i.e., whether there exists an estimator whose asymptotic variance is equal to this bound. The following theorem gives a sufficient condition for the existence of the optimal estimator meeting the bound.

THEOREM 2. When

$$y^* = w + \{\langle \dot{w}, s^* \rangle / \langle (D^n s^*)^2 \rangle\} D^n s^* \tag{4.5}$$

constructed from an $s^* \in S$ belongs to C_2, i.e., when y^* is free from ξ, it gives the optimal estimator meeting the lower bound.

Proof. When y^* belongs to C_2, we have

$$\langle y^{*2} \rangle = \langle w^2 \rangle + \{\langle \dot{w}, s^* \rangle^2 / \langle (D^n s^*)^2 \rangle\} = \bar{g}_{\theta\theta} + f(s^*),$$

and hence the asymptotic variance of the estimator given by y^* is

$$\bar{g}^{-1} + \bar{g}^{-2}\bar{f}(s^*).$$

This implies that the bound is met with this estimator and that $\bar{f}(s)$ is maximized at s^*.

The theorem shows that the normal part of the optimal y^* is derived by the mixture covariant derivative of $s^* \in S$. This manifests the very differential-geometrical aspect of the problem. Since a more systematic study on the structure of the set S is necessary for obtaining the condition on the existence of the optimal estimator in C_2, we transfer it to a forthcoming paper. Instead, in the following two sections, special types of $s \in S$ are searched for, which yield the optimal estimator in some cases.

5. TANGENTIAL RESTRICTION

Let T be a subset of S whose element $t(x; \theta, \xi)$ belongs to the tangent space $T_{\theta\xi}$ at any (θ, ξ). Such an element may be called the tangent field. Although it might be difficult to obtain the supremum of $\bar{f}(s)$ over S, it is easy to obtain it over the subset T. This also gives a lower bound, which is not necessarily attainable. It is a point that the latter bound can be calculated in an explicit form, which is related to a kind of local curvature of the statistical model M. Moreover, the bound can be met in some cases, as will be shown later.

For two random variables w and $v \in T_{\theta\xi}$, let us define a two-dimensional random variable vector by

$$h = (D^n w, \quad D^n v)^t, \tag{5.1}$$

which consists of the normal parts of the mixture covariant derivatives of w and v spanning $T_{\theta,\xi}$, t denoting the transposition of a vector so that h is a column vector. The vector h denotes the mixture curvature vector of the model M along the ξ-coordinate. The variance-covariance matrix of h is given by

$$H = \begin{bmatrix} h_{\theta\theta} & h_{\theta\xi} \\ h_{\xi\theta} & h_{\xi\xi} \end{bmatrix}, \tag{5.2}$$

where $h_{\theta\theta} = \langle (D^n w)^2 \rangle$, $h_{\theta\xi} = \langle D^n w, \ D^n v \rangle$, and $h_{\xi\xi} = \langle (D^n v)^2 \rangle$. The H is called the mixture curvature matrix of M.

We next define a two-dimensional column vector

$$c = (c_\theta, \ c_\xi)^t, \tag{5.3}$$

where $c_\theta = \langle \dot{w}, \ w \rangle$ and $c_\xi = \langle \dot{w}, \ v \rangle$. Since the vector c represents the coefficients of the exponential connection of M, it is called the exponential connection vector of M. From (5.2) and (5.3), we define the quadratic form

$$k = c^t H^{-1} c, \tag{5.4}$$

when H is non-singular. When H is singular, $k = c_\theta^2 h_{\theta\theta}^{-1}$ if $h_{\theta\theta} \neq 0$, and $k = c_\xi^2 h_{\xi\xi}^{-1}$ if $h_{\xi\xi} \neq 0$. We finally define

$$\bar{k} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} k_i, \tag{5.5}$$

where k_i is the value of k evaluated at $(\theta, \ \xi_i)$.

THEOREM 3. The supremum over T of $\bar{f}(s)$ is given by $\bar{k}$, so that $\bar{g}^{-1} + \bar{g}^{-2}\bar{k}$ is a lower bound for the asymptotic variance.

THEOREM 4. When the optimal y^* is derived from s^* belonging to T, it is directly given by

$$y^* = w + c^t H^{-1} h, \tag{5.6}$$

when H is non-singular and otherwise by

$$y^* = \begin{cases} w + c_\theta h_{\theta\theta}^{-1} h, & \text{if } h_{\theta\theta} \neq 0, \quad (5.7a) \\ w + c_\xi h_{\xi\xi}^{-1} h, & \text{if } h_{\xi\xi} \neq 0. \quad (5.7b) \end{cases}$$

That is, when the right-hand side does not depend on ξ, y^* is optimal with $s^* \in T$, and the lower bound of Theorem 3 is attained by this y^*.

Proof of Theorems 3 and 4. We search for the maximum of $\bar{f}(s)$ for $s \in T$. To this end, let us put

$$s = a_1 w + a_2 v,$$

where a_i are scalars depending on θ and ξ. The normal component $D^n s$ of the mixture covariant derivative of s is then given by $D^n s = a^t h$, where $a = (a_1, a_2)^t$. Hence, $f(s) = (a^t c)^2/(a^t H a)$. When H is non-singular, from the Cauchy-Schwarz inequality

$$(a^t c)^2 \leq (a^t H a)(c^t H^{-1} c)$$

follows Theorem 3. Since the equality holds when $a = H^{-1} c$, Theorem 4 follows. When H is singular, the proof is easy.

There are some examples in which the optimal estimators are obtained from Theorem 4.

Example 1. Let $x = (x_1, x_2)$ be a vector composed of two mutually independent random variables x_1 and x_2 from the normal distributions $N(\xi, 1)$ and $N(\theta\xi, 1)$, respectively. The density function $p(x; \theta, \xi)$ is given by

$$p(x; \theta, \xi) = (2\pi)^{-1} \exp\{-\tfrac{1}{2}(x_1 - \xi)^2 + (x_2 - \theta\xi)^2\}$$

From this, the following is easily calculated.

$$w = \xi(x_2 - \theta x_1)/\rho,\ v = (x_1 + \theta x_2 - \xi\rho),\ \text{where } \rho = \theta^2 + 1 > 0.$$

Meanwhile, the quantities related to Theorems 3, 4 are

$$D^n w = wv,\ c_\theta = \xi/\rho,\ c_\xi = 0,\ h_{\theta\theta} = \xi^2,\ h_{\theta\xi} = 0,\ h_{\xi\xi} \neq 0.$$

Hence, we have

$$y^* = w + c_\theta h_{\theta\theta}^{-1} D^n w = (x_2 - \theta x_1)(x_1 + \theta x_2)/\rho^2.$$

Since y^* is free from ξ, it is the optimal estimating function in C_2. As for the lower bound, we have

$$\bar{g}_{\theta\theta} = \xi^2/\rho,\ k = 1/\rho^2,$$

and thus, by letting $\mu^2 = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \xi_i^2$, we have the Cramér-Rao bound by $\bar{g}^{-1} = \rho/\mu^2$ and the best asymptotic variance by

$$\bar{g}^{-1} + \bar{g}^{-2}\bar{k} = (\rho\mu^2 + 1)/\mu^4.$$

Example 2. Let $x = (x_1,\ldots,x_p)$, where $x_j;\ j = 1,\ldots,p$ are p independent realizations from $N(\xi, \theta)$. The density function is

$$p(x;\ \theta, \xi) = (2\pi)^{-p/2} \exp\{-\frac{1}{2\theta} \sum_{j=1}^{p} (x_j - \xi)^2\}$$

The scores are given by

$$w = u = p\{z^2 + (x. - \xi)^2 - \theta\}/(2\theta^2),\ v = p(x. - \xi)/\theta,$$

where $x. = (\sum_{j=1}^{p} x_j)/p$, $z^2 = \{\sum_{j=1}^{p} (x_j - x.)^2\}/p$. The quantities of Theorems 3, 4 are

$$D^n v = p\{(p - 1)(x. - \xi)^2 - z^2\}/\theta^2,$$

$$c_\xi = -p/\xi^2,\ c_\theta = 0,\ h_{\xi\xi} = 2p(p - 1)/\theta^2,\ h_{\theta\xi} = 0,\ h_{\theta\theta} \neq 0.$$

Hence, we have

$$y^* = w + c_\xi h_{\xi\xi}^{-1} D^n v = p\{pz^2 - (p - 1)\theta\}/\{2\theta^2(p - 1)\},$$

which is again free from ξ, giving the optimal estimating function. We have for the lower bound

$$\bar{g}_{\theta\theta} = p/(2\theta^2),\ k = p/\{2(p - 1)\theta^2\},$$

and therefore

$$\bar{g}^{-1} = 2\theta^2 / p,\ \bar{g}^{-1} + \bar{g}^{-2}\bar{k} = 2\theta^2 / (p - 1).$$

We remark that $y^* = \hat{u}$ in Example 1, and $y^* = \hat{u} - <\hat{u}>$ in Example 2, where $\hat{u}$ is the estimating function of the maximum likelihood estimator defined by (2.3), i.e. efficient estimators are substantially given by the maximum likelihood method in these examples. We next show an example where y^* depends on ξ.

Example 3. Let $x = (x_1, x_2)$, where x_1 is subject to $N(\theta+\xi, 1)$, x_2 is subject to $N\{m(\xi), 1\}$ and $m(\xi)$ is a known function of ξ. Then,

$$p(x; \theta, \xi) = (2\pi)^{-1} \exp(-\tfrac{1}{2}[(x_1 - \theta - \xi)^2 + \{x_2 - m(\xi)\}^2]),$$

$$u = \tilde{x}_1,\ v = \tilde{x}_1 + \dot{m}(\xi)\tilde{x}_2,\ w = \dot{m}(\xi)(\dot{m}\tilde{x}_1 - \tilde{x}_2)/(1 + \dot{m}^2),$$

where we put $\tilde{x}_1 = x_1 - \theta - \xi$, $\tilde{x}_2 = x_2 - m(\xi)$. The quantities related to Theorems 3, 4 are

$$D^n w = \dot{m}(\dot{m}\tilde{x}_1 - \tilde{x}_2)(\tilde{x}_1 + \dot{m}\tilde{x}_2)/(1 + \dot{m}^2),\ D^n v = (\tilde{x}_1 + \dot{m}\tilde{x}_2)^2,$$

$$c_\theta = \dot{m}\ddot{m}(1 - \dot{m}^2)/(1 + \dot{m}^2)^2,\ c_\xi = \dot{m}\ddot{m}/(1 + \dot{m}^2),$$

$$h_{\theta\theta} = \dot{m}^2,\ h_{\theta\xi} = 0,\ h_{\xi\xi} = 3(1 + \dot{m}^2),$$

$$\bar{g}_{\theta\theta} = \dot{m}^2/(1 + \dot{m}^2),\ k = \ddot{m}^2(4\dot{m}^4 - 5\dot{m}^2 + 3)/\{3(1 + \dot{m}^2)^4\}.$$

Unfortunately, y^* depends on ξ except for the linear case $f(\xi) = a\xi + b$. In the linear case, $y^* = a(ax - y - b)/(1 + a^2)$, yielding the optimal estimating function, and since $k = 0$, the Cramér-Rao bound is attained. In the general case, we rather doubt the existence of the optimal estimator in C_2. Anyway, it is sure that the asymptotic variance of any estimator in C_2 is bounded by $\bar{g}^{-1} + \bar{g}^{-2}\bar{k}$.

6. SOME NON-TANGENTIAL CASE

It is generally difficult to search for the supremum of $\bar{f}(s)$ in the whole S. However, for some special statistical models, we can obtain the supremum and the optimal estimator, which are derived from a non-tangential $s \in S$.

To explain this, let us consider the following special exponential model whose score functions are given by

$$w = cw^*, \quad v = dv^* + e, \tag{6.1}$$

where c, d, e are functions of (θ, ξ) only, and w^* and v^* denote random variables independent of ξ, i.e. functions of x and θ only. For this model, let

$$y_* = w^*/v^*, \tag{6.2}$$

which does not depend on ξ, and assume that y_* belongs to C_2. Then, the optimality for the model (6.1) is delineated in a simple way.

THEOREM 5. When there exists the optimal estimating function for (6.1), it is given by y_*.

Proof. Let $\bar{y} = w + \bar{n}$ be the optimal estimating function in C_2. Then, as is easily shown, it satisfies

$$<\bar{y}, y> = <\bar{y}^2>$$

or equivalently

$$<\bar{n}, n> = <\bar{n}^2> \text{ for any } y = w + n \in C_2.$$

Since both $\bar{y}$ and y_* are assumed to belong to C_2, we have

$$<\bar{y}, w>/<w^2> = <y_*, w>/<w^2> = 1,$$

which becomes from (6.1) and (6.2)

$$<\bar{y}, w^*> = <w^{*2}/v^*>. \tag{6.3}$$

Nextly, by the assumption of $\bar{y}$, we have

$$<\bar{y}^2> = <\bar{y},\ y_*> = <\bar{y}w^*/v^*>. \tag{6.4}$$

By differentiating (6.3) and (6.4) with respect to ξ, once for (6.3), and twise for (6.4), the following is derived,

$$<\bar{y}w^*v^*> = <w^{*2}>,\ <(\bar{y}v^*)^2> = <\bar{y}w^*v^*>.$$

Then, in the Cauchy-Schwarz inequality

$$<(\bar{y}v^*)^2><w^{*2}> \geqq <\bar{y}w^*v^*>^2,$$

the above two relations guarantee the equality, which means $\bar{y}v^* \propto w^*$, or $\bar{y} \propto w^*/v^* = y_*$. Hence, the assumed $\bar{y}$ is nothing but the y_*, proving the theorem.

In the theorem, we presumed the existence of the optimal $\bar{y}$ for the model (6.1). As was suggested before, this existence problem will be studied in a forthcoming paper.

The optimal y_* in the theorem can also be constructed in line with Theorem 2. Let us define s_* by

$$s_* = y_*/<n_*^2>,$$

where n_* is the normal component of y_* in the decomposition $y_* = w + n_*$. Then, from the optimality of y_*, it follows that

$$<s_*,\ n> = <n_*,\ n>/<n_*^2> = 1$$

for any $y = w + n \in C_2$, showing that s_* belongs to the set S. Furthermore, a simple calculation plus the assumption $y_* \in C_2$ shows that the right-hand side of (4.5) in Theorem 2 becomes y_* when we substitute s_* for s^*.

The proposed s_* clearly does not belong to the tangential field T. We now show some examples for which the supremum of $\bar{f}(s)$ is not attained in T and hence the optimal estimating function cannot be found

by Theorems 3, 4, but is given by Theorem 5. We remark that following our study, the existence of the optimal estimating function is guaranteed in these examples.

Example 4. Let $x = (x_1,\ldots,x_p)$, where x_j; $j = 1,\ldots,p$ come independently from $N(\theta, \xi)$. The density function is

$$p(x; \theta, \xi) = (2\pi)^{-p/2} \exp\{-\frac{1}{2\xi}\sum_{j=1}^{p}(x_j - \theta)^2\}.$$

The scores are

$$w = u = p(x. - \theta)/\xi, \quad v = p\{z^2 + (x. - \theta)^2 - \xi\}/(2\xi^2),$$

which is the exponential model (6.1) with

$$c = p/\xi, \quad d = p/(2\xi^2), \quad e = -p/(2\xi),$$
$$w^* = x. - \theta, \quad v^* = z^2 + (x. - \theta)^2.$$

Let $y_* = w^*/v^*$, then, it is easy to examine $y_* \in C_2$ with $a_{y_*} = 1/p$, thus y_* is optimal in C_2. As for the lower bound, since $\bar{g}_{\theta\theta} = p/\xi$, by letting $\nu^2 = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\xi_i^{-1}$, we have

$$\bar{g}^{-1} = 1/(\nu^2 p), \quad \bar{g}^{-1} + \bar{g}^{-2}\bar{f}(s_*) = 1/\{\nu^2(p - 2)\}.$$

Example 5. Let $x = (x_1, x_2)$, where x_1 and x_2 are mutually independent, and subject to the exponential distributions $e\{0, 1/(\theta\xi)\}$ and $e(0, 1/\xi)$, respectively. The density function is

$$p(x; \theta, \xi) = \theta\xi\exp\{-(\theta x_1 + x_2)\xi\},$$

so that the scores are

$$w = \xi\{(x_2/\theta) - x_1\}/2, \quad v = -(\theta x_1 + x_2) + (2/\xi).$$

This again is the type (6.1) with

$$c = \xi/2, \quad d = -1, \quad e = 2/\xi, \quad w^* = (x_2/\theta) - x_1, \quad v^* = \theta x_1 + x_2.$$

Then, it is shown that $y_* \in C_2$ with $a_{y_*} = 2/3$, from which follows the optimality of y_*. Meanwhile, since $\bar{g}_{\theta\theta} = 1/(2\theta^2)$, we have for the lower bound

$$\bar{g}^{-1} = 2\theta^2, \ \bar{g}^{-1} + \bar{g}^{-2}\bar{f}(s_*) = 3\theta^2.$$

Note that in these examples, $y_* = \hat{u}$, i.e. the maximum likelihood estimators are efficient as in Examples 1, 2.

7. DISCUSSION

Various bounds can be obtained for the asymptotic variance of estimators $\hat{\theta}$ in various situations. They have proper meanings as follows. The simplest bound is the inverse of the Fisher information $v_1 = g^{-1}$, where $g = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} g_{\theta\theta}(\theta, \xi_i)$. This can be attained when all the ξ_i are identical, $\xi_i = \xi$ and the value of common ξ is known. The second bound is the inverse of the partial information $v_2 = \bar{g}^{-1}$ defined by (3.5). This bound is attained when all the ξ_i are identical but we do not know its value. Hence, the difference $v_2 - v_1 \geqq 0$ accounts for the loss of information which is carried by the exact knowledge of ξ. The third bound is given by Theorem 3, $v_3 = \bar{g}^{-1} + \bar{g}^{-2}\bar{k}$. This takes account that ξ_i are different and unknown, but the tangential restriction is imposed on S. The tangential restriction corresponds to the evaluation of the effect of distributed unknown ξ_i by using the local curvature of the statistical manifold. Hence, it cannot be attained unless the model is uniformly curved, i.e. the higher-order curvatures vanish. We can obtain stricter bounds further by evaluating the effect of higher-order curvatures successively. However, the results are tedious and complicated. We finally have the bound given by (4.4) in Theorem 1.

This bound is attainable, whenever there exists the optimal estimator in C_2. We have discussed the asymptotic efficiency of estimators in the class C_2 of the uniformly informative estimating functions. Meanwhile, Lindsay (1982) introduced the concept of information unbiasedness, which requires

$$< y^2 > + < \partial_\theta y > = 0$$

in the present terminology. By using the decomposition (2.10), this is rewritten as

$$a^2 \bar{g}_{\theta\theta} + < n^2 > = a \bar{g}_{\theta\theta}.$$

Then, it follows that an estimating function $y \in C_1$ is both information unbiased and uniformly informative, if and only if the ratio $< n^2 > / \bar{g}_{\theta\theta}$ is independent of ξ. We return to the examples to examine this point. In Examples 2, 4, and 5, the optimal uniformly informative $y^* \in C_2$ are shown to be information unbiased at the same time. They coincide with the estimators obtained by Lindsay's method of seeking the optimal weights for the conditional score function. However, the $y^* \in C_2$ in Example 1 is not information unbiased. It is proved that the asymptotic variance of the information unbiased one is larger than our uniformly informative y^* in Example 1. Of course, this is merely a warning to rely too much on the concept of the information unbiasedness. In general, it seems difficult to judge the superiority between the two criteria from the point of view of efficiency, because in some special examples treated by Godambe (1976), the optimal estimator exists in C_1 which is information unbiased but is not uniformly informative. Further researches are necessary in this respect.

REFERENCES

ANDERSON, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. J. Roy. Statist. Soc. B 32, 283 - 301.

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

AMARI, S. (1982 b). Differential geometry of curved exponential families -curvatures and information loss. Ann. Statist. 10, 357 - 85.

AMARI, S. & KUMON, M. (1983). Differential geometry of Edgeworth expansions in curved exponential family. Ann. Statist. Math. A 35, 1 - 24.

COX, D. R. (1975). Partial likelihood. Biometrika. 62, 269 - 76.

GODAMBE, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. Biometrika. 63, 277 - 84.

IBRAGIMOV, I. A. & KHASMINSKII, R. Z. (1982). Efficient estimation in the presence of infinite dimensional incidental parameters. IV USSR-JAPAN Sympo. Prob. Theo. Math. Statist. Abst. Comm. 1, 259-61.

KUMON, M. & AMARI, S. (1983). Geometrical theory of higher-order asymptotics of test, interval estimator and conditional inference. Pro. Roy. Soc. A, in press.

LINDSAY, B. G. (1980). Nuisance parameters, mixture models, and the efficiency of partial likelihood estimators. Phil. Trans. Roy. Soc. A 296, 639 - 65.

LINDSAY, B. G. (1982). Conditional score functions: Some optimality results. Biometrika 69, 503 - 12.

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