数理解析研究所講究録
第507巻 1983年 97-116

Geometrical Theory on Estimation of Structural Parameter in the Presence of Infinitely Many Nuisance Parameters

甘利俊一，公文雅之

Shun-ichi AMARI (Fac. of Eng., Univ. of Tokyo)

Masayuki KUMON (Fac. of Eng., Univ. of Tokyo)

1. Statement of the problem

Let $p(x; \theta, \xi)$ be a family of probability density functions of a (vector) random variable x with respect to some dominating measure P, specified by two scalar parameters θ and ξ. The set $M = \{ p(x; \theta, \xi) \}$ is a parametric statistical model which we presume. Let $x_1, x_2, \ldots, x_n$ be a sequence of independent observations such that the i-th observation x_i is a realization from the distribution $p(x; \theta, \xi_i)$, where both θ and ξ_i are unknown. In other words, the distributions of x_i are assumed to be specified by the common fixed but unknown parameter θ and also by the unknown parameter ξ_i whose value changes for each observation. We call θ the structural parameter and ξ the incidental or nuisance parameter. The problem is to find the asymptotic best estimator $\hat{\theta}_n = \hat{\theta}_n(x_1, x_2, \ldots, x_n)$ of the structural parameter θ, when the number n of observations are large.

An estimator $\hat{\theta}_n$ is said to be consistent, when $\hat{\theta}_n$ converges to the true parameter θ in the sense of mean square as n tends to infinity. The goodness of a consistent estimator is measured by its asymptotic variance defined by

$$\mathrm{Av}(\hat{\theta}, \Xi) = \lim_{n\to\infty} \mathrm{V}[\sqrt{n}(\hat{\theta}_n - \theta)]$$

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

$$\mathrm{Av}[\hat{\theta}, \Xi] \leq \mathrm{Av}[\tilde{\theta}, \Xi]$$

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

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

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

where $y(x, \theta)$ is a function of x and θ only, i.e., it does not depend on ξ. The function y is called the estimating function, and the above equation is called the estimating equation. Let C_1

be a subclass of C_0, consisting of all the consistent estimators in C_0. We can prove the following theorem.

Theorem 1. An estimator $\hat{\theta} \in C_0$ is consistent, if and only if its estimating function y satisfied

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

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

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

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

As will be shown later, the class C_1 is often too large to guarantee the existence of the best estimator. Some estimator is good for a specific sequence Ξ at the sacrifice of the bad performances for other Ξ. Hence, it is necessary to consider a more restricted class C_2 of estimators, which is a subclass of C_1. We can always decompose an estimating function $y(x, \theta)$ into the following sum

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

where

$$u(x; \theta, \xi) = \partial_\theta \ell(x; \theta, \xi) , \qquad v(x; \theta, \xi) = \partial_\xi \ell(x; \theta, \xi)$$

$$\mathcal{L}(x; \theta, \xi) = \log p(x; \theta, \xi) , \quad \partial_\xi = \partial/\partial\xi ,$$

and $n(x; \theta, \xi)$ is a random variable which is not correlated to u and v, i.e.,

$$E_{\theta,\xi}[un] = E_{\theta,\xi}[vn] = 0 .$$

It is remarked that $u = \partial_\theta \mathcal{L}(x; \theta, \xi)$ is the part which involves the necessary information for estimating θ. Hence, when $a(\theta, \xi)$ does not depend on ξ, the estimating function $y(x, \theta)$ includes the information with respect to the structural parameter θ uniformly in ξ. In this case, without loss of generality we can put $a(\theta, \xi) = 1$, because $y(x, \theta)/a(\theta)$ can be adopted as the equivalent estimating function.

Hence, an estimating function $y(x, \theta)$ with $a(\theta, \xi) = 1$ or the estimator $\hat{\theta}$ derived therefrom, is said to be uniformly informative. The class C_2 consists of all the uniformly informative estimators belonging to C_1. It is also easy to prove that $y(x, \theta) \in C_2$ can be decomposed into

$$y(x, \theta) = w(x; \theta, \xi) + n(x; \theta, \xi) \tag{1}$$

where

$$w(x; \theta, \xi) = u(x; \theta, \xi) = (E_{\theta,\xi}[uv] / E_{\theta\xi}[v^2]) v(x; \theta, \xi) .$$

Here,

$$E[w^2] = \bar{g}_{\theta\theta} = g_{\theta\theta} - g_{\theta\xi}^2/g_{\xi\xi} \tag{2}$$

is called the partial information, where $g_{\theta\theta}$, $g_{\xi\xi}$ and $g_{\theta\xi}$ are the respective components of the Fisher information matrix of the model M. By putting

$$g^0(\Xi) = \lim_{n\to\infty} \frac{1}{n} \sum E_{\theta,\xi_i} [n(x; \theta, \xi_i)^2] \tag{3}$$

$$\bar{g}(\Xi) = \lim_{n\to\infty} \frac{1}{n} \sum \bar{g}_{\theta\theta}(\theta, \xi_i) = \frac{1}{n} \sum \partial_\theta y(x_i; \theta, \xi_i) , \tag{4}$$

we have the following theorem.

Theorem 2. The asymptotic variance of an estimator $\hat{\theta}$ in C_2 is given by

$$Av[\hat{\theta}; \Xi] = \bar{g}^{-1} + \bar{g}^{-2} g^0 , \tag{5}$$

where $\bar{g}$ is common to all the estimators in C_2 and only g^0 depends on the estimator.

2. Geometrical preliminaries

In order to analyze the structures of the classes C_1 and C_2 of estimators, we need to introduce the geometrical notions connected with the manifold of the statistical model M.

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

be the vector space of all the random variables r(x) whose expectations with respect to $p(x; \theta, \xi)$ vanish and whose

variances are finite. This is a Hilbert space in which the inner product $\langle r(x), x(x) \rangle$ of two vectors of $R_{\theta,\xi}$ is defined by the covariance as

$$\langle r(x), s(x) \rangle = E_{\theta,\xi}[rs] .$$

It is easy to show that $u = u(x; \theta, \xi)$ and $v = v(x; \theta, \xi)$ belong to the $R_{\theta,\xi}$. Indeed,

$$E_{\theta,\xi}[u] = E_{\theta,\xi}[\partial_\theta \ell(x; \theta, \xi)] = 0 ,$$

$$E_{\theta,\xi}[v] = E_{\theta,\xi}[\partial_\xi \ell(x; \theta, \xi)] = 0$$

and $E_{\theta,\xi}[u^2] = g_{\theta\theta}$, $E_{\theta,\xi}[v^2] = g_{\xi\xi}$. The subspace

$$T_{\theta,\xi} = \{ au + bv \} \subset R_{\theta,\xi}$$

spanned by two vectors u and v is identified with the tangent space of the manifold at point (θ, ξ), because $u = \partial_\theta \ell$ and $v = \partial_\xi \ell$ can be identified with the natural basis vectors $\partial/\partial\theta$ and $\partial/\partial\xi$, respectively, of M associated with the coordinate system (θ, ξ). Obviously, $w(x; \theta, \xi)$ belongs to $T_{\theta,\xi}$. Let $O_{\theta,\xi}$ be the orthogonal complement of $T_{\theta,\xi}$ in $R_{\theta,\xi}$. Then, $R_{\theta,\xi}$ can be decomposed as the direct sum of the two orthogonal subspaces,

$$R_{\theta,\xi} = T_{\theta,\xi} \oplus O_{\theta,\xi} .$$

Let us attach one Hilbert space $R_{\theta,\xi}$ to each point (θ, ξ) of the manifold M. Mathematically speaking, such an aggregate is a fibre bundle with base space M. Let $r(x; \theta, \xi)$ be a random vector depending sufficiently smoothly on θ and ξ and let $r(x; \theta, \xi)$ belong to $R_{\theta,\xi}$ for each (θ, ξ). Such a random vector function $r(x; \theta, \xi)$ is called a vector field. Obviously, a vector field r satisfies $E_{\theta,\xi}[r(x; \theta, \xi)] = 0$. Now we introduce two differential operators ∇^e_ξ and ∇^m_ξ by

$$\nabla^e_\xi r(x; \theta, \xi) = \partial_\xi r - E_{\theta,\xi}[\partial_\xi r] , \tag{6}$$

$$\nabla^m_\xi r(x; \theta, \xi) = \partial_\xi r + ru . \tag{7}$$

It is easy to show that $\nabla^e_\xi r$ and $\nabla^m_\xi r$ again satisfies $E[\nabla^e_\xi r] = E[\nabla^m_\xi r] = 0$, so that they are vector fields under a certain regularity conditions. The operators ∇^e_ξ and ∇^m_ξ are called, respectively, the exponential covariant derivative and mixture covariant derivative with respect to ξ. A vector field $r(x; \theta, \xi)$ is said to be e-invariant when $\nabla^e_\xi r = 0$ holds. A vector field $r(x; \theta)$ which does not depend on ξ automatically satisfies $\nabla^e r = 0$. A field satisfying $\partial_\xi r = 0$ is said to be strongly e-invariant or shortly se-invariant. A se-invariant field is e-invariant.

We next define e- and m-parallel displacement of a vector. Let a(x) be a vector belonging to $R_{\theta,\xi'}$ of a point $(\theta, \xi') \in M$. We can extend it along the ξ-axis for all the points (θ, ξ) having the same fixed θ-coordinate such that the extended fields

are e- and m-invariant. Indeed the extended fields $a^e(x, \xi)$ and $a^m(x, \xi)$ are defined by solving the equation

$$\nabla^e_\xi a(x, \theta) = 0 \quad \text{and} \quad \nabla^m_\xi a(x, \theta) = 0 ,$$

respectively, where θ is fixed. We call $a^e(x, \xi)$ and $a^m(x, \xi)$, respectively, the e- and m-parallel displacements of $a(x)$ from (θ, ξ') to (θ, ξ) along the ξ-axis, and denote the parallel displacement operators by $\overset{e}{\Pi}{}^{\xi}_{\xi'}$ and $\overset{m}{\Pi}{}^{\xi}_{\xi'}$ as

$$\overset{e}{\Pi}{}^{\xi}_{\xi'} a(x) = a^e(x, \xi) ,$$

$$\overset{m}{\Pi}{}^{\xi}_{\xi'} a(x) = a^m(x, \xi) .$$

It is easy to show that the parallel displacements can be written explicitly as

$$\overset{e}{\Pi}{}^{\xi}_{\xi'} a(x) = a(x) \quad E_{\theta,\xi} [a(x)] , \tag{8}$$

$$\overset{m}{\Pi}{}^{\xi}_{\xi'} a(x) = \frac{p(x; \theta, \xi')}{p(x; \theta, \xi)} a(x) . \tag{9}$$

A se-invariant field $a(x)$ is by itself invariant under $\overset{e}{\Pi}{}^{\xi}_{\xi'}$,

$$\overset{e}{\Pi}{}^{\xi}_{\xi'} a(x) = a(x) .$$

The length of a vector or the angle of two vectors does not in general preserved by the e- and m-parallel displacements.

This is because the underlying e- and m-connections of the fibre bundle is not metric. However, the two connections are dual in the sense treated in Amari or more fundamentally in Nagaoka and Amari, so that the following important relation holds,

$$\langle r, s \rangle_{\xi'} = \left\langle \overset{e}{\Pi}{}^{\xi}_{\xi'} r, \ \overset{m}{\Pi}{}^{\xi}_{\xi'} s \right\rangle_{\xi} , \tag{10}$$

where $\langle r, s \rangle_{\xi'}$ denotes the inner product at (θ, ξ'), i.e., $\langle r, s \rangle_{\xi'} = E_{\theta, \xi'}[rs]$, where θ is fixed. The differential form of the above relation is

$$\partial_{\xi} \langle r, s \rangle = \langle \overset{e}{\nabla}_{\xi} r, s \rangle + \langle r, \overset{m}{\nabla}_{\xi} s \rangle . \tag{11}$$

Now, let us fix a point (θ, ξ) arbitrarily, and study the structure of the Hilbert space $R_{\theta, \xi}$ attached to this point. First, let $R^{T}_{\theta, \xi}$ be the subspace of $R_{\theta, \xi}$ spanned by the vectors

$$\left\{ \bigcup_{\xi'} \overset{m}{\Pi}{}^{\xi}_{\xi'} a(x) \;\middle|\; a(x) \in T_{\theta, \xi'} \right\} ,$$

i.e., $R^{T}_{\theta, \xi}$ is spanned by the vectors which are m-parallel displacements, to (θ, ξ), of the tangent vectors spanned by $u(x; \theta, \xi')$ and $v(x; \theta, \xi')$ at some (θ, ξ'). We call $R^{T}_{\theta, \xi}$ the tangential subspace, which obviously includes the tangent space $T_{\theta, \xi}$. We can decompose $R_{\theta, \xi}$ into the direct sum

$$R_{\theta, \xi} = R^{T}_{\theta, \xi} \oplus R^{O}_{\theta, \xi} , \tag{12}$$

where $R^O_{\theta,\xi}$ is the orthogonal complement of $R^T_{\theta,\xi}$. We call $R^O_{\theta,\xi}$ the orthogonal subspace. Let us further decompose R^T as follows. Let $R^N_{\theta,\xi}$ be the subspace of $R^T_{\theta,\xi}$ spanned by the vectors $\Pi^{\xi}_{\xi'} v(x; \theta, \xi')$ for all ξ' where θ is fixed. We call $R^N_{\theta,\xi}$ the nuisance subspace, because it is composed of the m-parallel displacements of all the v-vectors $v(x; \theta, \xi')$ which are responsible for changes in the nuisance parameter ξ. Then, we have the following orthogonal decomposition

$$R^T_{\theta,\xi} = R^N_{\theta,\xi} \oplus R^I_{\theta,\xi} , \tag{13}$$

where $R^I_{\theta,\xi}$ is the orthogonal complement of $R^N_{\theta,\xi}$ in $R^T_{\theta,\xi}$. It is called the information subspace, because it carries the information included in $u = u(x; \theta, \xi')$ but not in $R^N_{\theta,\xi}$ by m-parallelly displacing it to (θ, ξ). We thus have the full decomposition

$$R_{\theta,\xi} = R^I_{\theta,\xi} + R^N_{\theta,\xi} + R^O_{\theta,\xi} . \tag{14}$$

The above decomposition can be done at each point $(\theta, \xi) \in M$. Then, $R^I_{\theta,\xi}$, $R^N_{\theta,\xi}$ and $R^O_{\theta,\xi}$ are defined for all (θ, ξ). Let $R^A_{\theta,\xi}$ be a subspace of $R_{\theta,\xi}$ defined at every (θ, ξ). It is said to be e-closed or m-closed, respectively, when the e- or m-parallel displacements of a vector belonging to $R^A_{\theta,\xi'}$ from (θ, ξ') to (θ, ξ) remain in $R^A_{\theta,\xi}$ for all ξ' and ξ. It is clear from the definition that $R^N_{\theta,\xi}$ and $R^T_{\theta,\xi} = R^N_{\theta,\xi} \oplus R^I_{\theta,\xi}$ are m-closed. When $R^A_{\theta,\xi}$ is m-closed (e-closed), its orthogonal complement $\bar{R}^A_{\theta,\xi}$ is e-closed (m-closed).

This can be proved from the relation (10). Hence, $R^O_{\theta,\xi}$ and $R^O_{\theta,\xi} \oplus R^I_{\theta,\xi}$ are e-closed. However, $R^I_{\theta,\xi}$ is in general neither e-closed nor m-closed.

3. Conditions for existence of estimators in C_1 and C_2

Let $y(x, \theta)$ be an estimating function belonging to C_1. Then, it is a se-invariant vector field, because of the consistency condition $E_{\theta,\xi}[y(x, \theta)] = 0$. By differentiating this with respect to ξ, we have $\langle y, v \rangle = 0$ for all ξ. Since $y(x, \theta)$ is invariant under the e-parallel displacement, we have

$$\langle y, v \rangle_{\xi'} = \langle \overset{e}{\Pi}{}^{\xi}_{\xi'} y, \overset{m}{\Pi}{}^{\xi}_{\xi'} v \rangle = \langle y, \overset{m}{\Pi}{}^{\xi}_{\xi'} v \rangle .$$

This shows that y is orthogonal to the nuisance subspace $R^N_{\theta,\xi}$, i.e., y belongs to $R^I_{\theta,\xi} \oplus R^O_{\theta,\xi}$. We can hence prove

Lemma 1. A se-field $y(x, \theta)$ can uniquely be decomposed into

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

$y^I \in R^I_{\theta,\xi}$, $y^O \in R^O_{\theta,\xi}$ for every (θ, ξ). Conversely, given two vectors $a(x; \theta) \in R^I_{\theta,\xi}$ and $b(x; \theta) \in R^O_{\theta,\xi}$ at a point (θ, ξ), the sum $a(x; \theta) + b(x; \theta)$ can uniquely be extended to a se-field

$$y(x, \theta) = a(x; \theta) + b(x; \theta)$$

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

where $y^I(x; \theta, \xi) = a(x; \theta)$, $y^O(x; \theta, \xi) = b(x; \theta)$ at the point ξ.

It should be noted that y^I and y^O depends on ξ, although their sum is free of ξ. This is because $R^I_{\theta,\xi}$ is not e-closed. Since $R^O_{\theta,\xi}$ is e-closed, the I-part $y^I(x; \theta, \xi)$ of a se-field $y(x, \theta)$ is identically 0, when it vanishes at a point ξ. Obviously, for any $y(x, \theta) \in R^O_{\theta,\xi}$, $E[\partial_\theta y(x, \theta)] = - \langle y, u \rangle = 0$ holds, so that it is not related to the information for estimating θ. Hence, we have

Theorem 3. The class C_1 of the consistent estimators is not void, if and only if $R^I_{\theta,\xi}$ contains a vector other than 0.

In order to obtain the condition for the existence of C_2 estimators, we define the vector

$$\bar{u}^I(x; \theta, \xi; \xi') = \bar{g}^{-1}_{\theta\theta}(\xi') P^I \Pi^{m\,\xi}_{\xi'} u(x; \theta, \xi') ,$$

where P^I is the projection to R^I. THis is the $\bar{g}^{-1}(\xi')$ times the projection to R^I of the m-parallel displacement of the vector $u(x; \theta, \xi')$ from $(\hat{\theta}, \xi')$ to (θ, ξ). Obviously, all the $\bar{u}^I$'s span the whole $R^I_{\theta,\xi}$. The set of vectors $\bar{u}^I(x; \theta, \xi; \xi')$ where ξ' is changing and (θ, ξ) is fixed, forms a curve in $R^I_{\theta,\xi}$. The curve

$\bar{u}^I$ is said to have coplanarity, when it is on a hyprplane in $R^I_{\theta,\xi}$, i.e., when there exists a vector $w \in R^I_{\theta\xi}$ such that

$$\left\langle w, \bar{u}^I(x; \theta, \xi; \xi') \right\rangle_\xi = 1 . \tag{17}$$

We call such a w the information vector and denote it by $w^I(x; \theta, \xi)$, when it exists. We can prove that the information vector field $w^I(x; \theta, \xi)$ is unique, when it exists. The following theorem is important.

Theorem 4. The class C_2 of the uniformly informative estimators is not void, if and only if the curve $\bar{u}^I$ is coplanar. When C_2 is not void, any estimating function $y(x, \theta) \in C_2$ can be decomposed into

$$y(x, \theta) = w^I(x; \theta, \xi) + y^O(x; \theta, \xi) , \tag{18}$$

where $y^O \in R^O_{\theta,\xi}$. In other words, the I-part of any C_2 estimator is unique and is given by the information vector.

The proof is omitted. The theorem elucidates the structure of the uniformly informative estimators.

4. n-exponential family of distributions

Before obtaining the best estimators in C_1 and C_2, it is wise to obtain the explicit form of the decomposition (14) by using some special but widely used type of distributions. We

consider the following family of distributions whose density functions are given by

$$p(x;\theta,\xi) = \exp\{\xi s(x,\theta) + r(x,\theta) - \psi(\theta,\xi)\} \quad . \tag{19}$$

This type of distributions is characterized by the fact that there exists a scalar sufficient statistic for the nuisance parameter ξ, when θ is known. Indeed, $s(x, \theta)$ is the sufficient statistic. The above type of distributions is called the nuisance-exponential family or shortly the n-exponential family.

In an n-exponential family of distributions, the tangent vectors are given by

$$u = \partial_\theta \ell = \xi \partial_\theta s - \partial_\theta r - \partial_\theta \psi ,$$

$$v = \partial_\xi \ell = s - \partial_\xi \psi .$$

The m-parallel displacement is expressed as

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

Hence, the parallel displacements of $u(x; \theta, \xi')$ and $v(x; \theta, \xi')$ from (θ, ξ') to (θ, ξ) are given, respectively, by

$$\Pi^{m\,\xi}_{\xi'} u = \{\xi' \partial_\theta s - \partial_\theta r - \partial_\theta \psi(\xi')\} \exp\{(\xi' - \xi)s - \psi(\xi') + \psi(\xi)\} ,$$

$$\Pi^{m\,\xi}_{\xi'} v = \{s - \partial_\xi \psi(\xi')\} \exp\{(\xi' - \xi)s - \psi(\xi') + \psi(\xi)\} .$$

The nuisance subspace $R^N_{\theta,\xi}$ is spanned by $\prod^m v$ for all ξ'. As is known from the theory of Laplace transformation, the linear combination of $\prod^m v$'s yields a function $f(s) - c(\theta, \xi)$, where $c(\theta, \xi) = E_{\theta,\xi}[f\{s(x, \theta)\}]$. Hence, the nuisance subspace $R^N_{\theta,\xi}$ is composed of the following random variables

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

for arbitrary functions f. Similarly, the subspace $R^T_{\theta,\xi}$ spanned by $\prod^m u$'s and $\prod^m v$'s are written as

$$R^T_{\theta,\xi} = \{f'(s)\partial_\theta s + f(s)(\xi\partial_\theta s + \partial_\theta r) + h(s) - c\},$$

where

$$c(\theta, \xi) = E_{\theta,\xi}[f'(s)\partial_\theta s + f(s)(\xi\partial_\theta s + \partial_\theta r) + h(s)],$$

f' is the derivative of f with respect to s, f and h are arbitrary functions in s and they may depend on θ and ξ. The information subspace $R^I_{\theta,\xi}$ is the orthogonal complement of $R^N_{\theta,\xi}$ in $R^T_{\theta,\xi}$. Since $R^N_{\theta,\xi}$ is generated by the random variable $s(x, \theta)$, the projection of a random variable a(x) to $R^N_{\theta,\xi}$ is given by the conditional expectation $E[a(x) \mid s(x, \theta)]$, which is a function of s. Hence, the projection of $a(x) \in R^T_{\theta,\xi}$ to $R^I_{\theta,\xi}$ is given by

$$P^I a(x) = a(x) - E[a \mid s] . \tag{20}$$

Theorem 5. The information subspace R^I of an n-exponential family is given by

$$R^{I}_{\theta,\xi} = \left\{ (f'(s) + \xi f(x))P^{I}\partial_{\theta} s + f(s)P^{I}\partial_{\theta} r \right\} . \qquad (21)$$

It is easy to see that the class C_1 is not void in the present case, unless both $\partial_\theta s$ and $\partial_\theta r$ are functionally dependent on s. Hence, there always exists a consistent estimator in an n-exponential family. However, this is not true for a general family. We give an example.

Example. Let $x = (y_1, y_2)$ be a pair of random variables y_1, y_2, which are independent and take on two values 0 and 1. We assume the following probability law:

$$P(y_1 = 0) = 1/(1 + \exp\{\theta + \xi\}) ,$$

$$P(y_2 = 0) = 1/(1 + \exp\{f(\xi)\}) ,$$

where $f(\xi)$ is a known function, and $P(y_1 = 1) = 1 - P(y_1 = 0)$, $P(y_2 = 1) = 1 - P(y_2 = 0)$. By the use of the function $\delta_1(z)$ which is equal to 1 when $z = 1$ and otherwise equal to 0, the above probability can be written as

$$\ell(x; \theta, \xi) = (\theta + \xi)\delta_1(y_1) + f(\xi)\,\delta_1(y_2)$$

$$- \log\{1 + \exp(\theta + \xi)\}\{1 + \exp(f(\xi))\} .$$

Hence, the distributions are of the n-exponetial type only when $f(\xi)$ is a linear function. It is easy to show that $R_{\theta,\xi}$ can be spanned by three random variables,

$$\delta_1(y_1)\delta_1(y_2) - c_{11} , \qquad \delta_1(y_1)\{1 - \delta_1(y_2)\} - c_{10} ,$$

$$\{1 - \delta_1(y_1)\}\delta_1(y_2) - c_{01} ,$$

where the constants c_{ij} are added such that the expectation of the above vanish. Hence, it is a three-dimensional vector space. We can prove that the nuisance subspace $R^N_{\theta,\xi}$ is also three-dimensional, when $f(\xi)$ is not linear. Hence, $R^I_{\theta,\xi} = \{0\}$, and there exist no consistent estimators, unless $f(\xi)$ is linear. When f is linear, for example $f(\xi) = \xi$, the distributions are of the n-exponential type with $s(x, \theta) = \delta_1(y_1) + \delta_1(y_2)$, $r(x, \theta) = \delta_1(y_2)$. Hence, $\partial_\theta s = 0$, $\partial_\theta r = \delta_1(y_2)$, and $R^N_{\theta,\xi}$ is composed of random variables $f(s) - c$. Since

$$P^1 \partial_\theta s = 0 , \qquad \text{i.e.,} \qquad \partial_\theta s = E[\partial_\theta s \mid s] ,$$

$R^I_{\theta,\xi}$ is composed of the random variables of the type $f(s)E[\delta_1(y_2) \mid s]$.

Let w^u be the vector field obtained by projecting the θ-score $u = \partial_\theta \ell$ to the information subspace,

$$w^u(x; \theta, \xi) = P^I u(x; \theta, \xi) . \tag{22}$$

In the case of an n-exponential family, it is given by

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

We call it the u-vector field. It is e-invariant, when and only when $E[\partial_\theta s \mid s] = \partial_\theta s$. The u-vector field is, in this case, equal to the conditional score,

$$w^u(x;\ \theta\ ,\ \xi\ ) = P^I \partial_\theta r = \partial_\theta r - E[\partial_\theta r \mid s]$$

conditioned on s.

We next study the coplanarity of the vectors $\bar{u}^I(x;\ \theta\ ,\ \xi\ ;\ \xi')$ in the n-exponential family. When they are coplanar, there exists the information vector field $w^I(x;\ \theta\ ,\ \xi\ )$. Since any vector in $R^I_{\theta,\xi}$ can be written in the form

$$w(x;\ \theta\ ,\ \xi) = \left\{ f'(x) + \xi f(x) \right\} P^I \partial_\theta s + f(s) P^I \partial_\theta r\ , \qquad (23)$$

by using a suitable function $f(s;\ \theta\ ,\ \xi)$, the equation determing the information vector w^I

$$\left\langle w(x;\ \theta,\ \xi)\ ,\quad P^I \prod_{\xi'}^{m} u(x;\ \theta, \xi') \right\rangle_\xi = \bar{g}_{\theta\theta}(\xi')$$

can be rewritten as

$$\left\langle w,\ P^I(\xi' \partial_\theta s + \partial_\theta r) \right\rangle_{\xi'} = \bar{g}_{\theta\theta}(\xi')\ . \qquad (24)$$

This is an integro-differential equation obtaining $f(s; \theta, \xi)$, and the information vector field $w^I(s; \theta, \xi)$ is given by (23).

When $P^I\partial_\theta r = 0$ or $\partial_\theta r = E[\partial_\theta r \mid s]$, i.e., when $\partial_\theta r$ is functionally dependent on s, the equation reduces to

$$E_{\theta,\xi'}[g(s)V[\partial_\theta s \mid s]] = \bar{g}_{\theta\theta}(\xi')/\xi' , \tag{25}$$

where we put $g(s) = f'(s) + \xi f(s)$. Obviously, the solution of the above integral equation $g(s)$ does not depend on ξ. Therefore, when $P^I\partial_\theta r = 0$, the information vector field w^I is e-invariant.

5. Asymptotically best estimators in C_1 and C_2

We have so far elucidated the structures of the estimating functions in C_1 and C_2. We can now prove the following two fundamental theorems.

Theorem 6. There exists the best estimator in C_1, when and only when the u-vector field w^u is e-invariant. The best estimating function $y(x, \theta)$ is given by the e-invariant u-vector w^u.

Theorem 7. There exists the best estimator in C_2, when and only when the information vector $w^I(x; \theta, \xi)$ is e-invariant. The best estimating function y is given by the invariant information vector w^I.

The proofs are omitted. Instead, we apply the theorems to the n-exponential family to get the following results.

Theorem 8. In the n-exponential family, the optimal estimator exists in C_1, when and only when $P^I\partial_\theta s = 0$, and the optimal estimating function y is given by the conditional score function $P^I\partial_\theta \ell = P^I\partial_\theta r = \partial_\theta r - E[\partial_\theta r|s]$.

This includes the distributions treated by Godambe as a special case. It is easy to show that the best estimator is always information unbiased in the sense of Lindsay in this case. The example of the previous section belongs to this type, when f is linear.

Theorem 9. In the n-exponential family, the optimal estimator exists in C_2, when $P^I\partial_\theta r = 0$, i.e., when $\partial_\theta r$ is functionally dependent on s, and the best estimating function is given by the e-invariant information vector w^I.

We can solve various examples by this method. The distributions treated by Lindsay belong to a special case of $P^I\partial_\theta r = 0$. It should be remarked that the best estimator in C_2 is not necessarily information unbiased.