Bernoulli **3**(1), 1997, 29–54

Information geometry of estimating functions in semi-parametric statistical models

SHUN-ICHI AMARI[1,2]* and MOTOAKI KAWANABE[2]

[1]*Department of Mathematical Engineering and Information Physics, Faculty of Engineering, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan*
[2]*Frontier Research Program, Institute of Physical and Chemical Research (RIKEN), Wako-shi, 351-01 Japan*

For semi-parametric statistical estimation, when an estimating function exists, it often provides an efficient or a good consistent estimator of the parameter of interest against nuisance parameters of infinite dimensions. The present paper elucidates the structure of estimating functions, based on the dual differential geometry of statistical inference and its extension to fibre bundles. The paper studies the following problems. First, when does an estimating function exist and what is the set of all the estimating functions? Second, how are the asymptotic variances of the estimators derived from estimating functions and when are the estimators efficient? Third, how do we adaptively choose a practically good (quasi-)estimating function from the observed data? The concept of m-curvature freeness plays a fundamental role in solving the above problems.

Keywords: dual geometry; dual parallel transport; efficient score function; estimating function; Hilbert fibred structure; m-curvature free; semi-parametric model

1. Introduction

A semi-parametric statistical model treats a family of probability distributions $\{p(x, \theta, k)\}$ specified by a finite-dimensional parameter θ of interest and an infinite-dimensional nuisance parameter k. Estimation of the parameter of interest in such a model has attracted statisticians for many years, because various important problems are formulated in terms of semi-parametric models. When the nuisance parameter is finite-dimensional, a fundamental role is played by the efficient or projected score function, which is the projection of the score function on the space orthogonal to the score functions of the nuisance parameters. The Cramér–Rao-type inequality has been established in terms of the efficient Fisher information and the bound is asymptotically attainable.

It is not easy to generalize these results to the semi-parametric case. Levit (1978), Begun *et al.* (1983) and Small and McLeish (1988; 1989) defined the efficient Fisher information in

* To whom correspondence should be addressed.

1350-7265

the semi-parametric model by using the projected score, and showed that the Cramér–Rao-type inequality holds. Only recently has its asymptotic attainability been elucidated under certain regularity conditions (Bickel *et al.* 1993) by various efforts based on functional analysis (for example, Ritov and Bickel 1990; van der Vaart 1991; see also Groeneboom and Wellner 1992; Pfanzagl 1990). It is also known that there exist cases where the bound is not asymptotically attainable (see Hasminskii and Ibragimov 1983; Ritov and Bickel 1990). Estimation procedures are generally very complicated because of the infinite dimensionality of the nuisance parameter. Moreover, as is clearly shown in Bickel *et al.* (1993), the rigorous analytical foundation involves difficulties, which we do not state in the present paper.

The present paper aims to elucidate the differential geometrical structure underlying the semi-parametric model from the point of view of the dual differential geometry of statistical inference (Amari 1985; Barndorff-Nielsen 1986; Murray and Rice 1993). This may complement the analytical and geometrical treatments of Bickel *et al.* (1993). It should be noted that our mathematical treatments are not rigorous in the sense of pure mathematics, because we do not give rigorous regularity conditions for theorems. Also, we do not enter into the rigorous formulation of the dual differential geometry of the relevant function space, since the appropriate pure mathematical background has not yet been developed. The present paper belongs to the field of 'theoretical mathematics' in the sense of Jaffe and Quinn (1993). The paper proposes interesting geometrical ideas and some heuristic arguments which lead to useful statistical procedures but are not necessarily mathematically rigorous.

There is a class of estimators obtained by solving a simple equation of the type

$$\sum_{i=1}^{n} \mathbf{y}(x_i, \theta) = 0, \tag{1.1}$$

where $x_1, \cdots, x_n$ are independent observations from an identical distribution $p(x, \theta, k)$ of a semi-parametric model. Here, the vector function $\mathbf{y}(x, \theta)$ should satisfy

$$\mathrm{E}_{\theta,k}[\mathbf{y}(x, \theta)] = 0$$

for all k, where $\mathrm{E}_{\theta,k}$ denotes the expectation with respect to the distribution specified by θ and k. Such a function $\mathbf{y}(x, \theta)$ which does not depend on the unknown nuisance parameter k is called an *estimating function* (Godambe 1976) and such an estimator is also called an M-estimator (Huber 1981; Bickel *et al.* 1993). It gives a practically tractable method of estimation (see, for example, McLeish and Small 1988; Godambe 1991; Godambe and Heyde 1987). Further regularity conditions will be imposed later. The present paper aims to elucidate estimating functions and their efficiency from the geometrical point of view by introducing Hilbert bundles and their dual parallel transports.

The class of M-estimators does not necessarily include efficient estimators. This is one of the points discussed by Pfanzagl (1990), who criticized the method of estimating functions for this reason. On the other hand, it is widely known that the class includes efficient estimators in many practically important cases, and moreover, it gives tractable and robust estimating procedures. Therefore, it is not wise to ignore this class of estimators. On the contrary, it is important to study such fundamental problems as the following:

(1) When does an estimating function exist?
(2) What is the set of all the estimating functions?
(3) What is the best estimator derived from estimating functions?
(4) When does the best estimating function give an efficient estimator? In the case where the best one is not efficient, what is the amount of loss of information caused by using the estimation function method?
(5) How to choose a good (quasi-)estimating function based on the observed data?

The present paper treats these problems. Here, a quasi-estimating function implies that it is one obtained from the observed data adaptively. The concept of m-information-curvature freeness plays a fundamental role.

The method of the present paper is based on informal notes by Amari (1987a) and is an extension of Amari and Kumon (1988). The theory is motivated by the information geometry (Amari 1985; Nagaoka and Amari 1982; Barndorff-Nielsen 1986; Murray and Rice 1993), which studies the structure of the manifold of probability distributions or a statistical model by introducing a Riemannian metric due to the Fisher information and a pair of dual affine connections. It has been proved to be a powerful method in various fields of information science (Amari 1985; 1987b; Amari and Han 1989; Amari and Kumon 1988; Okamoto *et al.* 1991; Amari *et al.* 1992), although the pure mathematical background has been established mostly in finite-dimensional cases. Infinite-dimensional cases are being studied by mathematicians. See Friedrich (1991), Lafferty (1988) and Kanbayashi (1994).

2. Semi-parametric statistical models and estimating functions

Let $p(x, \theta, k)$ be a probability density function of a random variable x with respect to a common dominating measure $\mu(\mathrm{d}x)$, specified by two kinds of parameters $\theta = (\theta^1, \cdots, \theta^m)$ and k, where $\theta \in \Theta$ is a finite-dimensional vector, Θ is an open set of $\mathbb{R}^m$ and $k \in K$ is an infinite-dimensional parameter, typically occupying a space of functions. The set of distributions $S = \{ p(x, \theta, k) \}$ is called a semi-parametric statistical model, where θ is called the parameter of interest and k is called the nuisance parameter.

Let $\mathbf{y}(x, \theta) = [y_i(x, \theta],\ i = 1, \cdots, m$, be a vector-valued smooth function of θ, not depending on k, of the same dimension as θ. Such a function is called an estimating function when it satisfies the following conditions,

$$\mathrm{E}_{\theta, k}[\mathbf{y}(x, \theta)] = 0, \tag{2.1}$$

$$\det |\mathrm{E}_{\theta, k}[\partial_\theta \mathbf{y}(x, \theta)]| \neq 0, \tag{2.2}$$

$$\mathrm{E}_{\theta, k}[\, \|\mathbf{y}(x, \theta)\|^2] < \infty, \qquad \mathrm{E}_{\theta, k}[\, \|\partial_\theta \mathbf{y}(x, \theta)\|^2] < \infty, \tag{2.3}$$

for all θ and k, where $\mathrm{E}_{\theta, k}$ denotes the expectation with respect to the distribution $p(x, \theta, k)$, $\partial_\theta \mathbf{y}$ is the gradient of $\mathbf{y}$ with respect to θ, i.e., the matrix whose elements are $(\partial y_i / \partial \theta^j)$ in the component form, $\det |\cdot|$ denotes the determinant of a matrix, and $\|\mathbf{y}\|^2$ is the squared norm of the vector $\mathbf{y}$, $\|\mathbf{y}\|^2 = \sum (y_i)^2$. We further need that $\int \mathbf{y} p \mathrm{d}\mu$ is differentiable with respect to θ (Godambe 1976). Condition (2.1) is essential, as is shown

below. Condition (2.2) guarantees that $\mathbf{y}$ depends substantially on θ. This excludes trivial functions. Condition (2.3) guarantees the applicability of the law of large numbers and the central limit theorem shown below.

When a function satisfying (2.1)–(2.3) exists, by replacing the expectation in (2.1) by the empirical sum, we have an estimator $\hat{\theta}$ of θ by solving

$$\sum_{i=1}^{n} \mathbf{y}(x_i, \hat{\theta}) = 0, \tag{2.4}$$

where $x_1, \cdots, x_n$ are n independently and identically distributed observations. This is called the estimating equation and such an estimator is called an M-estimator.

The asymptotic behaviour of the M-estimator $\hat{\theta}$ is obtained from the expansion

$$0 = \sum_{i=1}^{n} \mathbf{y}(x_i, \hat{\theta}) = \sum_{i=1}^{n} \mathbf{y}(x_i, \theta) + \sum_{i=1}^{n} \partial_{\theta} \mathbf{y}(x_i, \theta)(\hat{\theta} - \theta) + O_p(\|\hat{\theta} - \theta\|^2), \tag{2.5}$$

by applying the law of large numbers to $(1/n) \sum \partial_{\theta} \mathbf{y}(x_i, \theta)$ and the central limit theorem to $(1/\sqrt{n}) \sum \mathbf{y}(x_i, \theta)$,

$$\frac{1}{n} \sum \partial_{\theta} \mathbf{y}(x_i, \theta) \approx A, \tag{2.6}$$

$$\frac{1}{\sqrt{n}} \sum \mathbf{y}(x_i, \theta) \sim \varepsilon,$$

where $A = \mathrm{E}_{\theta, k}[\partial_{\theta} \mathbf{y}(x, \theta)]$, ε is a normal random variable subject to $N(0, V)$ with

$$V = \mathrm{E}_{\theta, k}[\mathbf{y}\mathbf{y}^{\mathrm{T}}], \qquad (V_{ij} = \mathrm{E}_{\theta, k}[y_i y_j]), \tag{2.8}$$

$\mathbf{y}$ is a column vector and $\mathbf{y}^{\mathrm{T}}$ is its transposition, and $\approx$ and $\sim$ denote convergence in probability and in distribution, respectively. From this, by neglecting higher-order terms, we have

$$\hat{\theta} - \theta \sim A^{-1} \varepsilon. \tag{2.9}$$

Condition (2.2) guarantees the existence of A^{-1} and (2.3) guarantees the existence of V. The following proposition holds.

Proposition 1. *Under the ordinary regularity conditions, the estimator* $\hat{\theta}$ *obtained from an estimating function* $\mathbf{y}(x, \theta)$ *is consistent and asymptotically normally distributed, with the asymptotic covariance matrix*

$$\mathrm{AV}[\hat{\theta}; \mathbf{y}] = A^{-1} \mathrm{E}_{\theta, k}[\mathbf{y}\mathbf{y}^{\mathrm{T}}] (A^{\mathrm{T}})^{-1}, \tag{2.10}$$

where the asymptotic covariance matrix is defined by

$$\mathrm{AV}[\hat{\theta}; \mathbf{y}] = \lim_{n \to \infty} n \mathrm{E}_{\theta, k}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^{\mathrm{T}}]. \tag{2.11}$$

Let $T(\theta)$ be a non-singular $m \times m$ matrix smoothly depending on θ. It should be noted that $\mathbf{y}^*(x, \theta) = T(\theta)\mathbf{y}(x, \theta)$ gives an estimating function, equivalent to $\mathbf{y}$ in the sense of yielding the same estimates.

We give some examples of semi-parametric statistical models for later use. See Bickel *et al.* (1993) and Groeneboom and Wellner (1992) for many interesting models.

Example 1. *Linear dependence*. We consider the following simple but interesting example of linear regression known as one of the Neyman–Scott problems (Neyman and Scott 1948). This example is analysed in Section 7. Let $x = (\alpha, \beta)$ be composed of two components α and β. It is supposed that the two variables satisfy the linear relation

$$\beta = \theta\alpha$$

in the ideal case, but their observations are contaminated by normal noises. Let $\boldsymbol{\xi}^* = (\xi_1, \xi_2, \ldots)$ be an unknown infinite sequence, and we assume that the ith observation $x_i = (\alpha_i, \beta_i)$ is a noisy version of $(\xi_i, \theta\xi_i)$ given by

$$\begin{aligned} \alpha_i &= \xi_i + n_i, \\ \beta_i &= \theta\xi_i + n_i' \end{aligned} \tag{2.12}$$

where n_i and n_i' $(i = 1, 2, \ldots)$ are independent and are subject to the normal distribution $N(0, \sigma^2)$.

There are many estimators of θ. A very simple one is

$$\hat{\theta}' = \frac{\sum \alpha_i \beta_i}{\sum \alpha_i^2},$$

which minimizes the regression error along the β-axis $\sum(\beta_i - \theta\alpha_i)^2$. However, this $\hat{\theta}'$ is not good, and is not even a consistent estimator. Another estimator is

$$\hat{\theta}_\infty = \frac{\sum \beta_i}{\sum \alpha_i},$$

which is the solution of the estimating function

$$y_\infty(x, \theta) = \beta - \theta\alpha.$$

The estimator given by the estimating function

$$y_0(x, \theta) = (\beta - \theta\alpha)(\theta\beta + \alpha)$$

is the θ component of the joint maximum likelihood estimator of all the parameters $(\theta, \xi_1, \xi_2, \cdots, \xi_n)$ and is known to be equal to the least-squares estimator with respect to the distance of the observed data to the regression line in the orthogonal direction. Which of the two latter consistent estimators is better? There is no definite answer because it depends on the sequence $\boldsymbol{\xi}^*$.

When $\boldsymbol{\xi}^*$ is considered to consist of independently and identically distributed observations from an unknown distribution $k(\xi)$, the present problem can be written in the semi-parametric form

$$p(x, \theta, k) = \int q(x, \theta, \xi) k(\xi) \mathrm{d}\xi,$$

where

$$q(x, \theta, \xi) = c \exp\left[-\frac{1}{2\sigma^2}\{(\alpha - \xi)^2 + (\beta - \theta\xi)^2\}\right].$$

This problem is popular but is still very interesting. We propose later an interesting class of estimators. It is the class of c-estimators $\hat{\theta}_c$, obtained from the following family of estimating functions

$$y_c(x, \theta) = (\beta - \theta\alpha)(\theta\beta + \alpha + c)$$

where c denotes a real constant. A good estimator is obtained by determining c adaptively from the observed data.

The problem is a special case of general mixture models discussed in the following.

Example 2. *Mixture models.* Let $\{q(x, \theta, \boldsymbol{\xi})\}$ be a regular statistical model, where both the parameter of interest θ and the nuisance parameter $\boldsymbol{\xi}$ are of finite dimensions. Let x_i, $i = 1, 2, \cdots, n$, be n independent observations from $q(x, \theta, \boldsymbol{\xi}_i)$, where θ is common but $\boldsymbol{\xi}_i$ takes a different value at each observation. Moreover, we assume that the unknown $\boldsymbol{\xi}_i$ are independently generated subject to a common but unknown probability distribution having a density function $k(\boldsymbol{\xi})$. Then, the x_i are regarded as independent observations from the semi-parametric model

$$p(x, \theta, k) = \int q(x, \theta, \boldsymbol{\xi}) k(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}, \tag{2.13}$$

where $k(\boldsymbol{\xi})$ is the infinite-dimensional nuisance parameter. This model is called the mixture model. This type of problem was studied by Neyman and Scott (1948) and has attracted many researchers (among them Andersen 1970; Lindsay 1982; Kumon and Amari 1984; Amari and Kumon 1988; Pfanzagl 1990). There are a lot of interesting and important examples in this class. Most researchers have treated the distributions of the following exponential form as examples,

$$q(x, \theta, \boldsymbol{\xi}) = \exp\{\boldsymbol{\xi} \cdot \mathbf{s}(x, \theta) + r(x, \theta) - \psi(\theta, \boldsymbol{\xi})\}, \tag{2.14}$$

where $\mathbf{s}(x, \theta)$ is a vector not depending on $\boldsymbol{\xi}$ and $\cdot$ is the inner product. Here, the distribution is of exponential type for $\boldsymbol{\xi}$ when θ is fixed. Models of this type have the m-flat nuisance structure (Amari 1987a) or the convex structure (Bickel *et al.* 1993), to be explained later, so that they possess nice properties. The linear dependence model (Example 1) is a special case of this type.

Example 3. *Semi-parametric additive regression.* Let $(\mathbf{x}, t)$ be covariates to which y is connected by

$$y = \theta \cdot \mathbf{x} + k(t) + \epsilon, \tag{2.15}$$

where θ is the parameter of interest, $k(t)$ is an unknown nuisance smooth function of t, and ϵ is a noise term. We assume here that ϵ is subject to $N(0, 1)$. The problem is to estimate θ based on n observations $(y_i, \mathbf{x}_i, t_i)$, $i = 1, 2, \cdots, n$, where we assume that $(\mathbf{x}, t)$ is subject to a

known joint distribution $q(\mathbf{x}, t)$ (see Cuzick 1992). This model is analysed later to show that our theory is applicable also to non-*m*-flat models.

3. Hilbert fibre space and score function

Given a probability density function $p(x)$, its small deviation in the direction of $a(x)$ can be represented by a curve $p(x, t)$ starting from $p(x)$,

$$p(x, t) = p(x)\{1 + ta(x)\}, \tag{3.1}$$

where $t(0 \leq t < \epsilon)$ is the parameter of the curve. Here

$$\mathrm{E}[a(x)] = 0$$

holds, where E is the expectation with respect to $p(x)$, because of

$$\int p(x)\{1 + ta(x)\}\mathrm{d}\mu(x) = 1. \tag{3.2}$$

In order to be specific, we consider the linear space of functions which satisfy

$$\mathrm{E}[a(x)] = 0, \qquad \mathrm{E}[\{a(x)\}^2] < \infty. \tag{3.3}$$

This set is a Hilbert space H_p, often denoted by $L_2^0(p)$, with the inner product of $a(x)$ and $b(x)$ defined by

$$\langle a(x), b(x)\rangle = \mathrm{E}[a(x)b(x)]. \tag{3.4}$$

We call the random variable

$$a(x) = \frac{\mathrm{d}}{\mathrm{d}t}\log p(x, t)\bigg|_{t=0} \tag{3.5}$$

the tangent vector of the curve (3.1). This is the score function for the one-dimensional statistical model (3.1) parametrized by t. Refer to Pfanzagl (1990), van der Vaart (1991), Groeneboom and Wellner (1992) and Bickel *et al.* (1993) for mathematical details on the rigorous construction of the tangent space.

Given a semi-parametric model $S = \{p(x, \boldsymbol{\theta}, k)\}$, the Hilbert space $H_p = H_{\boldsymbol{\theta}, k}$ is associated with each point $(\boldsymbol{\theta}, k)$, that is, with each distribution $p(x) = p(x, \boldsymbol{\theta}, k)$ specified by $(\boldsymbol{\theta}, k)$. A collection of such $H_{\boldsymbol{\theta}, k}$ is called a *fibred structure*, where the fibres are the Hilbert spaces.

We first define the tangent directions along the parameter of interest. Let

$$u_i(x, \boldsymbol{\theta}, k) = \frac{\partial}{\partial\theta^i}\log p(x, \boldsymbol{\theta}, k) \tag{3.6}$$

be the score function with respect to the ith component θ^i of $\boldsymbol{\theta}$. Obviously,

$$\mathrm{E}_{\boldsymbol{\theta}, k}[u_i] = 0 \tag{3.7}$$

and we further assume that u_i is square-integrable. Then it belongs to $H_{\theta,k}$. We call the subspace spanned by these u_i the tangent subspace $T^{\mathrm{I}}_{\theta,k}$ along the parameter of interest. The vector score function is $\mathbf{u} = (u_1, \cdots, u_m)$.

We next define the tangent directions along the nuisance parameter. Let us assume that, for any $\tilde{k}$ in a small neighbourhood of k in the set K of the nuisance parameter, there exists a curve $c(t)$ connecting them, such that $c(0) = k$ and $c(\epsilon) = \tilde{k}$, and that the score function for the one-dimensional statistical model $p\{x, \theta, c(t)\}$ parametrized by t,

$$v(x, \theta, k, c) = \frac{\mathrm{d}}{\mathrm{d}t} \log p\{x, \theta, c(t)\}\Big|_{t=0} \tag{3.8}$$

belongs to $H_{\theta,k}$. This v is the tangent vector along $c(t)$ of the nuisance parameter. Let $T^{\mathrm{N}}_{\theta,k}$ be the smallest closed subspace including all such vs. We call it the nuisance tangent space.

Now, let us project the score function u_i onto the subspace $(T^{\mathrm{N}}_{\theta,k})^{\perp}$ which is the orthogonal complement of $T^{\mathrm{N}}_{\theta,k}$. The result is the function $u_i^{\mathrm{E}} = u_i - v$ that minimizes $\mathrm{E}[|u_i - v|^2]$, $v \in T^{\mathrm{N}}_{\theta,k}$. The vector function $\mathbf{u}^{\mathrm{E}} = (u_i^{\mathrm{E}})$ is called the efficient score function and the u_i^{E} are called the components of the efficient or projected score functions (see Begun *et al.* 1983; Amari and Kumon 1988; Small and McLeish 1989). Let $T^{\mathrm{E}}_{\theta,k}$ be the subspace of $H_{\theta,k}$ spanned by the components u_i^{E} of the efficient score function.

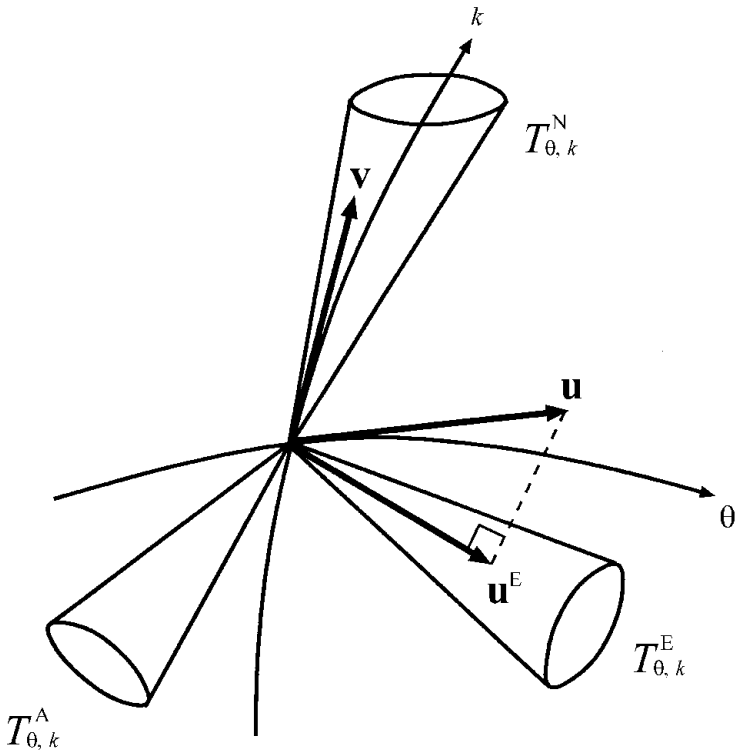

Figure 1. Orthogonal decomposition of $H_{\theta,k}$

Let $T^{\mathrm{A}}_{\theta,k}$ be the orthogonal complement of $T^{\mathrm{N}}_{\theta,k} \oplus T^{\mathrm{E}}_{\theta,k}$. This is called the ancillary subspace and spans directions orthogonal to any changes in the parameter of interest and the nuisance parameter. We thus have the orthogonal decomposition of the Hilbert fibre space (Fig. 1; see Amari 1987a; Amari and Kumon 1988; see also Small and McLeish 1988),

$$H_{\theta,k} = T^{\mathrm{E}}_{\theta,k} \oplus T^{\mathrm{A}}_{\theta,k} \oplus T^{\mathrm{N}}_{\theta,k}. \tag{3.9}$$

The matrix $G^{\mathrm{E}} = (g^{\mathrm{E}}_{ij})$ defined by using the efficient score function

$$g^{\mathrm{E}}_{ij}(\theta,k) = \mathrm{E}_{\theta,k}[u^{\mathrm{E}}_i u^{\mathrm{E}}_j] \tag{3.10}$$

is called the efficient Fisher information matrix. Begun *et al.* (1983) proved that G^{E} gives the Cramér–Rao bound of the asymptotic covariance of estimators $\hat{\theta}$,

$$\lim_{n\to\infty} n\mathrm{E}[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^{\mathrm{T}}] \geq (G^{\mathrm{E}})^{-1} \tag{3.11}$$

for any asymptotically normally distributed unbiased estimator in a semi-parametric model. Ritov and Bickel (1990) showed that the bounds may be unattainable in some cases. However, the above bounds are attainable in many important cases under mild regularity conditions (Bickel *et al.* 1993).

4. Invariant decomposition of Hilbert fibres due to dual parallel transports

An estimating function $\mathbf{y}(x,\theta)$ satisfies the unbiasedness condition (2.1) for all k. Such a global structure is elucidated by introducing two parallel transports of the Hilbert fibres along the nuisance space.

Let $a(x)$ be a random variable belonging to $H_{\theta,k}$. Let us fix θ, and consider the subset $S_\theta = \{p(x,\theta,k)|k \in K\}$. We define two parallel transports of a vector $a(x)$ from $H_{\theta,k}$ to $H_{\theta,k'}$ (Amari 1987a). Then

$$\prod^{(e)}{}^{k'}_{k} a(x) = a(x) - \mathrm{E}_{\theta,k'}[a(x)], \tag{4.1}$$

$$\prod^{(m)}{}^{k'}_{k} a(x) = \frac{p(x,\theta,k)}{p(x,\theta,k')} a(x) \tag{4.2}$$

are called the e-parallel transport and the m-parallel transport of $a(x)$ from (θ,k) to (θ,k'), respectively. It should be noted that the e-parallel transport exists only when the expectation of $a(x)$ at (θ,k') exists. It is easy to show that

$$\mathrm{E}_{\theta,k'}\left[\prod^{(m)}{}^{k'}_{k} a(x)\right] = 0$$

always holds, and that

$$\mathrm{E}_{\theta,k'}\left[\prod_{k}^{(e)k'} a(x)\right] = 0$$

holds when $\mathrm{E}_{\theta,k'}[a(x)]$ exists. However, the e- and/or m-parallel transports of $a(x)$ do not necessarily belong to $H_{\theta,k'}$. They belong to $H_{\theta,k'}$ only when they are square-integrable at (θ,k') with respect to $p(x,\theta,k')$.

The parallel transports are generalizations of the dual geometrical structures derived from the underlying e- and m-connections or e- and m-covariant derivatives (Amari 1985; see also Amari and Kumon 1988), but we will not go into mathematical details of differential geometry.

The following lemma shows the most important property connecting the two parallel transports. The proof is immediate and hence omitted.

Lemma 1. *The two parallel transports are dual in the sense that, for any two* $a(x), b(x) \in H_{\theta,k}$, *the inner product*

$$\langle a,b\rangle_{\theta,k} = \left\langle \prod_{k}^{(e)k'} a, \prod_{k}^{(m)k'} b \right\rangle_{\theta,k'}, \tag{4.3}$$

where the suffix (θ,k) *denotes that the expectation is taken with respect to* $p(x,\theta,k)$, *is kept invariant when their parallel transports belong to* $H_{\theta,k'}$.

It is remarked that an estimating function is e-invariant,

$$\prod_{k}^{(e)k'} \mathbf{y}(x,\theta) = \mathbf{y}(x,\theta)$$

because of (2.1) where $\prod^{(e)}$ operates componentwise. Let us consider a curve $k = k(t)$, $k_0 = k(0)$, in the nuisance space. By differentiating (2.1) with respect to t along the curve $k = k(t)$ and exchanging the integral and differentiation, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\int p\{x,\theta,k(t)\}\mathbf{y}(x,\theta)\mathrm{d}\mu(x)\Big|_{t=0} \\
&\quad = \int v\{x,\theta,k_0\}p\{x,\theta,k_0\}\mathbf{y}(x,\theta)\mathrm{d}\mu(x) \\
&\quad = \langle v,\mathbf{y}(x,\theta)\rangle_{\theta,k_0} = 0,
\end{aligned}$$

where

$$v = \frac{\mathrm{d}}{\mathrm{d}t}\log p\{x,\theta,k(t)\}\Big|_{t=0}$$

is the nuisance tangent direction at k_0. This holds for any k_0 so that any estimating function $\mathbf{y}(x, \theta)$ is orthogonal to v at any point (θ, k). However, from

$$\begin{aligned} \langle v, \mathbf{y}(x, \theta) \rangle_{\theta, k'} &= \left\langle \prod\nolimits_{k'}^{k(m)} v, \prod\nolimits_{k'}^{k(e)} \mathbf{y} \right\rangle_{\theta, k} \\ &= \left\langle \prod\nolimits_{k'}^{k(m)} v, \mathbf{y} \right\rangle_{\theta, k}, \end{aligned} \tag{4.4}$$

where v is a nuisance tangent direction at k', the orthogonality condition at k' is transferred to that at k by the m-parallel transports of v at k'. This shows that an estimating function $\mathbf{y}$ is orthogonal, not only to the nuisance tangent direction at any k, but to the m-parallel transports from k' to k of the nuisance tangent directions at any k'.

Motivated by the above discussion, we now reorganize the decomposition (3.9) of $H_{\theta, k}$ by taking account of the global structure induced by the parallel transports. The information fibre space $F^{\mathrm{I}}_{\theta, k}$ at (θ, k) is constructed from $T^{\mathrm{E}}_{\theta, k}$ such that its elements are orthogonal not only to the nuisance tangent space $T^{\mathrm{N}}_{\theta, k}$ at (θ, k) but also to the m-parallel transports from k' to k of $T^{\mathrm{N}}_{\theta, k'}$ at points (θ, k') for all $k' \in K$. To this end, we first consider a vector

$$r(x) \in T^{\mathrm{E}}_{\theta, k} \oplus T^{\mathrm{A}}_{\theta, k}$$

whose e-transport exists in the Hilbert space $H_{\theta, k'}$ for every k', that is,

$$\mathrm{E}_{\theta, k'}[\{r(x)\}^2] < \infty. \tag{4.5}$$

Suppose its e-transport to (θ, k') is orthogonal to $T^{\mathrm{N}}_{\theta, k'}$ for every $k' \in K$, that is,

$$\left\langle v, \prod\nolimits_{k}^{k'(e)} r(x) \right\rangle_{\theta, k'} = \left\langle \prod\nolimits_{k'}^{k(m)} v, r(x) \right\rangle_{\theta, k} = 0, \qquad v \in T^{\mathrm{N}}_{\theta, k'}. \tag{4.6}$$

We express this by saying that $r(x)$ is free of any nuisance tangent directions at every k' when it is e-transported. In this case $r(x)$ is a candidate for an estimating function. This implies that $r(x)$ is orthogonal to not only $T^{\mathrm{N}}_{\theta, k}$ but also all the m-transports from k' to k of $T^{\mathrm{N}}_{\theta, k'}$. This suggests that we enlarge the nuisance space $T^{\mathrm{N}}_{\theta, k}$ to

$$\mathrm{span}\left\{ \bigcup_{k'} \prod\nolimits_{k'}^{k(m)} T^{\mathrm{N}}_{\theta, k'} \right\},$$

and define the directions orthogonal to it. Any estimating functions should be orthogonal to the enlarged subspace.

To define such a space formally, we consider the closed subspace of $H_{\theta, k}$ consisting of the vectors satisfying the above conditions (4.5) and (4.6) and refer to it tentatively by $F^{\mathrm{IA}}_{\theta, k}$,

where I denotes the information part and A denotes the ancillary part. Obviously, the following hold:

$$F^{\mathrm{IA}}_{\theta,k} \subset T^{\mathrm{E}}_{\theta,k} \oplus T^{\mathrm{A}}_{\theta,k}.$$

$$F^{\mathrm{IA}}_{\theta,k} \perp \bigcup_{k'} \overset{(m)}{\prod}{}^{k}_{k'}\, T^{\mathrm{N}}_{\theta,k'}.$$

The nuisance fibre space $F^{\mathrm{N}}_{\theta,k}$ is defined as the orthogonal complement of $F^{\mathrm{IA}}_{\theta,k}$ in $H_{\theta,k}$. Obviously

$$F^{\mathrm{N}}_{\theta,k} \supset \overset{(m)}{\prod}{}^{k}_{k'}\, T^{\mathrm{N}}_{\theta,k'}. \tag{4.7}$$

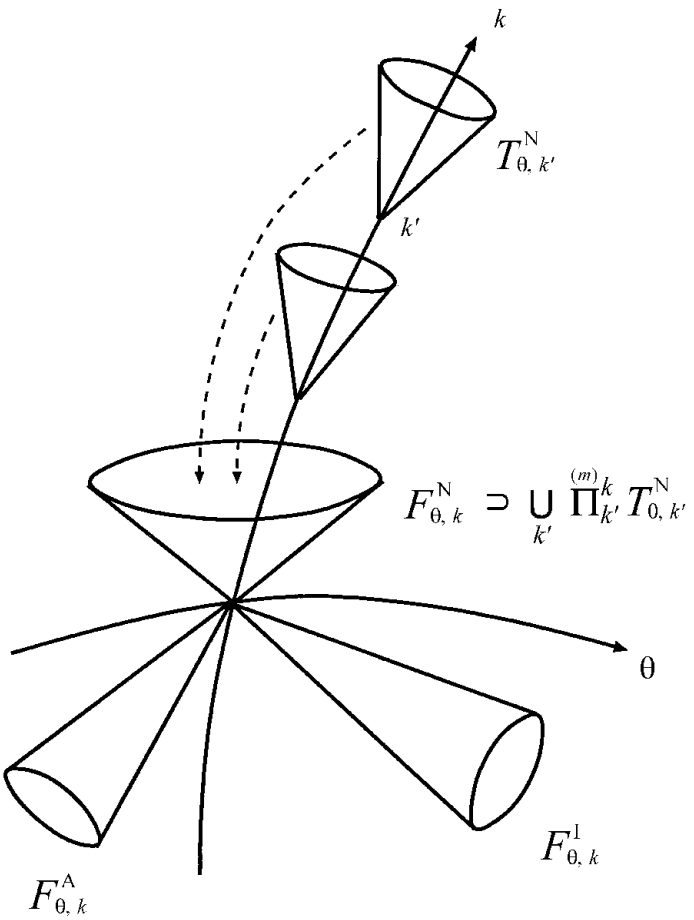

Figure 2. Alternative orthogonal decomposition of $H_{\theta,k}$ based on m-parallel transports

We have

$$H_{\theta,k} = F^{\mathrm{IA}}_{\theta,k} \oplus F^{\mathrm{N}}_{\theta,k}. \tag{4.8}$$

We now decompose $F^{\mathrm{IA}}_{\theta,k}$. Let u^{I}_i be the projection of the score u_i onto $F^{\mathrm{IA}}_{\theta,k}$. This is called the information score. The information fibre space denoted by $F^{\mathrm{I}}_{\theta,k}$ is the subspace spanned by the information score functions $u^{\mathrm{I}}_i(x,\theta,k)$. The ancillary fibre space is the orthogonal complement of $F^{\mathrm{I}}_{\theta,k}$ in $F^{\mathrm{IA}}_{\theta,k}$ and is denoted by $F^{\mathrm{A}}_{\theta,k}$,

$$F^{\mathrm{IA}}_{\theta,k} = F^{\mathrm{I}}_{\theta,k} \oplus F^{\mathrm{A}}_{\theta,k}.$$

We thus have another orthogonal decomposition of $H_{\theta,k}$,

$$H_{\theta,k} = F^{\mathrm{I}}_{\theta,k} \oplus F^{\mathrm{A}}_{\theta,k} \oplus F^{\mathrm{N}}_{\theta,k}, \tag{4.9}$$

which represents a more global structure of $H_{\theta,k}$ (Fig. 2). The subspace $F^{\mathrm{N}}_{\theta,k}$ includes all the m-parallel transports of $T^{\mathrm{N}}_{\theta,k'}$ from k' to k because of the relation (4.6). It is the information fibre $F^{\mathrm{I}}_{\theta,k}$ that plays an important role.

It is helpful to compare the two decompositions (3.9) and (4.9). $F^{\mathrm{N}}_{\theta,k}$ is enlarged from $T^{\mathrm{N}}_{\theta,k}$ to include all the m-parallel transports of $T^{\mathrm{N}}_{\theta,k'}$. The $F^{\mathrm{I}}_{\theta,k}$ is constructed from the projections of the θ-score functions to the orthogonal complement of the enlarged $F^{\mathrm{N}}_{\theta,k}$, while $T^{\mathrm{E}}_{\theta,k}$ is constructed from the projections to the orthogonal complement of smaller $T^{\mathrm{N}}_{\theta,k}$. Therefore, when $T^{\mathrm{N}}_{\theta,k} = F^{\mathrm{N}}_{\theta,k}$ holds, $F^{\mathrm{I}}_{\theta,k} = T^{\mathrm{E}}_{\theta,k}$ holds. This is an important special case.

5. Estimating functions and their efficiency

Based on the decomposition (4.9) of the Hilbert space $H_{\theta,k}$, we can now characterize the set of all the estimating functions. We first answer the two important questions when an estimating function exists and what is the set of all the estimating functions. To this end, we prove the following two important lemmas.

Lemma 2. *Any component of an estimating function* $\mathbf{y}(x,\theta)$ *belongs to* $F^{\mathrm{I}}_{\theta,k} \oplus F^{\mathrm{A}}_{\theta,k}$ *for any* k. *Let* $y^{\mathrm{I}}_i(x,\theta,k)$ *be the projection of the* i*th component of* $\mathbf{y}(x,\theta)$ *onto* $F^{I}_{\theta,k}$. *Then, the* $y^{I}_i(x,\theta,k), i = 1,\cdots,m$, *span* $F^{\mathrm{I}}_{\theta,k}$.

Proof. Let $\mathbf{y}(x,\theta)$ be an estimating function. Then, its e-transport always exists in the corresponding Hilbert space because of (2.3), and it is e-invariant:

$$\prod^{(e)} \mathbf{y}(x,\theta) = \mathbf{y}(x,\theta). \tag{5.1}$$

We have already shown in (4.4) that

$$\left\langle \prod^{(m)}{}^{k}_{k'} v(x,\theta,k'),\ \mathbf{y}(x,\theta) \right\rangle_{\theta,k} = 0 \tag{5.2}$$

for all k and k'. Therefore, $\mathbf{y}$ is included in $F^{\mathrm{I}}_{\theta,k} \oplus F^{\mathrm{A}}_{\theta,k}$. Moreover, by differentiating (2.1) with respect to θ, we have

$$\mathrm{E}_{\theta,k}[\partial_\theta \mathbf{y}(x,\theta)] + \langle \mathbf{u}, \mathbf{y}(x,\theta)\rangle = 0, \tag{5.3}$$

where $\langle \mathbf{u}, \mathbf{y}\rangle$ denotes a matrix whose elements are $\langle u_i, y_j\rangle$. This shows that

$$\mathrm{E}_{\theta,k}[\partial_\theta \mathbf{y}] = -\langle \mathbf{u}, \mathbf{y}\rangle.$$

Since $\mathbf{y}$ belongs to $F^{\mathrm{I}}_{\theta,k} \oplus F^{\mathrm{A}}_{\theta,k}$ and the projection of $\mathbf{u}$ onto the space $F^{\mathrm{I}}_{\theta,k} \oplus F^{\mathrm{A}}_{\theta,k}$ includes no $F^{\mathrm{A}}_{\theta,k}$ part, we have

$$\langle \mathbf{u}, \mathbf{y}\rangle = \langle \mathbf{u}^{\mathrm{I}}, \mathbf{y}\rangle,$$

where $\mathbf{u}^{\mathrm{I}}$ is the projection of $\mathbf{u}$ onto $F^{\mathrm{I}}_{\theta,k}$. Therefore, (2.2) implies that the determinant of $\langle u^{\mathrm{I}}_i, y_j\rangle$ does not vanish and that the projections of vectors y_i onto $F^{\mathrm{I}}_{\theta,k}$ span $F^{\mathrm{I}}_{\theta,k}$. This also shows that $F^{\mathrm{I}}_{\theta,k}$ is non-degenerate, that is, its dimension is the same as that of the parameter θ of interest. □

We now prove what is in essence a converse of Lemma 2.

Lemma 3. *Any vector* $\mathbf{w}(x,\theta)$ *belonging to* $F^{\mathrm{I}}_{\theta,k} \oplus F^{\mathrm{A}}_{\theta,k}$ *for some* $k \in K$ *is an estimating function provided* $F^{\mathrm{I}}_{\theta,k'}$ *is non-degenerate for every* k' *and the projections of the components* w_i *of* $\mathbf{w}$ *onto* $F^{\mathrm{I}}_{\theta,k'}$ *span* $F^{\mathrm{I}}_{\theta,k'}$ *for every* k'.

Remark. The first part of the proof, for which the conditions of the lemma are not needed, will show that the space $F^{\mathrm{I}}_{\theta,k} \oplus F^{\mathrm{A}}_{\theta,k}$ is e-invariant.

Proof. Let $c(t)$ be a curve connecting two points k and k', $k = c(0)$, and put

$$\mathbf{f}(t) = \mathrm{E}_{\theta,c(t)}[\mathbf{w}(\mathbf{x},\theta)].$$

Since the e-parallel transport of w from k to $c(t)$ is written as

$$\prod_{k}^{c(t)(e)} \mathbf{w} = \mathbf{w} - \mathbf{f}(t),$$

$\mathbf{w}$ is e-invariant if $\mathbf{f}(t) = 0$ holds. Hence, we need to prove $\mathbf{f}(t) = 0$. Obviously $\mathbf{f}(0) = 0$. By differentiation, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{f}(t) &= \int \frac{\mathrm{d}}{\mathrm{d}t} p\{x,\theta,c(t)\}\mathbf{w}(x,\theta)\mathrm{d}\mu(x) \\ &= \mathrm{E}_{\theta,c(t)}[v(t)\mathbf{w}(x,\theta)] = \langle v, \mathbf{w}\rangle_{\theta,c(t)},\end{aligned}$$

where

$$v(t) = \frac{\mathrm{d}}{\mathrm{d}t}\log p\{x,\theta,c(t)\}.$$

Since $\mathbf{w} \in F^{\mathrm{IA}}_{\theta,k}$, its e-parallel transport is orthogonal to $v(t)$, and hence we have

$$0 = \left\langle v(t), \prod_{k}^{(c)\,c(t)} \mathbf{w} \right\rangle_{\theta, c(t)} = \langle v(t), \mathbf{w} \rangle_{\theta, c(t)} - \mathbf{f}(t) \mathrm{E}_{\theta, c(t)}[v(t)],$$

$$= \langle v(t), \mathbf{w} \rangle_{\theta, c(t)}.$$

Hence $\mathrm{d}\mathbf{f}(t)/\mathrm{d}t = 0$, so that $\mathbf{f}(t) = 0$. This proves the e-invariance of $\mathbf{w}$,

$$\prod_{k}^{(e)\,k'} \mathbf{w} = \mathbf{w}$$

and

$$\mathrm{E}_{\theta, k'}[\mathbf{w}(x, \theta)] = 0$$

for all k'. Since the projections $\mathrm{w}_i^{\mathrm{I}}$ of the components w_i of $\mathbf{w}$ onto $F^{\mathrm{I}}_{\theta,k'}$ do, by assumption, span $F^{\mathrm{I}}_{\theta,k'}$ for any k', then $\mathrm{E}_{\theta,k'}[\partial_\theta \mathbf{w}(x,\theta)] = -\langle \mathbf{u}^{\mathrm{I}}, \mathbf{w} \rangle_{\theta,k'} = -\langle \mathbf{u}^{\mathrm{I}}, \mathbf{w}^{\mathrm{I}} \rangle_{\theta,k'}$ is of full rank. Therefore, $\mathbf{w}$ is an estimating function satisfying (2.1)–(2.3). □

By combining Lemmas 2 and 3, we have the following proposition.

Proposition 2. *Any estimating function* $\mathbf{y}(x, \theta) = \{y_i(x, \theta)\}$ *can be decomposed for any k as a sum*

$$\mathbf{y}(x, \theta) = T(\theta, k)\mathbf{u}^{\mathrm{I}}(x, \theta, k) + \mathbf{a}(x, \theta, k), \tag{5.4}$$

where the component $a_i(x, \theta, k)$ *of* $\mathbf{a}$ *belongs to* $F^{\mathrm{A}}_{\theta,k}$ *and* $T(\theta, k)$ *is a non-singular matrix. Conversely, any function* $\mathbf{y}(x, \theta)$ *defined in the form of* (5.4) *at a fixed* k_0 *gives an estimating function provided the projections of the components* $y_i(x, \theta)$ *onto* $F^{\mathrm{I}}_{\theta,k'}$ *span* $F^{\mathrm{I}}_{\theta,k'}$ *for every* k'.

It is possible to choose a basis for the efficient scores such that $T(\theta, k)$ becomes the identity at some k. The proposition also gives a simple sufficient condition for the existence of an estimating function. This is shown by putting $\mathbf{a}(x, \theta, k) = 0$ in Proposition 2. Lemma 2 gives a necessary condition. We summarize these.

Proposition 3. *A necessary condition for the existence of an estimating function is that* $F^{\mathrm{I}}_{\theta,k}$ *is non-degenerate, that is, m-dimensional. A sufficient condition is that, for a fixed* k_0, *the projections of* $u_i^{\mathrm{I}}(x, \theta, k_0)$ $(i = 1, \cdots, m)$ *onto* $F^{\mathrm{I}}_{\theta,k'}$ *span* $F^{\mathrm{I}}_{\theta,k'}$ *for every* k'. *The vector* $\mathbf{u}^{\mathrm{I}}(x, \theta, k_0)$ *is an estimating function in this case.*

Remark. We discuss informally a necessary and sufficient condition. It is clear that any estimating function can be written in the form

$$y_i(x, \theta) = \int \sum_{i=1}^{m} \varphi_{ij}(k) u_j^{\mathrm{I}}(x, \theta, k) \mathrm{d}\mu(k) + a_i(x, \theta),$$

where $a_i(x, \theta)$ is purely ancillary, that is orthogonal to $F^{\mathrm{I}}_{\theta,k}$ at any k. Therefore, a necessary and sufficient condition for the existence of estimating functions is the existence of

functionals $\varphi_{ij}(k)$ such that the projections $y_i^{\mathrm{I}}(x,\theta,k)$ of the above y_i onto $F^{\mathrm{I}}_{\theta,k}$ $(i=1,\ldots,m)$ span $y_i^{\mathrm{I}}(x,\theta,k)$ for all k. Let us define

$$\rho_{ij}(k,k') = \langle u_i^{\mathrm{I}}(x,\theta,k), u_j^{\mathrm{I}}(x,\theta,k')\rangle_{\theta,k'}.$$

Then, the condition is equivalent to the existence of $\varphi_{ij}(k)$ such that the matrix

$$B_{ij} = \sum_{n=1}^{m} \int \varphi_{in}(k)\rho_{nj}(k,k')\mathrm{d}\mu(k)$$

is non-degenerate for all k'. The sufficient condition of Proposition 3 guarantees the existence of such φ_{ij}. When $u_j^{\mathrm{I}}(x,\theta,k')$ $(i=1,\ldots,m)$ are non-degenerate, because of continuity, there exists a neighbourhood $N_K(k)$ of k such that the projections of $u_i^{\mathrm{I}}(x,\theta,k)$, $(i=1,\ldots,m)$ onto $F^{\mathrm{I}}_{\theta,k'}$ span $F^{\mathrm{I}}_{\theta,k'}$ for all elements k' of $N_K(k)$. If a function $\mathbf{y}(x,\theta)$ satisfies the conditions of estimating functions where the global condition (2.2) is replaced by the local one,

$$\det|\mathrm{E}_{\theta,k'}[\partial_\theta \mathbf{y}(x,\theta)]| \neq 0,$$

at least in a neighbourhood of a point k, such a function is called a local estimating function. The non-degeneracy of $\mathbf{u}^{\mathrm{I}}(x,\theta,k)$ is a necessary and sufficient condition of the existence of local estimating functions at k.

We next calculate the asymptotic covariance matrix of an estimating function. This calculation leads us to the optimal estimating function. As discussed above, an estimating function $\mathbf{y}(x,\theta)$ may be decomposed as

$$\mathbf{y}(x,\theta) = \mathbf{u}^{\mathrm{I}}(x,\theta,k) + \mathbf{a}(x,\theta,k),$$

where $\mathbf{u}^{\mathrm{I}} = (u_i^{\mathrm{I}}) \in F^{\mathrm{I}}_{\theta,k}$ and $\mathbf{a} = (a_i) \in F^{\mathrm{A}}_{\theta,k}$. It is easy to show that

$$\mathrm{E}[\partial_\theta \mathbf{a}] = -\langle \mathbf{u}, \mathbf{a}\rangle = 0,$$

by differentiating $\mathrm{E}_{\theta,k}[\mathbf{a}(x,\theta,k)] = 0$. Therefore, we have

$$-\mathrm{E}[\partial_\theta \mathbf{y}] = -\mathrm{E}[\partial_\theta \mathbf{u}^{\mathrm{I}}] = \langle \mathbf{u}, \mathbf{u}^{\mathrm{I}}\rangle = \langle \mathbf{u}^{\mathrm{I}}, \mathbf{u}^{\mathrm{I}}\rangle, \tag{5.5}$$

$$\mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}}] = \mathrm{E}[\mathbf{u}^{\mathrm{I}}(\mathbf{u}^{\mathrm{I}})^{\mathrm{T}}] + \mathrm{E}[\mathbf{a}\mathbf{a}^{\mathrm{T}}] = G^{\mathrm{I}} + G^{\mathrm{A}}, \tag{5.6}$$

where we put

$$G^{\mathrm{I}} = \mathrm{E}[\mathbf{u}^{\mathrm{I}}(\mathbf{u}^{\mathrm{I}})^{\mathrm{T}}], \qquad G^{\mathrm{A}} = \mathrm{E}[\mathbf{a}\mathbf{a}^{\mathrm{T}}]. \tag{5.7}$$

So, by Proposition 1, we have the following result.

Proposition 4. *The asymptotic covariance matrix derived from an estimating function* $\mathbf{y}(x,\theta)$ *is given by*

$$\mathrm{AV}[\hat{\theta};\mathbf{y}] = (G^{\mathrm{I}})^{-1} + (G^{\mathrm{I}})^{-1}G^{\mathrm{A}}(G^{\mathrm{I}})^{-1}. \tag{5.8}$$

The estimating function given by

$$\mathbf{y}(x,\theta) = \mathbf{u}^{\mathrm{I}}(x,\theta,k_0) \tag{5.9}$$

where k_0 is a fixed point in K, is optimal among all the M-estimators when the true distribution happens to be specified by (θ, k_0) and $\mathbf{u}_i^{\mathrm{I}}$ span $F_{\theta,k}^{\mathrm{I}}$ for every k. The optimal asymptotic covariance is given by $(G^{\mathrm{I}})^{-1}$ in this case.

It is shown that the information fibre $F_{\theta,k}^{\mathrm{I}}$ plays the fundamental role, being defined through the m-parallel transports of $T_{\theta,k'}^{\mathrm{N}}$. However, in many important cases, $F_{\theta,k}^{\mathrm{I}}$ is equal to the simpler $T_{\theta,k}^{\mathrm{E}}$, which is easier to obtain. We show when such a simplification occurs.

We fix θ, and consider the statistical submodel $S_\theta = \{p(x, \theta, k)\}$, where $k \in K$ is the only free parameter. The tangent vectors of S_θ compose the nuisance tangent space $T_{\theta,k}^{\mathrm{N}}$. Let us consider the m-parallel transport of $T_{\theta,k'}^{\mathrm{N}}$ from (θ, k') to (θ, k) and see how it is different from $T_{\theta,k}^{\mathrm{N}}$. A manifold in general is said to be flat or curvature-free when its tangent directions are the same at all the points. In the present case, we can compare two tangent spaces $T_{\theta,k}^{\mathrm{N}}$ and $T_{\theta,k'}^{\mathrm{N}}$ by the m-parallel transport of one to the other. We give formal definitions of m-flatness and m-information-curvature-freeness.

Definition 1. *A semi-parametric statistical model S is said to be* m-flat *or* m-convex*, when the $T_{\theta,k'}^{\mathrm{N}}$ are invariant under the m-parallel transports, that is*

$$\prod_{k'}^{k(m)} T_{\theta,k'}^{\mathrm{N}} \subset T_{\theta,k}^{\mathrm{N}}$$

for any k, k' and θ. When the m-parallel transport of $T_{\theta,k'}^{N}$ from (θ, k') to (θ, k) does not include the $T_{\theta,k}^{\mathrm{E}}$ components for any k, k' and θ, that is,

$$\prod_{k'}^{k(m)} T_{\theta,k'}^{\mathrm{N}} \subset T_{\theta,k}^{\mathrm{N}} \oplus T_{\theta,k}^{\mathrm{A}}, \tag{5.10}$$

the model S is said to be m-curvature free in the information directions, or for short, m-information-curvature free.

It is easy to see that, when S is m-flat, it is m-information-curvature free. When S_θ is not m-flat, S_θ is curved in general, because its tangent directions change as k changes. However, when S_θ is m-information-curvature free, the changes in the tangent directions are restricted to the $T_{\theta,k}^{\mathrm{A}}$ (and $T_{\theta,k}^{\mathrm{N}}$) directions, showing that S_θ is not curved in the direction of $T_{\theta,k}^{\mathrm{E}}$. Now, by using m-information-curvature freeness, we show a necessary and sufficient condition that $F_{\theta,k}^{\mathrm{I}}$ coincides with $T_{\theta,k}^{\mathrm{E}}$ at every k.

Lemma 4. *If and only if S_θ is m-information-curvature free, we have, for every k,*

$$F_{\theta,k}^{\mathrm{I}} = T_{\theta,k}^{\mathrm{E}}. \tag{5.11}$$

Proof. By definition, when S_θ is m-information-curvature free

$$\bigcup_{k'} \prod_{k'}^{k(m)} T_{\theta,k'}^{\mathrm{N}} \subset T_{\theta,k}^{\mathrm{N}} \oplus T_{\theta,k}^{\mathrm{A}} \tag{5.12}$$

holds for every k and vice versa. At any k, (5.12) is equivalent to

$$F_{\theta,k}^{\mathrm{IA}} = (F_{\theta,k}^{\mathrm{N}})^{\perp} \supset T_{\theta,k}^{\mathrm{E}}, \tag{5.13}$$

or, in other words,

$$\left\langle \prod_{k'}^{k(m)} v(x,\theta,k'),\ u_i^{\mathrm{E}}(x,\theta,k) \right\rangle_{\theta,k} = 0$$

holds for every k', $v \in T^{\mathrm{N}}_{\theta,k'}$ and for all $i = 1, \cdots, m$, because of (4.6). The information fibre $F^{\mathrm{I}}_{\theta,k}$ is constructed by projecting the components of the θ-score function $\mathbf{u}$ onto $F^{\mathrm{IA}}_{\theta,k}$. It is easy to show that $F^{\mathrm{I}}_{\theta,k}$ is the projection of $T^{\mathrm{E}}_{\theta,k}$ onto $F^{\mathrm{IA}}_{\theta,k}$ because $T^{\mathrm{E}}_{\theta,k} \oplus T^{\mathrm{A}}_{\theta,k} \supset F^{\mathrm{IA}}_{\theta,k}$. Therefore, (5.13) leads to $F^{\mathrm{I}}_{\theta,k} = T^{\mathrm{E}}_{\theta,k}$. The converse is obvious because (5.11) implies (5.13) which is equivalent to (5.12). □

We can now show the amount of loss of information caused by adopting the simple method of estimating functions (M-estimators), relative to the Cramér–Rao-type bound. We also give a necessary and sufficient condition for the best M-estimator to be lossless, that is, efficient in the sense that it attains the Cramér–Rao-type bound. Here, we define the optimal estimating function. Let the true distribution be $p(x, \theta, k_0)$. An estimating function $\mathbf{y}(x, \theta)$ is said to be optimal at a point k_0 when the asymptotic variance $\mathrm{AV}[\hat{\theta}; \mathbf{y}]$ is minimal (in the sense of matrices) among the estimators obtained from all the estimating functions. The optimal estimating function depends on k_0, and there is no guarantee that we can find it because k_0 is unknown.

Proposition 5. *When S_θ is m-information-curvature free, $\mathbf{y}(x, \theta) = \mathbf{u}^{\mathrm{I}}(x, \theta, k_0) = \mathbf{u}^{\mathrm{E}}(x, \theta, k_0)$ is the optimal estimating function at k_0 and is efficient at k_0. When S_θ is not m-information-curvature free and $\mathbf{u}^{\mathrm{I}}(x, \theta, k_0)$ is not equal to $\mathbf{u}^{\mathrm{E}}(x, \theta, k_0)$, the minimal loss of information by using the M-estimators is*

$$G^{\mathrm{E}} - G^{\mathrm{I}} = \mathrm{E}_{\theta,k_0}[(\mathbf{u}^{\mathrm{E}} - \mathbf{u}^{\mathrm{I}})(\mathbf{u}^{\mathrm{E}} - \mathbf{u}^{\mathrm{I}})^{\mathrm{T}}]. \tag{5.14}$$

Proof. When S_θ is m-information-curvature free, we have $\mathbf{u}^{\mathrm{I}} = \mathbf{u}^{\mathrm{E}}$ from (5.11). Therefore, $G^{\mathrm{I}} = G^{\mathrm{E}}$, showing that $\mathbf{u}^{\mathrm{I}}(x, \theta, k_0)$ is optimal and efficient at k_0. On the other hand, $\mathbf{u}^{\mathrm{E}} = \mathbf{u}^{\mathrm{I}} + (\mathbf{u}^{\mathrm{E}} - \mathbf{u}^{\mathrm{I}})$ is an orthogonal decomposition so that the loss of information is given by (5.14). □

It should be noted that most semi-parametric models so far treated by many researchers are m-flat. The important role of m-flatness in the estimation function method is noted by Amari and Kumon (1988), Amari (1987a), and also by Bickel *et al.* (1993) under the name of convexity. The present result shows that the m-information-curvature freeness is essential, establishing a necessary and sufficient condition for the estimating function method to be lossless. We later give an example in which the model is not m-flat but m-information-curvature free, and the optimal estimating function is efficient. However, the optimal estimating function depends on the true k so that there is still a serious problem of choosing a good k_0 from observed data to derive a good estimating function. It is a merit of estimating functions that, even if we misspecify the true k and choose an incorrect k_0, the estimator is still $\sqrt{n}$-consistent. The next section proposes a method of choosing a simple but good quasi-estimating function based on the observed data.

6. Simple method of choosing good quasi-estimating functions

The final problem is how to choose a good estimating function. It is clear that when $\mathbf{u}^{\mathrm{I}}(x, \theta, k)$ does not depend on k, this $\mathbf{u}^{\mathrm{I}}(x, \theta)$ gives the best estimating function without any loss of Fisher information. There are many examples belonging to this class (Amari and Kumon 1988).

When $\mathbf{u}^{\mathrm{I}}(x, \theta, k)$ includes an unknown k, one orthodox idea is first to find a consistent estimator $\hat{k} = \hat{k}(x_1, \cdots, x_n, \theta)$ of k where θ is fixed. Given $x_1, \cdots, x_n$ and hence $\hat{k}$, we have that $\mathbf{u}^{\mathrm{I}}(x, \theta, \hat{k})$ belongs to $F^{\mathrm{I}}_{\theta,k} \oplus F^{\mathrm{A}}_{\theta,k}$ for any k. This is the optimal estimating function at $k = \hat{k}$ when it is applied to observations x_i other than those used for estimating $\hat{k}$. However, when we apply this estimating function to x_i from which $\hat{k}$ is obtained, we encounter a problem in analysing the estimating equation

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

because the $\mathbf{u}^{\mathrm{I}}(x_i, \theta, \hat{k})$ are not independent but are dependent through random variables $\hat{k}(x_1, \cdots, x_n, \theta)$. We call $\mathbf{u}^{\mathrm{I}}(x, \theta, \hat{k})$ a quasi-estimating function. Theoretically, we can often avoid the difficulty by dividing n observations into $\sqrt{n}$ and $n - \sqrt{n}$ disjoint subsets and by using $\sqrt{n}$ observations for estimating k and the other $n - \sqrt{n}$ observations for estimating θ. We do not discuss this problem further (see, for example, Pfanzagl 1990; Bhanja and Ghosh 1992). From a practical point of view, computer simulations show that the use of quasi-estimating functions $u_i(x, \theta, \hat{k})$ is justified.

It is the point of an estimating function $\mathbf{u}^{\mathrm{I}}(x, \theta, k)$ that even a misspecified k still gives a $\sqrt{n}$-consistent estimator. Therefore, it is wise for practical purposes to choose a simple but good k. We propose the following new method. Since we know at least in principle the set of all estimating functions, we choose a parametric family $\mathbf{y}(x, \theta; \eta)$ of estimating functions, where η is a finite-dimensional parameter. We then obtain an adequate $\hat{\eta}$ based on observed data $x_1, \cdots, x_n$. There are a number of methods for doing so. There again remains the problem of justifying the application of the quasi-estimating function $y(x, \theta; \hat{\eta})$ to the data x_i from which $\hat{\eta}$ is obtained. However, since $\hat{\eta}$ is finite-dimensional, it is easier to justify this by using expansions similar to (2.5). We do not discuss this point further. In order to choose $\hat{\eta}$, one idea is to use a parametrized subset of K,

$$M = \{k(\eta)\}, \qquad M \subset K,$$

where η is a finite-dimensional parameter that specifies k. Since the true k is not necessarily included in the subset M, we have a statistical model

$$S^* = \{p(x, \theta, \eta) = p[x, \theta, k(\eta)]\}$$

parametrized by a finite number of parameters (θ, η) which might not include the true distribution. It is not difficult to obtain an estimate $(\tilde{\theta}, \tilde{\eta})$, say, by the maximum likelihood method. However, this $\tilde{\theta}$ is not consistent in general and the idea of using $\tilde{\theta}$ was dismissed long ago. The new idea is to use the quasi-estimating function

$$\mathbf{u}^{\mathrm{I}}\{x, \theta, k(\tilde{\eta})\}$$

to obtain a good $\sqrt{n}$-consistent estimator $\hat{\theta}$. This type of idea is similar to one proposed by Lindsay (1985).

7. Examples

In order to explain the basic concepts, we use the simple mixture model of (2.13).

Example 2. *Mixture model.* A mixture model is m-flat and is hence m-information-curvature free. In particular, when a model is given by (2.13) and (2.14), the θ-score, the information score $\mathbf{u}^{\mathrm{I}}$, and the nuisance score in the direction of $a(\boldsymbol{\xi})$ can be calculated explicitly.

We first calculate the θ-score $\mathbf{u}$,

$$\begin{aligned}\mathbf{u} &= \frac{\partial}{\partial\theta}\log p \\ &= \frac{1}{p(x,\theta,k)}\int(\partial_\theta \mathbf{s}\cdot\boldsymbol{\xi} + \partial_\theta r - \partial_\theta\psi)k(\boldsymbol{\xi})\exp\{\boldsymbol{\xi}\cdot\mathbf{s} - r - \psi\}\mathrm{d}\boldsymbol{\xi}.\end{aligned}$$

Noting that the conditional distribution $p(\boldsymbol{\xi}|\mathbf{s})$ of $\boldsymbol{\xi}$ conditioned on $\mathbf{s}$ is written as

$$p(\boldsymbol{\xi}|\mathbf{s}) = \frac{k(\boldsymbol{\xi})\exp\{\boldsymbol{\xi}\cdot\mathbf{s} - \psi\}}{\displaystyle\int k(\boldsymbol{\xi})\exp\{\boldsymbol{\xi}\cdot\mathbf{s} - \psi\}\mathrm{d}\boldsymbol{\xi}},$$

the θ-score may be written as

$$\mathbf{u} = \partial_\theta \mathbf{s}\cdot\mathrm{E}[\boldsymbol{\xi}|\mathbf{s}] + \partial_\theta r - \mathrm{E}[\partial_\theta\psi|\mathbf{s}],$$

where $\mathrm{E}[\cdot|\mathbf{s}]$ is the conditional expectation. Similarly, the nuisance score in the direction of $a(\boldsymbol{\xi})$ is given by

$$v[a] = \mathrm{E}\left[\frac{a(\boldsymbol{\xi})}{k(\boldsymbol{\xi})}\middle|\mathbf{s}\right].$$

Therefore, $v[a]$ depends on x only through $\mathbf{s}$ so that the nuisance subspace $T^{\mathrm{N}}_{\theta,k}$ is generated by the random variable $\mathbf{s}(x,\theta)$.

It is known that the projection of a random variable t onto the space generated by s_i is given by the conditional expectation $\mathrm{E}[t|s_i]$ and the projection onto the orthogonal complement is $t - \mathrm{E}[t|s_i]$. Hence, the efficient score, which is the same as the information score in this case, is given by

$$\begin{aligned}\mathbf{u}^{\mathrm{I}} = \mathbf{u}^{\mathrm{E}} &= \mathbf{u} - \mathrm{E}[\mathbf{u}|\mathbf{s}] \\ &= \{\partial_\theta \mathbf{s} - \mathrm{E}[\partial_\theta \mathbf{s}|\mathbf{s}]\}\cdot\mathrm{E}[\boldsymbol{\xi}|\mathbf{s}] + \{\partial_\theta r - \mathrm{E}[\partial_\theta r|\mathbf{s}]\},\end{aligned}$$

where the vector notation should be understood appropriately. This gives the efficient

estimating functions. There is an interesting special case when $\partial_\theta \mathbf{s}$ is a function of $\mathbf{s}$. In this case, $\partial_\theta \mathbf{s} = \mathrm{E}[\partial_\theta \mathbf{s}|\mathbf{s}]$. The efficient score is then given by

$$\mathbf{u}^{\mathrm{E}} = \partial_\theta r - \mathrm{E}[\partial_\theta r|\mathbf{s}].$$

This does not depend on $k(\boldsymbol{\xi})$, so that $\mathbf{u}^{\mathrm{E}}$ is the optimal estimation function at any $k(\boldsymbol{\xi})$. There is no information loss.

Example 1. *Linear dependence model.* We now apply our general theory to the linear dependence problem of (2.12), where we put $\sigma^2 = 1$ for simplicity's sake. We further put $x = (\alpha, \beta)$, and

$$\begin{aligned} r(x,\theta) &= -\tfrac{1}{2}(\alpha^2 + \beta^2), \\ s(x,\theta) &= \alpha + \theta\beta, \\ \partial_\theta s(x,\theta) &= \beta, \\ \psi(\xi,\theta) &= \tfrac{1}{2}\xi^2(1+\theta^2) + \log 2\pi. \end{aligned}$$

The θ-score is

$$u = \beta \mathrm{E}[\xi|s] - \theta \mathrm{E}[\xi^2|s].$$

Since

$$\mathrm{E}[\beta|s] = \frac{\theta}{1+\theta^2} s,$$

we have

$$u^{\mathrm{I}} = u^{\mathrm{E}} = \frac{1}{1+\theta^2}\{\beta - \theta\alpha\}\mathrm{E}[\xi|s].$$

In order to obtain a good quasi-estimating function, we need to estimate $\mathrm{E}[\xi|s]$, which might not be easy.

We study a special case first. When $k(\boldsymbol{\xi})$ is a normal distribution with mean μ_ξ and variance σ_ξ^2, we have

$$p(s,\xi;\theta,k) = c(s)\exp\left\{\xi s - \frac{(\xi-\mu_\xi)^2}{2\sigma_\xi^2} - \frac{1}{2}(1+\theta^2)\xi^2\right\}.$$

Hence, after some calculations, we have

$$\mathrm{E}[\xi|s] = \mathrm{const.}\left(s + \frac{\mu_\xi}{\sigma_\xi^2}\right).$$

Therefore, the optimal estimating function is

$$\begin{aligned} u^{\mathrm{I}} &= \left(s + \frac{\mu_\xi}{\sigma_\xi^2}\right)(\beta - \theta\alpha) \\ &= \left(\alpha + \theta\beta + \frac{\mu_\xi}{\sigma_\xi^2}\right)(\beta - \theta\alpha). \end{aligned}$$

Motivated by this (see also Pfanzagl 1990), we consider a family of estimating functions parametrized by c,

$$y(x, \theta; c) = (\alpha + \theta\beta + c)(\beta - \theta\alpha).$$

The estimator obtained from $y(x, \theta; c)$ is called the c-estimator and is denoted by $\hat{\theta}_c$. This class includes various estimators. When $c \to \infty$, the estimator is

$$\hat{\theta}_\infty = \frac{\sum \beta_i}{\sum \alpha_i}.$$

We calculate the maximum likelihood estimators $(\hat{\theta}; \hat{\xi}_1, \hat{\xi}_2, \cdots, \hat{\xi}_n)$ of all the parameters. Here $\hat{\theta}$ is the least-squares estimator in the orthogonal direction, and is given by the 0-estimator $\hat{\theta}_0$. This estimator is not optimal. When the sequence $(\xi_1, \xi_2, \ldots)$ can be regarded as a realization from $N(\mu_\xi, \sigma_\xi^2)$, the estimator obtained from $c = \mu_\xi / \sigma_\xi^2$ is optimal.

The idea of the c-estimator springs from the following observation. Let us temporarily assume that $k(\xi)$ is subject to a normal distribution $N(\mu_\xi, \sigma_\xi^2)$. Then, the nuisance function k is parametrized by $\boldsymbol{\eta} = (\mu_\xi, \sigma_\xi^2)$. From the observed data, we have an estimator $\hat{\boldsymbol{\eta}} = (\hat{\mu}_\xi, \hat{\sigma}_\xi^2)$ in which

$$\hat{\mu}_\xi = \frac{1}{n} \sum \alpha_i$$

$$\hat{\sigma}_\xi^2 = \frac{1}{n} \sum \alpha_i^2 - \hat{\mu}_\xi^2 - 1.$$

We use the $\hat{c}$-quasi-estimating function, where $\hat{c}$ is given by

$$\hat{c} = \frac{\hat{\mu}_\xi}{\hat{\sigma}_\xi^2}.$$

The estimator $\hat{\theta}_{\hat{c}}$ is consistent and improves the maximum likelihood estimator even when $k(\xi)$ is not normal.

We now analyse the c-estimators. Given a sequence $(\xi_1, \xi_2, \cdots)$, the asymptotic variance of the c-estimator $\hat{\theta}_c$ is calculated from

$$V_c = \frac{(1+\theta^2)\{(c + (1+\theta^2)\bar{\xi})^2 + (1+\theta^2)^2(\overline{\xi^2} - \bar{\xi}^2) + (1+\theta^2)\}}{\{c\bar{\xi} + (1+\theta^2)\overline{\xi^2}\}^2},$$

where

$$\bar{\xi} = \frac{1}{n} \sum_{i=1}^{n} \xi_i,$$

$$\overline{\xi^2} = \frac{1}{n} \sum_{i=1}^{n} \xi_i^2.$$

The optimal value of c is given by

$$c_{\text{opt}} = \frac{\bar{\xi}}{\overline{\xi^2} - (\bar{\xi})^2}.$$

The unknown $\bar{\xi}$ and $\overline{\xi^2}$ are estimated by

$$\hat{\xi} = \frac{1}{n}\sum \alpha_i,$$

$$\widehat{\xi^2} = \frac{1}{n}\sum \alpha_i^2 - 1.$$

The estimator $\hat{\theta}_{\hat{c}}$ is not necessarily efficient, but is close to the optimal among the c-estimators. We show how it improves the maximum likelihood estimator. For example, when $\bar{\xi} = 1$, $\overline{\xi^2} = 2$, $\sigma^2 = 1$, and $\theta = 0$, the variances of the MLE $\hat{\theta}_0$ and of $\hat{\theta}_\infty$ are, respectively,

$$V[\bar{\theta}_{\mathrm{MLE}}] \approx \frac{3}{4n},$$

$$V[\hat{\theta}_\infty] \approx \frac{1}{n},$$

while that of the c_{opt} estimator is

$$V[\hat{\theta}_c] \approx \frac{2}{3n}.$$

Computer simulations show that $\hat{\theta}_{\hat{c}}$ attains this.

Example 3. *Semi-parametric additive regression*. This model, given by (2.15), is analysed here because it is not m-flat but is m-information-curvature free. The probability density function is written as

$$p(y, \mathbf{x}, t; \boldsymbol{\theta}, k) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\{y - \boldsymbol{\theta}\cdot\mathbf{x} - k(t)\}^2\right],$$

which is not linear in k. The $\boldsymbol{\theta}$-score vector at $(\boldsymbol{\theta}, k)$ is written as

$$\mathbf{u} = \mathbf{x}\{y - \boldsymbol{\theta}\cdot\mathbf{x} - k(t)\},$$

while the nuisance score in the direction of $a(t)$ of a change in function $k(t)$ is

$$v = a(t)\{y - \boldsymbol{\theta}\cdot\mathbf{x} - k(t)\}.$$

The efficient score is given by

$$\mathbf{u}^{\mathrm{E}} = \{\mathbf{x} - \mathrm{E}[\mathbf{x}|t]\}\{y - \boldsymbol{\theta}\cdot\mathbf{x} - k(t)\}.$$

Since the model is not m-flat, the m-parallel transport of the nuisance tangent space $T^{\mathrm{N}}_{\boldsymbol{\theta},k'}$ to $(\boldsymbol{\theta}, k)$ is different from $T^{\mathrm{N}}_{\boldsymbol{\theta},k}$. The m-parallel transport of v from k' to k is given by

$$\begin{aligned}\prod\nolimits^{(m)}_{k'}{}^{k} v &= \frac{p(y, \mathbf{x}, t; \boldsymbol{\theta}, k')}{p(y, \mathbf{x}, t; \boldsymbol{\theta}, k)} v \\ &= a(t)\{y - \boldsymbol{\theta}\cdot\mathbf{x} - k'(t)\} \exp\left\{(k' - k)\left(y - \boldsymbol{\theta}\cdot\mathbf{x} - \frac{k + k'}{2}\right)\right\}.\end{aligned}$$

This does not belong to $T^{\mathrm{N}}_{\theta,k}$. The information score $\mathbf{u}^{\mathrm{I}}$ is the projection of $\mathbf{u}$ onto the space orthogonal to that spanned by all of $\prod^{(m)\,k}_{k'} T^{\mathrm{N}}_{\theta,k'}$.

In the present case, we can show that

$$\prod^{(m)\,k}_{k'} T^{\mathrm{N}}_{\theta,k'} \subset T^{\mathrm{N}}_{\theta,k} \oplus T^{\mathrm{A}}_{\theta,k}.$$

This is proved by showing, for any $v \in T^{\mathrm{N}}_{\theta,k'}$,

$$\begin{aligned}
&\left\langle \mathbf{u}^{\mathrm{E}}, \prod^{(m)\,k}_{k'} v \right\rangle_{\theta,k} \\
&= \mathrm{E}\left[\{\mathbf{x} - \mathrm{E}[\mathbf{x}|t]\}a(t)(y - \theta\cdot\mathbf{x} - k)(y - \theta\cdot\mathbf{x} - k')\exp\left\{(k'-k)\left(y - \theta\cdot\mathbf{x} - \frac{k+k'}{2}\right)\right\}\right] \\
&= 0
\end{aligned}$$

To prove this, we first calculate the expectation of the above with respect to y, giving

$$\begin{aligned}
&\left\langle \mathbf{u}^{\mathrm{E}}, \prod^{(m)\,k}_{k'} v \right\rangle_{\theta,k} \\
&\quad = \mathrm{E}_{\mathbf{x},t}\left[(\mathbf{x} - \mathrm{E}[\mathbf{x}|t])a(t)\mathrm{E}_{\epsilon}\left[\epsilon(\epsilon - k' + k)\exp\left\{(k'-k)\left(\epsilon - \frac{k'-k}{2}\right)\right\}\right]\right],
\end{aligned}$$

where E_{ϵ} is the expectation with respect to the normal random variable

$$\epsilon = y - \theta\cdot\mathbf{x} - k(t)$$

and $\mathrm{E}_{\mathbf{x},t}$ is the expectation with respect to $q(\mathbf{x}, t)$. Since the term in E_{ϵ} depends on t but not on $\mathbf{x}$, by taking the conditional expectation $\mathrm{E}_{\mathbf{x}|t}$, we have

$$\left\langle \mathbf{u}^{\mathrm{E}}, \prod^{(m)\,k}_{k'} v \right\rangle_{\theta,k} = 0$$

which shows that S_{θ} is curved only in the direction of $T^{\mathrm{A}}_{\theta,k}$. Hence, we have

$$\mathbf{u}^{\mathrm{I}} = \mathbf{u}^{\mathrm{E}}.$$

Any estimating function is equivalently, or to within a matrix $T(\theta, k)$, written as

$$\begin{aligned}
\mathbf{y} &= \mathbf{u}^{\mathrm{I}} + \mathbf{a} \\
&= \{\mathbf{x} - \mathrm{E}[\mathbf{x}|t]\}\{y - \theta\cdot\mathbf{x} - k(t)\} + \mathbf{a}
\end{aligned}$$

for any $k(t)$ and $\mathbf{a} \in F^{\mathrm{A}}_{\theta,k}$. From (5.8), $G^{\mathrm{A}} = 0$ for the estimator with $\mathbf{a} = 0$, so that this gives an efficient estimator. It is noteworthy that, even when $k(t)$ is misspecified, this $\mathbf{y}$ gives a $\sqrt{n}$-consistent estimator provided $\mathrm{E}[\mathbf{x}|t]$ is known or estimated well from the data. Cuzick (1992) discussed efficient estimation of $\mathrm{E}[\mathbf{x}|t]$ and $k(t)$ by smoothing.

Acknowledgements

The authors thank the editor and reviewers for their constructive comments, suggestions and criticisms, which led to significant improvements to the paper.

References

Amari, S. (1985) *Differential-Geometrical Method in Statistics*. Lecture Notes in Statistics 28. New York: Springer-Verlag.

Amari, S. (1987a). Dual connections on the Hilbert bundles of statistical models. In C.T.J. Dodson (ed.), *Geometrization of Statistical Theory*, pp. 123–152. Lancaster: University of Lancaster Department of Mathematics.

Amari, S. (1987b) Differential geometry of a parametric family of invertible linear systems – Riemannian metric, dual affine connections and divergence. *Math. Systems Theory*, **20**, 53–82.

Amari, S. and Han, T.S. (1989) Statistical inference under multi-terminal rate restrictions. *IEEE Trans. Inform. Theory*, **35**, 217–227.

Amari, S. and Kumon, M. (1988) Estimation in the presence of infinitely many nuisance parameters – geometry of estimating functions. *Ann. Statist.*, **16**, 1044–1068.

Amari, S., Kurata, K. and Nagaoka, H. (1992) Information geometry of Boltzmann machines. *IEEE Trans. Neural Networks*, **3**, 260–271.

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

Barndorff-Nielsen, O.E. (1986) Likelihood and observed geometries. *Ann. Statist.*, **14**, 856–873.

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.

Bhanja, J. and Ghosh, J.K. (1992) Efficient estimation with many nuisance parameters. *Sankhyā Ser. A*, **54**, 1–39, 135–156.

Bickel, P.J., Klaassen, C.A.J., Ritov, Y. and Wellner, J.A. (1993) *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore, MD: Johns Hopkins University Press.

Cuzick, J. (1992) Semiparametric additive regression. *J. Roy. Statist. Soc. B*, **54**, 831–843.

Friedrich, T. von (1991) Die Fisher-Information und symplektische Strukturen. *Math. Nachr.*, **153**, 273–296.

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

Godambe, V.P. (ed.) (1991) *Estimating Functions*. New York: Oxford University Press.

Godambe, V.P. and Heyde, C.C. (1987) Quasi-likelihood and optimal estimation. *Internat. Statist. Rev.*, **55**, 231–244.

Groeneboom, P. and Wellner, J.A. (1992) *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Basel: Birkhäuser.

Hasminskii, R.Z. and Ibragimov, I.A. (1983) Efficient estimation in the presence of infinite dimensional incidental parameters. In *Probability Theory and Mathematical Statistics*. Lecture Notes in Math. 1021, pp. 195–229, New York: Springer-Verlag.

Huber, P.J. (1981) *Robust Statistics*. New York: Wiley.

Jaffe, A. and Quinn, F. (1993) Theoretical mathematics. *Bull. Amer. Math. Soc.*, **29**, 1–13.

Kanbayashi, T. (1994) Statistical manifold of infinite dimensions (in Japanese). *Trans. Japan SIAM*, **4**, 211–228.

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

Lafferty, J.D. (1988) The density manifold and configuration space, quantization. *Trans. Amer. Math. Soc.*, **305**, 699–741.

Levit, B.Y. (1978) Infinite-dimensional information inequalities. *Theory Probab. Appl.*, **23**, 371–377.

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

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

McLeish, D.L. and Small, C.G. (1988) *The Theory and Applications of Statistical Inference Functions*. Lecture Notes in Statistics 44. New York: Springer-Verlag.

Murray, M.K. and Rice, J.W. (1993) *Differential Geometry and Statistics*. New York: Chapman & Hall.

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

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

Okamoto, I., Amari, S. and Takeuchi, K. (1991) Asymptotic theory of sequential estimation: Differential geometrical approach. *Ann. Statist.*, **19**, 961–981.

Pfanzagl, J. (1990) *Estimation in Semiparametric Models: Some Recent Developments*. Lecture Notes in Statistics 63. New York: Springer-Verlag.

Ritov, Y. and Bickel, P.J. (1990) Achieving information bounds in non and semiparametric models. *Ann. Statist.*, **18**, 925–938.

Small, C.G. and McLeish, D.L. (1988) Generalization of ancillarity, completeness and sufficiency in an inference function space. *Ann. Statist.*, **16**, 534–551.

Small, C.G. and McLeish, D.L. (1989) Projection as a method for increasing sensitivity and eliminating nuisance parameters. *Biometrika*, **73**, 693–703.

van der Vaart, A.W. (1991) On differentiable functionals. *Ann. Statist.*, **19**, 178–205.

Received July 1994 and revised April 1996