Differential Geometry of Estimating functions in Semiparametric Statistical Models

Shun-ichi Amari and Motoaki Kawanabe

METR 93-20 November 1993

Differential Geometry of Estimating functions in Semiparametric Statistical Models

Shun-ichi Amari, Motoaki Kawanabe,
University of Tokyo*

November 30, 1993.

Abstract

For semiparametric statistical estimation, when an estimating function exists, it often provides robust and efficient estimation of the parameter of interest against nuisance parameters of infinite dimensions. The present paper elucidates the estimating function method by solving the following problems. 1)When does an estimating function exist and what is the set of all the estimating functions? 2)How is the efficiency of the estimators derived from estimating functions and when are they fully efficient? 3)How to construct an efficient or nearly efficient estimating function? The present theory is motivated by the dual differential geometry of statistical inference and its extension to fibre bundles, although we do not mention geometrical details.

Keywords

semiparametric model, estimating function, Hilbert fibred structure, effective score function, dual parallel transport, mixture m-flat.

*Department of Mathematical Engineering and Information Physics, Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan.

1 Introduction

A semiparametric statistical model treats a family of probability distributions specified by a finite-dimensional parameter θ of interest together with a nuisance parameter z of function-degrees of freedom. Estimation of the parameter of interest in such a model has attracted statisticians for long years, because various important problems are formulated in terms of semiparametric models. When the nuisance parameter is finite-dimensional, a fundamental role is played by the effective 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 is established in terms of the effective Fisher information and the bound is asymptotically attainable.

It is not easy to generalize these results to the semiparametric case. Levit(1978), Begun et al.(1983), Small and McLeish(1989) defined the effective Fisher information in the semiparametric model by using the projected score, and showed that the Cramér-Rao type inequality holds. It is only recently that its asymptotic attainability is established under a certain regularity conditions [see Groeneboom and Wellner(1992), Bickel, Klaassen, Ritov and Wellner(1993)] based on various efforts on functional analysis [e.g., Ritov and Bickel(1990), van der Vaart(1991)]. It is also known that there exists a pathological case where the bound is not asymptotically attainable [see Hasminskii and Ibragimov(1983)]. Estimation procedures are generally very complicated because of the infinite dimensionality of the nuisance parameter.

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

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

where $x_1, \cdots, x_n$ are i.i.d. observations from a semiparametric model. Here, the function $y(x, \theta)$ should satisfy

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

for all z, where $\mathrm{E}_{\theta,z}$ denotes the expectation with respect to the distribution specified by θ and z. Such a function $y(x,\theta)$ which does not include the nuisance parameter is called an estimating function [Godambe(1976)]. It gives a practically tractable method of estimation [see, for example McLeish and Small(1988), Godambe(1991), Godambe and Heyde(1987)]. The present paper aims at elucidating the estimating function method and its efficiency. The estimating function method has a merit of robustness in the sense that it always gives a consistent estimator whatever y we choose, while the maximum-likelihood method does not necessarily have this property. Therefore, it is important to know when an estimating function exists and how efficient such a simple estimator is.

We prove in the present paper that the Cramér-Rao bound of the effective Fisher information is attainable by using an adequate estimating function, when a statistical model has the m-flat nuisance structure. This is the case with many important semiparametric models. In this case, the estimating function method combined with the adaptive method gives a fully efficient estimator. The problem is again how to choose the best estimating function. However, the point is that even if the estimated nuisance parameter $\hat{z}$ is wrong or is arbitrarily assigned, the estimating function method gives an estimator which might not be fully efficient but is still consistent.

The present paper studies the structure and efficiency of estimating functions by answering the following fundamental problems :

1. When does an estimating function exist?
2. What is the set of all the estimating functions?
3. What is the best estimating function?
4. When does the best estimating function gives a fully efficient estimator and what is the amount of loss of information caused by using the estimation function method?
5. How to construct the best or a very good estimating function?

The method of the present paper is an extension of Amari and Kumon(1988), and Amari(1987). They studied a special but important type of the semiparametric models where the number of nuisance parameters increases in proportion to the number of observations. There are lots of research on this subject [Neyman and Scott(1948), Andersen(1970), Lindsay(1982,1985), Bhanja and Ghosh(1992)]. They obtained a condition that guarantees the existence of the best estimating function that gives a uniformly efficient estimator whatever z is. Although this theory is applicable only to the very limited cases that the best estimating function exists irrespective of z, the theory elucidated the geometrical structure of semiparametric models. The theory is motivated by the information geometry [Amari(1985), Nagaoka and Amari(1982)], 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 sciences [Amari(1985,1987), Amari and Han(1989), Amari and Kumon(1988) Okamoto, Amari and Takeuchi(1990), Amari, Kurata and Nagaoka(1992)]. However, we do not enter in details of the dual differential geometry of the manifold of function-degrees of freedom, nor details of functional analysis [see Bickel et al.(1992)], because it is still mathematically not easy to construct differential geometry of function spaces rigorously.

2 Semiparametric statistical model and estimating function

Let $p(x, \boldsymbol{\theta}, z)$ be a probability density functions of a random variable x with respect to a common measure $\mu(x)$, specified by two kinds of parameters $\boldsymbol{\theta} = (\theta^1, \cdots, \theta^k)$ and z, where $\boldsymbol{\theta} \in \Theta$ is a finite-dimensional vector and $z \in Z$ is a parameter having function-degrees of freedom. The set of distributions $S = \{p(x, \boldsymbol{\theta}, z)\}$ is called a semiparametric statistical model, where $\boldsymbol{\theta}$ is called the parameter of interest and z is called the nuisance parameter.

Let $\boldsymbol{y}(x, \boldsymbol{\theta}) = [y_i(x, \boldsymbol{\theta})],\ i = 1, \cdots, k$, be a vector-valued smooth function

of x and $\boldsymbol{\theta}$ having the same dimensions as $\boldsymbol{\theta}$, not depending on z. Such a function is called an estimating function, when it satisfies the following conditions,

$$\begin{aligned} \mathrm{E}_{\boldsymbol{\theta},z}[\boldsymbol{y}(x,\boldsymbol{\theta})] &= 0, & (2.1)\\ \det|\mathrm{E}_{\boldsymbol{\theta},z}[\partial_{\boldsymbol{\theta}}\boldsymbol{y}(x,\boldsymbol{\theta})]| &\neq 0, & (2.2)\\ \mathrm{E}_{\boldsymbol{\theta},z}[\| \boldsymbol{y}(x,\boldsymbol{\theta}) \|^2] &< \infty, & (2.3) \end{aligned}$$

for all $\boldsymbol{\theta}$ and z, where $\mathrm{E}_{\boldsymbol{\theta},z}$ denotes the expectation with respect to the distribution $p(x,\boldsymbol{\theta},z)$, $\partial_{\boldsymbol{\theta}}\boldsymbol{y}$ is the gradient of $\boldsymbol{y}$ with respect to $\boldsymbol{\theta}$, i.e., the matrix whose elements are $(\partial y_i/\partial\theta^j)$ in the component form, $\det|\ \ |$ denotes the determinant of a matrix, and $\| \boldsymbol{y} \|^2$ is the squared norm of the vector $\boldsymbol{y}$, $\| \boldsymbol{y} \|^2 = \sum (y_i)^2$.

When a function satisfying (2.1) $\sim$ (2.3) exists, we have an estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ by solving

$$\sum_{i=1}^{n} \boldsymbol{y}(x_i,\boldsymbol{\theta}) = 0, \qquad (2.4)$$

where $x_1,\cdots,x_n$ are n i.i.d. observations. This is called the estimating equation.

The asymptotic behavior of the estimator $\hat{\boldsymbol{\theta}}$ is obtained from the expansion

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

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

$$\begin{aligned} \frac{1}{n}\sum \partial_{\boldsymbol{\theta}}\boldsymbol{y}(x_i,\boldsymbol{\theta}) &\sim \mathrm{E}_{\boldsymbol{\theta},z}[\partial_{\boldsymbol{\theta}}\boldsymbol{y}(x,\boldsymbol{\theta})] = A, & (2.6)\\ \frac{1}{\sqrt{n}}\sum \boldsymbol{y}(x_i,\boldsymbol{\theta}) &\sim \boldsymbol{\varepsilon}, & (2.7) \end{aligned}$$

where $\boldsymbol{\varepsilon}$ is the normal random variable subject to $N(0,V)$ with

$$V = E_{\boldsymbol{\theta},z}[\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}], \qquad (V_{ij} = E_{\boldsymbol{\theta},z}[y_i y_j]) \qquad (2.8)$$

and $\boldsymbol{y}$ is a column vector and $\boldsymbol{y}^{\mathrm{T}}$ is its transposition. From this, by neglecting higher order terms, we have

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

The following theorem was immediately proved under the ordinary regularity conditions [Godambe(1976)].

Theorem 1 *The estimator $\hat{\boldsymbol{\theta}}$ obtained from an estimating function $\boldsymbol{y}(x, \boldsymbol{\theta})$ is asymptotically consistent and asymptotically normally distributed, with the asymptotic covariance matrix*

$$\mathrm{AV}[\boldsymbol{y}] = A^{-1}E_{\boldsymbol{\theta},z}[\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}](A^{\mathrm{T}})^{-1}, \tag{2.10}$$

where the asymptotic covariance matrix is defined by

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

Let $T(\boldsymbol{\theta})$ be a non-singular $k \times k$ matrix. It should be noted that $\boldsymbol{y}^*(x, \boldsymbol{\theta}) = T(\boldsymbol{\theta})\boldsymbol{y}(x, \boldsymbol{\theta})$ gives an estimating function equivalent to $\boldsymbol{y}$.

We give some examples of semiparametric statistical models. See Bickel et al.(1993) and Groeneboom and Wellner(1992) for many other interesting models.

1. Location-scale model :

   When density functions of $x \in \mathbf{R}^1$ are given by

   $$p(x, \mu, \sigma) = \frac{1}{\sigma} z\left(\frac{x-\mu}{\sigma}\right), \tag{2.12}$$

   where $\boldsymbol{\theta} = (\mu, \sigma)$ is the two-dimensional location and scale parameters of interest and z is an unknown shape of the density function, the family is called the location-scale model. For the sake of convenience, we represent the density function relative to the normal measure

   $$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\},$$

so that we treat the following representation

$$p(x, \mu, \sigma) = \frac{1}{\sigma} z\left(\frac{x-\mu}{\sigma}\right) \varphi\left(\frac{x-\mu}{\sigma}\right), \tag{2.13}$$

instead of (2.12), where z is assumed to be smooth and square integrable with respect to the normal measure and $z(x)\varphi(x)$ is rapidly decreasing. In order that the parameters μ and σ are identifiable and that it represents a probability distribution, we pose the following conditions,

$$\begin{aligned} z(x) &> 0, \\ \int z(x) d\mu^*(x) &= 1, \\ \int x z(x) d\mu^*(x) &= 0, \\ \int x^2 z(x) d\mu^*(x) &= 1 \end{aligned} \tag{2.14}$$

where

$$d\mu^*(x) = \varphi(x) d\mu(x).$$

This model will be analyzed in detail by Amari et al.(1993).

2. Mixture model :

Let $\{q(x, \boldsymbol{\theta}, \boldsymbol{\xi})\}$ be a regular statistical model, where both the parameter of interest $\boldsymbol{\theta}$ and the nuisance parameter $\boldsymbol{\xi}$ are of finite dimensions. Let x_i, $i = 1, 2, \cdots, n$, be n independent observations from $q(x, \boldsymbol{\theta}, \boldsymbol{\xi}_i)$, where $\boldsymbol{\theta}$ is common but $\boldsymbol{\xi}_i$ takes different values 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 $z(\boldsymbol{\xi})$. Then, x_i are regarded as i.i.d. observations from the semiparametric model

$$p(x, \boldsymbol{\theta}, z) = \int q(x, \boldsymbol{\theta}, \boldsymbol{\xi}) z(\boldsymbol{\xi}) d\boldsymbol{\xi}, \tag{2.15}$$

where $z(\boldsymbol{\xi})$ is the nuisance parameter of function-degrees of freedom. This model is called the mixture model. When $\boldsymbol{\xi}_i$ are not random

but a fixed unknown sequence, we may regard the empirical distribution $Z(\boldsymbol{\xi})$ as the governing probability law [see van der Vaart(1991)]. This problem was studied by Neyman and Scott(1948) and has attracted many researchers [Andersen(1970), Lindsay(1982), Kumon and Amari(1984), Amari and Kumon(1988), etc.]. Most researchers have treated the distributions of the following exponential form as examples,

$$q(x, \boldsymbol{\theta}, \boldsymbol{\xi}) = \exp\{\boldsymbol{\xi} \cdot \boldsymbol{s}(x, \boldsymbol{\theta}) + r(x, \boldsymbol{\theta}) - \psi(\boldsymbol{\theta}, \boldsymbol{\xi})\}, \tag{2.16}$$

where $\boldsymbol{s}(x, \boldsymbol{\theta})$ is a vector with the same dimensions as $\boldsymbol{\xi}$, not depending on $\boldsymbol{\xi}$ and $\cdot$ is the inner product. Here, the distribution is of exponential type for $\boldsymbol{\xi}$ when $\boldsymbol{\theta}$ is fixed. This type of models has the m-flat nuisance structure to be explained later so that they possess nice properties as will be shown later.

3. Linear binary choice model :

   Let $\mathbf{x} \in \mathbf{R}^m$ be a random variable subject to the normal distribution$N(\mathbf{0}, I)$, where I is the identity matrix. Let q be a binary random variable taking values on 1 and -1. Given $\boldsymbol{x}$, the conditional probability of q is assumed to depend only on $\boldsymbol{\theta} \cdot \boldsymbol{x}$,

$$p(q|\boldsymbol{x}; \boldsymbol{\theta}, z) = 0.5 + qz(\boldsymbol{\theta} \cdot \boldsymbol{x}), \tag{2.17}$$

   where $\boldsymbol{\theta} \in \mathbf{R}^m$ is a unit vector ($|\boldsymbol{\theta}| = 1$) which is the parameter of interest and z is a monotonically increasing smooth function satisfying

$$\begin{aligned} z(0) &= 0, \\ z'(0) &> 0, \\ |z(u)| &< 0.5. \end{aligned}$$

   This model shows that, above the hyperplane passing through the origin

$$\boldsymbol{\theta} \cdot \boldsymbol{x} = 0,$$

the probability of success ($q = 1$) is larger than 0.5 and below the hyperplane the probability of failure is larger than 0.5, depending on the distance of $\boldsymbol{x}$ from the unknown hyperplane. We want to estimate the hyperplane or $\boldsymbol{\theta}$ in the case where the function z is unknown. Such a problem occurs in the fields of economics, pattern recognition and neural networks [Nawata(1989), Manski(1985), Kawanabe, Amari and Hiroshige(1993)].

4. Semiparametric dose-response model :

   When x is one-dimensional and the separating hyperplane (point) does not necessarily pass through the origin, we have the following simple model

$$p(q|x, \theta, z) = 0.5 + qz(x - \theta), \tag{2.18}$$

   where θ is the separation point. We may regard this a one-dimensional version of the binary choice model. This is a non-parametric dose-response model, if we interpret x as the amount of dose and $q = \pm 1$ is the result after treatment where $q = 1$ represents the success of treatment and $q = -1$ failure. Here θ is the point that the probabilities of success and failure are fifty-fifty.

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 starting from $p(x)$,

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

where t ($\varepsilon > t \geq 0$) 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

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

In order to be specific, we consider the linear space of rapidly decreasing smooth functions satisfying

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

The closure of such a set is a Hilbert space H_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

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

the tangent vector of the curve (3.1). This is the score function for the one-dimensional statistical model (3.1) parameterized by t. More precisely, in order to define the tangent vector at $p(x) \in S$ in an infinite-dimensional space S, we need to use the Fréchet derivative in the sense of the Hellinger distance instead of the pointwise differentiation of (3.5). Refer to Bickel at el.(1993) and Groeneboom and Wellner(1992) for mathematical details.

Given a semiparametric model $S = \{p(x, \boldsymbol{\theta}, z)\}$, a Hilbert space $H_p = H_{\boldsymbol{\theta},z}$ is associated at each point $(\boldsymbol{\theta}, z)$, i.e., at each distribution $p(x) = p(x, \boldsymbol{\theta}, z)$ specified by $(\boldsymbol{\theta}, z)$. A collection of such $H_{\boldsymbol{\theta},z}$ is called a fibred structure, where the fibres are the Hilbert spaces.

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

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

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

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

is satisfied and we further assume that it is square-integrable. Then it belongs to $H_{\boldsymbol{\theta},z}$. We call the subspace spanned by these u_i's the tangent subspace $T^I_{\boldsymbol{\theta},z}$ along the parameter of interest. (A more rigorous theory requires the pathwise differentiability of $\boldsymbol{\theta}$ in the sense of van der Vaart(1991).) The score function vector is $\boldsymbol{u} = (u_1, \cdots, u_k)$.

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

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

belongs to the $H_{\boldsymbol{\theta},z}$. This v is the tangent vector along $c(t)$ of the nuisance parameter. Let $T^N_{\boldsymbol{\theta},z}$ be the smallest closed subspace including all such v's. We call it the nuisance tangent space.

Now, let us project the score functions u_i to the subspace $\left(T^N_{\boldsymbol{\theta},z}\right)^{\perp}$ which is the orthogonal complement of $T^N_{\boldsymbol{\theta},z}$. The result is the function $u_i^E = u_i - v$ that minimizes

$$|u_i - v|^2, \qquad v \in T^N_{\boldsymbol{\theta},z}.$$

These functions u_i^E are called the effective or projected score functions [see Begun et al.(1983), Amari and Kumon(1988), Small and McLeish (1989)]. Let $T^E_{\boldsymbol{\theta},z}$ be the subspace spanned by the effective score functions u_i^E's.

Let $T^A_{\boldsymbol{\theta},z}$ be the orthogonal complement of $T^N_{\boldsymbol{\theta},z} \oplus T^E_{\boldsymbol{\theta},z}$ called the ancillary subspace which represents 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 [Amari and Kumon(1988); see also Small and McLeish(1988)],

$$H_{\boldsymbol{\theta},z} = T^E_{\boldsymbol{\theta},z} \oplus T^A_{\boldsymbol{\theta},z} \oplus T^N_{\boldsymbol{\theta},z}. \tag{3.9}$$

The matrix $G^E = (g^E_{ij})$ defined by using the effective score functions

$$g^E_{ij} = \mathrm{E}_{\boldsymbol{\theta},z}[u_i^E u_j^E] \tag{3.10}$$

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

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

for any asymptotically normally distributed unbiased estimators in a semi-parametric model. Recently, an exact theory is constructed to show that the above bound is asymptotically attainable under mild regularity conditions [Bickel et al.(1992)]. It is important to show when the efficient estimator is given by an estimating function.

4 Invariant decomposition of Hilbert fibres due to dual parallel transports

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

Let $a(x)$ be a random variable belonging to $H_{\boldsymbol{\theta},z}$. Let us fix $\boldsymbol{\theta}$, and consider a subset $S_{\boldsymbol{\theta}} = \{p(x, \boldsymbol{\theta}, z) | z \in Z\}$. We define two parallel transports of a vector $a(x)$ at $H_{\boldsymbol{\theta},z}$ to $H_{\boldsymbol{\theta},z'}$. The following

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

$$\prod_{z}^{(e)z'} a(x) = \frac{p(x, \boldsymbol{\theta}, z)}{p(x, \boldsymbol{\theta}, z')} a(x) \tag{4.2}$$

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

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

always holds, and

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

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

The parallel transports are generalizations of the dual geometrical structures derived from the underlying e- and m-connections or e- and m-covariant derivatives in the finite dimensional case [see Amari(1985), see also Amari and Kumon(1988)], but we do 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 $a(x), b(x) \in H_{\boldsymbol{\theta},z}$, the inner product is kept invariant when their parallel transports belong to $H_{\boldsymbol{\theta},z'}$,*

$$\langle a, b\rangle_{\boldsymbol{\theta},z} = \left\langle \prod\nolimits_{z}^{(e)\,z'} a, \prod\nolimits_{z}^{(e)\,z'} b \right\rangle_{\boldsymbol{\theta},z'}, \tag{4.3}$$

where the suffix $(\boldsymbol{\theta}, z)$ denotes that the expectation is taken with respect to $p(x, \boldsymbol{\theta}, z)$.

We now reorganize the decomposition (3.9) of $H_{\boldsymbol{\theta},z}$ by taking account of the global structure induced by the parallel transports. The information fibre $F^I_{\boldsymbol{\theta},z}$ at $(\boldsymbol{\theta}, z)$ is constructed such that it is orthogonal to the nuisance spaces at any points $(\boldsymbol{\theta}, z')$, $z' \in Z$. To this end, we first consider a vector

$$r(x) \in T^E_{\boldsymbol{\theta},z} \oplus T^A_{\boldsymbol{\theta},z}$$

whose e-transport exists in $H_{\boldsymbol{\theta},z'}$ for any z', that is

$$\mathrm{E}_{\boldsymbol{\theta},z'}[r(x)^2] < \infty. \tag{4.4}$$

If its e-transport to $(\boldsymbol{\theta}, z')$ is orthogonal to $T^N_{\boldsymbol{\theta},z'}$ for any $z' \in Z$, that is,

$$\left\langle v,\ \prod\nolimits_{z}^{(e)\,z'} r(x) \right\rangle_{\boldsymbol{\theta},z'} = 0, \qquad v \in T^N_{\boldsymbol{\theta},z'}, \tag{4.5}$$

it is free of any nuisance directions at any z' when it is e-transported. We consider the closed subspace of $H_{\boldsymbol{\theta},z}$ consisting of the vectors satisfying the

above conditions (4.4) and (4.5) and denote it tentatively by $F^{I+A}_{\boldsymbol{\theta},z}$, where I denotes the information part and A denotes the ancillary part. Obviously

$$F^{I+A}_{\boldsymbol{\theta},z} \subset T^{E}_{\boldsymbol{\theta},z} \oplus T^{A}_{\boldsymbol{\theta},z}. \tag{4.6}$$

The nuisance fibre space $F^{N}_{\boldsymbol{\theta},z}$ is defined by the orthogonal complement of $F^{I+A}_{\boldsymbol{\theta},z}$ in $H_{\boldsymbol{\theta},z}$. Obviously

$$F^{N}_{\boldsymbol{\theta},z} \supset T^{N}_{\boldsymbol{\theta},z}. \tag{4.7}$$

We have

$$H_{\boldsymbol{\theta},z} = F^{I+A}_{\boldsymbol{\theta},z} \oplus F^{N}_{\boldsymbol{\theta},z}. \tag{4.8}$$

We now decompose $F^{I+A}_{\boldsymbol{\theta},z}$. Let u^{I}_{i} be the projection of the score u_i to $F^{I+A}_{\boldsymbol{\theta},z}$. The information fibre space denoted by $F^{I}_{\boldsymbol{\theta},z}$ is the subspace spanned by the information score functions $u^{I}_{i}(x,\boldsymbol{\theta},z)$. Its orthogonal subspace in $F^{I+A}_{\boldsymbol{\theta},z}$ is the ancillary fibre space denoted by $F^{A}_{\boldsymbol{\theta},z}$,

$$F^{IA}_{\boldsymbol{\theta},z} = F^{I}_{\boldsymbol{\theta},z} \oplus F^{A}_{\boldsymbol{\theta},z}.$$

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

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

which represents a more global structure of $H_{\boldsymbol{\theta},z}$. The $F^{N}_{\boldsymbol{\theta},z}$ includes all the m-parallel transports of $v \in T^{N}_{\boldsymbol{\theta},z'}$ from any $(\boldsymbol{\theta},z')$ to $(\boldsymbol{\theta},z)$ when it belongs to $H_{\boldsymbol{\theta},z}$, because of the relation (4.5). It is the information fibre $F^{I}_{\boldsymbol{\theta},z}$ that plays an important role.

However, in many cases, the e-parallel transports of $T^{E}_{\boldsymbol{\theta},z} \oplus T^{A}_{\boldsymbol{\theta},z}$ belongs to $H_{\boldsymbol{\theta},z'}$ for any z'. Moreover, it is orthogonal to $T^{N}_{\boldsymbol{\theta},z'}$. In such a case, the situation becomes very simple, because

$$\begin{aligned} F^{N}_{\boldsymbol{\theta},z} &= T^{N}_{\boldsymbol{\theta},z}, \\ F^{I}_{\boldsymbol{\theta},z} &= T^{E}_{\boldsymbol{\theta},z}, \\ F^{A}_{\boldsymbol{\theta},z} &= T^{A}_{\boldsymbol{\theta},z}, \end{aligned} \tag{4.10}$$

and the information fibre coincides with the effective score space. In most examples we treat here, the conditions are satisfied. The following is a typical case.

By fixing $\boldsymbol{\theta}$, we have the statistical model $S_{\boldsymbol{\theta}} = \{p(x, \boldsymbol{\theta}, z)\}$ where $z \in Z$ is the only parameter. The submodel $S_{\boldsymbol{\theta}}$ is said to be m-flat when, for any z_1, $z_2 \in Z$, the mixture family connecting them,

$$\begin{aligned} p(x,t) &= (1-t)p(x,\boldsymbol{\theta},z_1) + tp(x,\boldsymbol{\theta},z_2) \\ &= p(x,\boldsymbol{\theta},(1-t)z_1 + tz_2), \end{aligned} \tag{4.11}$$

also belongs to $S_{\boldsymbol{\theta}}$ and that the tangent vector of the curve connecting z_1 and z_2,

$$v = \frac{p(x,\boldsymbol{\theta},z_2) - p(x,\boldsymbol{\theta},z_1)}{p(x,\boldsymbol{\theta},z_1)},$$

belongs to $H_{\boldsymbol{\theta},z_1}$. In this case,

$$\prod\nolimits_{z}^{(m)\,z'} T_{\boldsymbol{\theta},z} = T_{\boldsymbol{\theta},z'} \tag{4.12}$$

holds, and $F^N_{\boldsymbol{\theta},z} = T^N_{\boldsymbol{\theta},z}$, $F^I_{\boldsymbol{\theta},z} = T^E_{\boldsymbol{\theta},z}$, provided the e-transport of u^E_i belongs to $H_{\boldsymbol{\theta},z'}$.

5 Estimating functions and their efficiency

Based on the decomposition of the Hilbert space $H_{\boldsymbol{\theta},z}$, 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.

Theorem 2 *An estimating function exists when and only when $F^I_{\boldsymbol{\theta},z}$ is non-degenerate at any z and the projections of $\prod^{(e)}{}_{z}^{z'} u^E_i$ span $F^I_{\boldsymbol{\theta},z'}$ at any z'.*

Theorem 3 *Any estimating function $\boldsymbol{y}(x,\boldsymbol{\theta}) = (\, y_i(x,\boldsymbol{\theta})\,)$ can be represented by a sum*

$$y_i(x,\boldsymbol{\theta}) = u^I_i(x,\boldsymbol{\theta}) + a_i(x,\boldsymbol{\theta}),\ a \in F^A_{\boldsymbol{\theta},z} \tag{5.1}$$

and conversely such a sum gives an estimating function when the projection of the components y_i to $F^I_{\boldsymbol{\theta},z}$, spans it.

The theorems are proved by the following lemmas.

Lemma 2 *Any estimating function $\boldsymbol{y}(x,\boldsymbol{\theta})$ belongs to $F^I_{\boldsymbol{\theta},z} \oplus F^A_{\boldsymbol{\theta},z}$. Let $y^I_i(x,\boldsymbol{\theta})$ be the projection of the i-th component of $\boldsymbol{y}(x,\boldsymbol{\theta})$ to $F^I_{\boldsymbol{\theta},z}$. Then, $y^I_i(x,\boldsymbol{\theta})$, $i=1,\cdots,n$, span $F^I_{\boldsymbol{\theta},z}$.*

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

$$\prod^{(e)} \boldsymbol{y}(x,\boldsymbol{\theta}) = \boldsymbol{y}(x,\boldsymbol{\theta}), \tag{5.2}$$

because of (2.1) and (4.1). By differentiating

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

along a curve $z = z(t)$ whose direction is $v \in T^N_{\boldsymbol{\theta},z}$, we have, for any $z \in Z$,

$$\begin{aligned}
&\frac{d}{dt}\int p(x,\boldsymbol{\theta},z(t))\boldsymbol{y}(x,\boldsymbol{\theta})d\mu(x) \\
&\quad = \int v(x,\boldsymbol{\theta},z(t))p(x,\boldsymbol{\theta},z(t))\boldsymbol{y}(x,\boldsymbol{\theta})d\mu(x) \\
&\quad = \langle v, \boldsymbol{y}(x,\boldsymbol{\theta})\rangle_{\boldsymbol{\theta},z} = 0.
\end{aligned} \tag{5.3}$$

This shows that $\boldsymbol{y}$ is included $F^I_{\boldsymbol{\theta},z} \oplus F^A_{\boldsymbol{\theta},z}$. Moreover, by differentiating (5.1) with respect to $\boldsymbol{\theta}$, we have

$$\mathrm{E}_{\boldsymbol{\theta},z}[\partial_{\boldsymbol{\theta}}\boldsymbol{y}(x,\boldsymbol{\theta})] + \langle \boldsymbol{u}, \boldsymbol{y}(x,\boldsymbol{\theta})\rangle = 0, \tag{5.4}$$

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

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

Since $\boldsymbol{y}$ belongs to $F^I_{\boldsymbol{\theta},z} \oplus F^A_{\boldsymbol{\theta},z}$ and the projection of $\boldsymbol{u}$ to the space $F^I_{\boldsymbol{\theta},z} \oplus F^A_{\boldsymbol{\theta},z}$ includes only $F^I_{\boldsymbol{\theta},z}$-part, we have

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

where $\boldsymbol{u}^I$ is the projection of $\boldsymbol{u}$ to $F^I_{\boldsymbol{\theta},z}$. Therefore, (2.2) implies that the determinant of $\langle u^I_i, y_j \rangle$ does not vanish, implying that the projections of vectors y_i's on $F^I_{\boldsymbol{\theta},z}$ span $F^I_{\boldsymbol{\theta},z}$ and also that $F^I_{\boldsymbol{\theta},z}$ is non-degenerate, that is, its dimension number is the same as that of the parameter of interest $\boldsymbol{\theta}$. □

Now we prove the converse of Lemma 2

Lemma 3 *When a vector $\boldsymbol{w}(x,\boldsymbol{\theta})$ belongs to $F^I_{\boldsymbol{\theta},z} \oplus F^A_{\boldsymbol{\theta},z}$ for some $z \in Z$, it automatically belongs to $F^I_{\boldsymbol{\theta},z'} \oplus F^A_{\boldsymbol{\theta},z'}$ for any $z' \in Z$. It is an estimating function when the projections of its components w_i to $F^I_{\boldsymbol{\theta},z'}$ span $F^I_{\boldsymbol{\theta},z'}$ for any z'.*

Proof We first prove that $\boldsymbol{w}(x,\boldsymbol{\theta})$ belonging to $F^I_{\boldsymbol{\theta},z} \oplus F^A_{\boldsymbol{\theta},z}$ is e-parallel invariant,

$$\prod^{(e)}{}^{z'}_{z} \boldsymbol{w}(t) = \boldsymbol{w}(t).$$

Let $c(t)$ be a curve connecting two points z and z' and we put

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

Obviously $\boldsymbol{f}(0) = 0$. By differentiation, we have

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

where

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

By m- and e-transporting v and w from $c(t)$ to z, respectively, we have

$$\langle v, \boldsymbol{w}\rangle_{c(t)} = \left\langle \prod^{(m)}{}^{z}_{c(t)} v, \prod^{(e)}{}^{z}_{c(t)} \boldsymbol{w} \right\rangle_{\boldsymbol{\theta},z}.$$

Since

$$\prod^{(e)}{}^{z}_{c(t)} \boldsymbol{w} = \boldsymbol{w} - \mathrm{E}_{\boldsymbol{\theta},z}[\boldsymbol{w}] \in F^I_{\boldsymbol{\theta},z} \oplus F^A_{\boldsymbol{\theta},z},$$

we have

$$\frac{d}{dt}\boldsymbol{f}(t) = 0.$$

Therefore,

$$\mathrm{E}_{\boldsymbol{\theta},c(t)}[\boldsymbol{w}(x,\boldsymbol{\theta})] = 0$$

holds, proving the e-invariancy of $\boldsymbol{w}$. From

$$\left\langle \prod^{(m)} {}^{z'}_{z''} v, \boldsymbol{w} \right\rangle_{z'} = \left\langle \prod^{(m)} {}^{z}_{z''} v, \boldsymbol{w} \right\rangle_{z} = 0,$$

it is shown that $\boldsymbol{w} \in F^I_{\boldsymbol{\theta},z'} \oplus F^A_{\boldsymbol{\theta},z'}$ for any z'. Since $\boldsymbol{w}$ spans $F^I_{\boldsymbol{\theta},z'}$ for any z', $\mathrm{E}_{\boldsymbol{\theta},z'}[\partial_{\boldsymbol{\theta}}\boldsymbol{w}(x,\boldsymbol{\theta})]$ is non-degenerate. Therefore, $\boldsymbol{w}$ is an estimating function satisfying (2.1) $\sim$ (2.3). □

The next important problem is to calculate the asymptotic covariance matrix of an estimating function. This calculation leads us to the optimal estimating function. An estimating function $\boldsymbol{y}(x,\boldsymbol{\theta})$ is decomposed into

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

where $\boldsymbol{u}^I = (u^I_i)$ and $\boldsymbol{a} = (a_i) \in F^A_{\boldsymbol{\theta},z}$. It is easy to show

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

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

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

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

where we put

$$G^I = \mathrm{E}[\boldsymbol{u}^I {\boldsymbol{u}^I}^{\mathrm{T}}], \quad G^A = \mathrm{E}[\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}]. \tag{5.7}$$

So we have the following theorem.

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

$$V[\boldsymbol{y}] = (G^I)^{-1} + (G^I)^{-1} G^A (G^I)^{-1}. \tag{5.8}$$

The estimating function given by

$$\boldsymbol{y}(x, \boldsymbol{\theta}) = \boldsymbol{u}^I(x, \boldsymbol{\theta}, z_0) \tag{5.9}$$

is the optimal at $(\boldsymbol{\theta}, z_0)$*, with the asymptotic covariance* $(G^I)^{-1}$*.*

We can now answer the question of what is the amount of loss of information by using the method of estimating functions compared to other estimating methods. We also show when the best estimating function is lossless, that is fully efficient, provided the best estimating function could be chosen.

Theorem 5 *The optimal estimating function is fully efficient when* $S_{\boldsymbol{\theta}}$ *is* m*-flat. When* $S_{\boldsymbol{\theta}}$ *is not* m*-flat, the loss of information is given by*

$$G^E - G^I = E[(\boldsymbol{u}^I - \boldsymbol{u}^E)(\boldsymbol{u}^I - \boldsymbol{u}^E)^{\mathrm{T}}]. \tag{5.10}$$

Proof When $S_{\boldsymbol{\theta}}$ is m-flat, $T^N_{\boldsymbol{\theta},z} = F^N_{\boldsymbol{\theta},z}$ so that

$$\boldsymbol{u}^I = \boldsymbol{u}^E$$

holds. Therefore, $G^I = G^E$. On the other hand

$$\boldsymbol{u}^I = \boldsymbol{u}^E + (\boldsymbol{u}^I - \boldsymbol{u}^E)$$

is an orthogonal decomposition so that the loss of information is given by (5.10). □

It should be noted that most semiparametric models so far treated by many researchers are m-flat. Therefore, the method of estimating functions produces no loss of information. However, there is still a serious problem of guessing a good z_0 from which we construct the estimating function $\boldsymbol{u}^I(x, \boldsymbol{\theta}, z_0)$ by fixing z_0. In spite of this, the point is that, even if we choose a wrong z_0, the estimator is still consistent although it is not fully efficient. The maximum likelihood estimator does not in general has this robust property.

6 Construction of estimating functions

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

When $\boldsymbol{u}^I(x, \boldsymbol{\theta}, z)$, $i = 1, 2, \cdots, n$, includes unknown z, one orthodox idea is first to find a consistent estimator $\hat{z}(x_1, \cdots, x_n, \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is fixed. We then have the optimal estimating function $\boldsymbol{u}^I(x, \boldsymbol{\theta}, \hat{z})$. However, this requires a formidable task and practically non-efficient. Moreover, there is a subtle problem in analyzing the estimating equation

$$\sum \boldsymbol{u}^I(x_i, \boldsymbol{\theta}, \hat{z}) = 0,$$

because $\boldsymbol{u}^I(x_i, \boldsymbol{\theta}, \hat{z})$ are not independent but are dependent through random variables $\hat{z}(x_1, \cdots, x_n, \boldsymbol{\theta})$. We do not touch upon this problem [see, for example, Bhanja and Ghosh(1992)].

It is the point of an estimating function that a misspecified z still gives a consistent estimator. Therefore, it is wise for practical purpose to choose a simple but good z. To this end, we propose to use a parameterized submodel of z,

$$M = \{z(\boldsymbol{\eta})\}, \qquad M \subset Z,$$

where $\boldsymbol{\eta}$ is a finite dimensional parameter to specify z. Since the true z is not necessarily included in the model M, we have an unfaithful statistical model

$$S^* = \{p(x, \boldsymbol{\theta}, \boldsymbol{\eta}) = p(x, \boldsymbol{\theta}, z(\boldsymbol{\eta})\}$$

parameterized by a finite number of parameters $(\boldsymbol{\theta}, \boldsymbol{\eta})$. It is not difficult to obtain an estimate $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\eta}})$, say the m.l.e. This $\tilde{\boldsymbol{\theta}}$ is not consistent in general. Instead of using this $\tilde{\boldsymbol{\theta}}$, use the estimating function

$$\boldsymbol{u}^I(x, \boldsymbol{\theta}, z(\tilde{\boldsymbol{\eta}}))$$

to obtain a good consistent estimator $\hat{\boldsymbol{\theta}}$. This type of idea was also used by Lindsay(1982).

The amount of loss of information by using a wrong z is given by the expectation of the square of the $F^A_{\boldsymbol{\theta},z}$-part of the m-parallel transport of $\boldsymbol{u}^I(z)$ from the true z_0 to z.

7 Examples

Example 1 Location-scale model

This is a nuisance m-flat model, because the probability density is linear in the nuisance function z. From

$$p(x,\mu,\sigma,z) = \frac{1}{\sigma} z\left(\frac{x-\mu}{\sigma}\right)\varphi\left(\frac{x-\mu}{\sigma}\right),$$

the score functions of the parameters of interest are

$$\begin{aligned} u_\mu &= \partial_\mu \log p = -\frac{1}{\sigma}\frac{z'}{z} + \frac{x-\mu}{\sigma^2}, \\ u_\sigma &= \partial_\sigma \log p = -\frac{x-\mu}{\sigma^2}\frac{z'}{z} + \frac{(x-\mu)^2}{\sigma^3} - \frac{1}{\sigma} \end{aligned}$$

where $\boldsymbol{\theta} = (\mu,\sigma)$, $\partial_\mu = \partial/\partial\mu$, $\partial_\sigma = \partial/\partial\sigma$. The tangent space $T_{\boldsymbol{\theta},z}$ is spanned by u_μ and u_σ.

We next construct a curve connecting $z(x)$ and $z(x)+r(x)$ by

$$z(x,t) = z(x) + tr(x), \tag{7.1}$$

where $z(x,0) = z(x)$. Then, the score function in the direction r is given by

$$\frac{d}{dt}\log p = \frac{r}{z}. \tag{7.2}$$

Here, we assume that Z consists of such functions that, for any $z_1, z_2 \in Z$,

$$\int (1-\frac{z_2}{z_1})^2 z_1 d\mu^*(x) < \infty. \tag{7.3}$$

This guarantees that $v = r/z$ belongs to $H_{\boldsymbol{\theta},z}$. From (2.14), $r(x)$ satisfies

$$\begin{aligned} \int r(x) d\mu^*(x) &= 0, \\ \int x r(x) d\mu^*(x) &= 0, \\ \int x^2 r(x) d\mu^*(x) &= 0. \end{aligned} \tag{7.4}$$

Let $H_i(x)$, $i = 0, 1, 2, \cdots$, be the Hermite polynomials which satisfy the orthonormality conditions

$$\int H_i(x) H_j(x) d\mu^*(x) = \delta_{ij}. \tag{7.5}$$

Since $H_i(x)$ is a polynomial in x of degree i, the conditions (7.4) show that any r can be expanded as

$$r(x) = \sum_{i=3}^{\infty} c_i H_i(x). \tag{7.6}$$

Therefore, the nuisance tangent space $T^N_{\boldsymbol{\theta},z}$ is spanned by $\{H_i/z, i \geq 3\}$. The ancillary space $T^A_{\boldsymbol{\theta},z}$ is composed of the other square integrable random variables $w(x)$ orthogonal to $T^N_{\boldsymbol{\theta},z}$ and $T^I_{\boldsymbol{\theta},z}$.

The subspace $F^I_{\boldsymbol{\theta},z} \oplus F^A_{\boldsymbol{\theta},z}$ is therefore spanned by $H_1(x)$ and $H_2(x)$, because these are orthogonal to (H_i/z), $i = 3, 4, \cdots$. Hence, we have the following estimating function,

$$\begin{aligned} y_1(x, \boldsymbol{\theta}) &= x - \mu, \\ y_2(x, \boldsymbol{\theta}) &= (x - \mu)^2 - \sigma^2. \end{aligned}$$

This is unique and optimal, since any other estimating functions are written as their linear combinations. It gives rather trivial estimators

$$\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum x_i, \\ \hat{\sigma} &= \frac{1}{n} \sum (x_i - \hat{\mu})^2 \end{aligned}$$

but they are the best in the semiparametric setting.

When other information is variable, we may have better estimators. For example, if z is known to be even function, $r(x)$ is also an even function. Hence, $r(x)$ is expanded as

$$r(x) = \sum_{i=2}^{\infty} c_{2i} H_{2i}(x), \quad n = 1, 2, \cdots. \tag{7.7}$$

Therefore, the $F^N_{\boldsymbol{\theta},z} = T^N_{\boldsymbol{\theta},z}$ is spanned by $\{H_i/z, i = 4, 6, 8, \cdots\}$, and $F^I_{\boldsymbol{\theta},z} \oplus F^I_{\boldsymbol{\theta},z}$ is spanned by $H_1, H_2, H_3, H_5, H_7, \cdots$. There are lots of candidates for estimating functions, and the best one is obtained by the projected score.

Example 2 Mixture model

These models have again the m-flat nuisance structure. The $\boldsymbol{\theta}$-score and the nuisance score in direction $a(\boldsymbol{\xi})$ of (2.15) and (2.16) can be calculated in a similar way but their expressions are rather complicated. For notational convenience, we define

$$L(\boldsymbol{s}; z) = \int \exp\{\boldsymbol{\xi} \cdot \boldsymbol{s}\}[z(\boldsymbol{\xi}) \exp\{-\psi(\boldsymbol{\xi}, \boldsymbol{\theta})\}]d\boldsymbol{\xi},$$

so that

$$p(x, \boldsymbol{\theta}, z) = L\{\boldsymbol{s}(x, \boldsymbol{\theta}); z\} \exp\{r(x, \boldsymbol{\theta})\}.$$

The $L(\boldsymbol{s}; z)$ is the Laplace transform of $z(\boldsymbol{\xi}) \exp\{-\psi(\boldsymbol{\xi}, \boldsymbol{\theta})\}$. By using this function, the $\boldsymbol{\theta}$-score is written as

$$\boldsymbol{u} = \frac{L(\boldsymbol{s}; \{\boldsymbol{\xi}\partial_{\boldsymbol{\theta}}\boldsymbol{s} + \partial_{\boldsymbol{\theta}} r - \partial_{\boldsymbol{\theta}}\psi\} z(\boldsymbol{\xi}))}{L(\boldsymbol{s}; z)},$$

where $\boldsymbol{s} = \boldsymbol{s}(x, \boldsymbol{\theta})$. The nuisance score is

$$v[a] = \frac{L(\boldsymbol{s}; a)}{L(\boldsymbol{s}; z)}.$$

Since a is an arbitrary function and $L(\boldsymbol{s}; a)$ is its Laplace transform, we can conclude that the nuisance subspace $T^N_{\boldsymbol{\theta},z}$ is the linear space generated by the random variable $\boldsymbol{s}(x, \boldsymbol{\theta})$.

It is known that the projection of a random variable t to the space generated by $\boldsymbol{s}$ is given by the conditional expectation $\mathrm{E}[t|\boldsymbol{s}]$ and the projection to the orthogonal complement is $t - \mathrm{E}[t|\boldsymbol{s}]$. Hence, the efficient score is given by

$$\begin{aligned} \boldsymbol{u}^E &= \boldsymbol{u} - \mathrm{E}[\boldsymbol{u}|\boldsymbol{s}] \\ &= \frac{1}{L(\boldsymbol{s}; z)} L[\boldsymbol{s}; \{\boldsymbol{\xi}(\partial_{\boldsymbol{\theta}}\boldsymbol{s} - \mathrm{E}[\partial_{\boldsymbol{\theta}}\boldsymbol{s}|\boldsymbol{s}]) - \partial_{\boldsymbol{\theta}}\psi\} z] \\ &\quad + \partial_{\boldsymbol{\theta}} r - \mathrm{E}[\partial_{\boldsymbol{\theta}} r|\boldsymbol{s}]. \end{aligned}$$

This gives the efficient estimating function. It is an interesting special case when $\partial_{\boldsymbol{\theta}}\boldsymbol{s}$ is a function of $\boldsymbol{s}$. In this case, $\partial_{\boldsymbol{\theta}}\boldsymbol{s} = \mathrm{E}[\partial_{\boldsymbol{\theta}}\boldsymbol{s}|\boldsymbol{s}]$. The efficient score is given in this case by

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

It does not depend on $z(\boldsymbol{\xi})$, so that the optimal estimation function exists in this case for any $z(\boldsymbol{\xi})$ and is given by the above $\boldsymbol{u}^E$. There is no information loss.

Example 3 Linear binary choice model
In this model, we assume that signals $\boldsymbol{x}$ are generated independently from the normal distribution with mean 0 and the unit covariance matrix,

$$\varphi(\boldsymbol{x}) = c \exp\left\{-\frac{1}{2}\boldsymbol{x}\cdot\boldsymbol{x}\right\}.$$

Then, the distribution is written as

$$p(\boldsymbol{x}, y, \boldsymbol{\theta}, z) = \varphi(\boldsymbol{x})\{0.5 + qz(\boldsymbol{\theta}\cdot\boldsymbol{x})\}.$$

The $\boldsymbol{\theta}$-score is

$$\boldsymbol{u} = \frac{q\boldsymbol{x}z'(\boldsymbol{\theta}\cdot\boldsymbol{x})}{0.5 + qz(\boldsymbol{\theta}\cdot\boldsymbol{x})}.$$

On the other hand, the nuisance-score in the direction of a is given by

$$v = \frac{qa(\boldsymbol{\theta}\cdot\boldsymbol{x})}{0.5 + qz(\boldsymbol{\theta}\cdot\boldsymbol{x})}$$

where

$$a(0) = 0.$$

The nuisance tangent space is spanned by these v vectors. Since $a(u)$ is an arbitrary smooth function, the nuisance tangent space is included in the σ-algebra generated by

$$s = \boldsymbol{\theta}\cdot\boldsymbol{x}$$

and q. Therefore, the projection of $\boldsymbol{u}$ to this space is given by the conditional expectation

$$\mathrm{E}[\boldsymbol{u}|s,q] = \mathrm{E}[\boldsymbol{x}|s]z'(s)\frac{q}{0.5 + qz(s)},$$

and it is included in $T^N_{\boldsymbol{\theta},z}$. We put

$$\mathbf{m}(\boldsymbol{\theta}\cdot\boldsymbol{x}) = \mathrm{E}[\boldsymbol{x}|\,\boldsymbol{\theta}\cdot\boldsymbol{x}\,].$$

When $\boldsymbol{x}$ is subject to $N(\mathbf{0}, I)$, it is easy to show that

$$\mathbf{m}(\boldsymbol{\theta} \cdot \boldsymbol{x}) = \mathrm{E}[\boldsymbol{x} | \, \boldsymbol{\theta} \cdot \boldsymbol{x} = s \,] = s \, \boldsymbol{\theta} = (\boldsymbol{\theta} \cdot \boldsymbol{x}) \, \boldsymbol{\theta}.$$

Then, the projected score is

$$\begin{aligned} \boldsymbol{u}^E &= \boldsymbol{u} - \mathrm{E}[\boldsymbol{u} | \, \boldsymbol{\theta} \cdot \boldsymbol{x}, \, q \,] \\ &= \frac{q\{\boldsymbol{x} - \mathbf{m}(\boldsymbol{\theta} \cdot \boldsymbol{x})\} z(\boldsymbol{\theta} \cdot \boldsymbol{x})}{0.5 + qz(\boldsymbol{\theta} \cdot \boldsymbol{x})}. \end{aligned}$$

This model is analyzed in detail in a forthcoming paper [Kawanabe, Amari and Hiroshige(1993)].

Example 4 Dose-response curve
The θ-score is given by

$$u = \frac{-qz(x - \theta)}{0.5 + qz(x - \theta)}.$$

The nuisance score along the direction a, is given by

$$v[a] = \frac{qa(x - \theta)}{0.5 + qz(x - \theta)}.$$

where

$$\begin{aligned} z(t) &= z(x) + ta(x), \\ a(0) &= 0, \end{aligned}$$

Since $z'(0) > 0$, u is not equal to any of $v[a]$ because $a(0) = 0$. However, u is included the closure $T^N_{\theta,z}$ of the space spanned by $v[a]$. Hence, the effective score is null, $u^E = 0$ and the effective Fisher information g^E is equal to 0. This implies that no $\sqrt{n}$-consistent estimator exist in this case. No estimating functions exist either.

Kawanabe, Amari and Hiroshige(1993) studied this problem by using asymptotic estimating functions, and obtained an $n^{p/(2p+1)}$-consistent estimator for any positive integer p.

Example 5

Not all the models are mixture m-flat. To show this, we give an artificial model of finite dimensions. It is a simple parametric model having a scalar parameter θ of interest and a scalar nuisance parameter ξ,

$$p(x,\theta,\xi) = (2\pi)^{-3/2} \exp\{-\frac{1}{2}[(x_1 - r_1)^2 + (x_2 - r_2)^2 + (x_3 - r_3)^2]\}$$

where

$$\begin{aligned} r_1 &= \theta, \\ r_2 &= \theta + \xi^2, \\ r_3 &= \xi. \end{aligned}$$

Since this is a finite-dimensional parametric model, it can easily be analyzed. The θ-score is given by

$$u = (x_1 - \theta) + (x_2 - \theta - \xi^2),$$

and the nuisance score (ξ-score) is given by

$$v = 2\xi(x_2 - \theta - \xi^2) + (x_3 - \xi).$$

From

$$\begin{aligned} \langle u, u \rangle &= 2, \qquad \langle v, v \rangle = 4\xi^2 + 1, \\ &\qquad \langle u, v \rangle = 2\xi, \end{aligned}$$

the projected score is

$$\begin{aligned} u^E &= u - \frac{\langle u, v \rangle}{|v|^2} v = u - \frac{2\xi}{4\xi^2 + 1} v \\ &= (x_1 - \theta) + \frac{1}{4\xi^2 + 1}(x_2 - \theta - \xi^2) - \frac{2\xi}{4\xi^2 + 1}(x_3 - \xi). \end{aligned}$$

Hence the effective Fisher information is

$$g^E = \langle u^E, u^E \rangle = 1 + \frac{1}{1 + 4\xi^2}.$$

When we fix θ, the submodel S_ξ is not m-flat since it is not linear in ξ. The m-transport of $v(\xi')$ to $v(\xi)$ is

$$\prod_{\xi'}^{(m)\,\xi} = \exp\{(\xi'^2 - \xi^2)x_2 + (\xi' - \xi)x_3 \\ +(\theta + \xi^2)^2 - (\theta + \xi'^2)^2 + \xi^2 - \xi'^2\},$$

The closed space spanned by all of these for any ξ' includes the random variables $x_2 - \theta - \xi^2$ and $x_3 - \xi$. Therefore, the information score u^I is different from the efficient score u^E and is given by

$$u^I = x_1 - \theta.$$

Its magnitude is

$$\langle u^I, u^I \rangle = 1 \leq g^E.$$

Therefore, in this simple case, if we can guess the value of ξ accurate enough, the m.l. equation

$$\sum u^E(x_i, \theta, \hat{\xi}(\theta)) = 0$$

gives an efficient estimator. However, if we use wrong ξ,

$$\sum u^E(x_i, \theta, \xi) = 0$$

does not give a consistent estimator. The best estimating function is $u^I = x_i - \theta$, giving the estimating equation

$$\sum_{i=1}^{n} (x_{i_1} - \theta) = 0$$

or

$$\hat{\theta} = \frac{1}{n} \sum x_{1i}.$$

This includes a loss of information, but is free of the precise estimation of ξ. In this finite-dimensional case, the estimation of ξ is easy. However, the point is that the estimating function method always gives a consistent estimator without any precise estimation of the nuisance function.

8 Conclusions

The present paper showed that the geometrical theory of estimating functions [Amari and Kumon(1988)] can be applicable to general semiparametric models which have nuisance parameters of functional degrees of freedom. By using the geometrical concept of Hilbert fibre space, Riemannian metric and dual parallel transports, the condition for existence of estimating function was derived, the space of estimating functions was specified, and the efficiency of the estimating function method was discussed. When a statistical model has m-flat nuisance structure, the effective score function itself is an estimating function, and estimating function method combined with the adaptive method gives a fully efficient estimator. This is the case with many important semiparametric models. For some of them the effective scores are calculated explicitly in the present paper.

Although the theory was established in a rather informal way manner, it is also necessary to study these problems further by functional analysis combined with the differential geometry of function space. The reason is not only that we need to make the theory mathematically rigorous, but also that the effective score functions, the space of estimating functions, etc. depend so much on the functional space of the nuisance parameter. Therefore it is important to study the relation between the structure of estimating functions and that of the space of the nuisance parameter, by applying the geometrical theory to many semiparametric models. It may be interesting to consider the case that some kind of symmetric or invariant restrictions are imposed on the space of the nuisance functions.

References

Amari, S. *Differential-Geometrical Method in Statistics*, volume 28 of *Lecture Note in Statistics*. Springer-Verlag, New York, 1985.

Amari, S. Differential geometry of a parametric family of invertible linear systems – riemannian metric, dual affine connections and divergence.

Mathematical Systems Theory, 20:53 – 82, 1987.

Amari, S. Dual connections on the hilbert bundles of statistical models. In Dodson, C.T.J., editor, *Geometrization of Statistical Theory*, pages 123 – 152. ULDM Lancaster UK., 1987.

Amari, S. and Han, T. S. Statistical inference under multi-terminal rate restrictions. *IEEE Trans. on Information Theory*, IT - 35:217 – 227, 1989.

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

Amari, S., Barndorff-Nielsen, O. E. and Labouriau, R. On the efficient score function for some semiparametric location-scale and regression models. 1993. to appear.

Amari, S., Kurata, K. and Nagaoka, H. Information geometry of boltzmann machines. *IEEE Trans. on Neural Networks*, 3:260 – 271, 1992.

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

Begun, J. M., Hall, W. J., Huang, W. M., and Wellner, J. A. Information and asymptotoic efficiency in parametric-nonparametric models. *Ann. Statist.*, 11:432 – 452, 1983.

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

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. *Efficient and Adaptive Estimation for Semiparametric Models.* Forthcoming monograph. Johns Hopkins University Press, Baltimore, 1992.

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

Godambe, V. P., editor. *Estimating Functions*, New York, 1991. Oxford University Press.

Godambe, V. P. and Heyde, C. C. Quasi-likelihood and optimal estimation. *Int. Stat. Rev.*, 55:231 – 244, 1987.

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

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

Kawanabe, M., Amari, S. and Hiroshige, T. Asymptotic estimating functions for semiparametric dose-response model. 1993. submitted to Biometrika.

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

Levit, B. Ya. Infinite-dimensional information inequalities. *Theory of Prob. Appl.*, 23:371 – 377, 1978.

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

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

Manski, C.F. Semiparametric analysis of discrete response. *Jounal of Econometrics*, 27:313 – 333, 1985.

McLeish, D. L. and Small, C. G. *The theory and applications of statistical inference functions*, volume 44 of *Lecture Notes in Statistics.* Springer Verlag, New York, 1988.

Nagaoka, H. and Amari, S. Differntial geometry of smooth families of probability distributions. Technical Report 82 - 7, Univ. Tokyo, 1982.

K. Nawata. Semiparametric estimation and efficiency bounds of binary choice models when the models contain one continuous variable. *Economics Letters*, 31:21–26, 1989.

Neyman, J. and Scott, E. L. Consistent estimates based on partially consitent obsevations. *Econometrica*, 32:1 – 32, 1948.

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

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

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

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

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