Institute of Mathematical Statistics

LECTURE NOTES — MONOGRAPH SERIES

ESTIMATING FUNCTIONS IN SEMIPARAMETRIC STATISTICAL MODELS

S. Amari
The Institute of Physical and Chemical Research
Saitama, Japan

M. Kawanabe
University of Tokyo

ABSTRACT

The geometrical structure of estimating functions is elucidated by information geometry in the framework of semiparametric statistical models. A condition which guarantees the existence of an estimating function is given. Moreover, the set of all the estimating functions is obtained explicitly when it is not null. The optimal estimating function is derived, and the maximum Godambe information is explicitly given. A geometrical condition is given which guarantees that the Godambe information is the maximal available information.

Key Words: Semiparametric models; estimating functions; dual parallel transport; Godambe information; m-curvature free; efficient score function.

1 Introduction

Godambe (1960, 1976) proposed the estimating function method as a generalization of the maximum likelihood method for parameter estimation. An estimating function gives a $\sqrt{n}$-consistent estimator by a simple and tractable procedure under certain regularity conditions. Moreover, the method is applicable even to semiparametric models. However, the class of estimators derived from estimating functions might not necessarily include the Fisher efficient estimator. Therefore, it is important to study efficiency of estimators derived from the estimating function method. It is another important problem to know how we can obtain the optimal estimating function. Recently researches on estimating functions have been developed and been applied to semiparametric models. It has been naturally understood that the optimal estimating function is given by projecting the score function to the linear space consisting of all the estimating functions (Small and McLeish (1989), Waterman and Lindsay (1996), Durairajan (1996), Chan and Ghosh (1996), Li (1996)). However, there still remain many important problems to be studied further. They are, for example, as follows :

1. To obtain a condition which guarantees the existence of estimating functions.
2. To obtain the linear spaces of all the estimating functions explicitly.
3. To obtain the amount of information, called the Godambe information, included in the optimal estimating function.
4. To obtain a condition which guarantees that the Godambe information is full, that is, equal to the maximal information in the general sense.

The present paper studies these problems by using the Hilbert bundle formalism of information geometry in the semiparametric context (Amari (1985), Amari and Kumon (1988)). This is a simplified version of the paper (Amari and Kawanabe (1996)), but this includes some new developments and a new example.

2 Estimating Functions in Semiparametric Models

Let $p(x, \boldsymbol{\theta}, \varphi)$ be a probability density function of a random variable x with respect to a common dominating measure $\mu(dx)$, specified by two kinds of parameters $\boldsymbol{\theta} = (\theta^1, \cdots, \theta^m)$ and φ, where $\boldsymbol{\theta} \in \Theta$ is a finite-dimensional vector, Θ is an open set of $\boldsymbol{R}^m$ and $\varphi \in \Phi$ is a finite or an infinite dimensional parameter, typically living in a space of functions. The set of distributions $S = \{p(x, \boldsymbol{\theta}, \varphi)\}$ is called a statistical model with a nuisance parameter and is called in particular a semiparametric statistical model when Φ is infinite-dimensional. Here $\boldsymbol{\theta}$ is called the parameter of interest and φ is called the nuisance parameter.

Let $\boldsymbol{y}(x, \boldsymbol{\theta}) = [y_i(x, \boldsymbol{\theta})]$, $i = 1, \cdots, m$, be a vector-valued smooth function of $\boldsymbol{\theta}$, not depending on φ, of the same dimension as $\boldsymbol{\theta}$. Such a function is called an estimating function (Godambe (1976, 1991)), when it satisfies the following conditions,

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

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

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

for all $\boldsymbol{\theta}$ and φ, where $\mathrm{E}_{\boldsymbol{\theta},\varphi}$ denotes the expectation with respect to the distribution $p(x, \boldsymbol{\theta}, \varphi)$, $\partial_{\boldsymbol{\theta}} \boldsymbol{y}$ 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 $|\quad|$ denotes the determinant of a matrix, and $\|\boldsymbol{y}\|^2$ is the squared norm of the vector $\boldsymbol{y}$, $\|\boldsymbol{y}\|^2 = \sum(y_i)^2$. We further need that $\int \boldsymbol{y}\, p\, \mathrm{d}\mu$ is differentiable with respect to $\boldsymbol{\theta}$ and that integration and differentiation are interchangeable. The condition (2.1) is called the unbiasedness condition. When the above conditions (2.1) – (2.3) hold in a neighborhood $N(\varphi_0)$ of φ_0, such $\boldsymbol{y}(x,\boldsymbol{\theta})$ is called a local estimating function at φ_0.

When an estimating function $\boldsymbol{y}(x,\boldsymbol{\theta})$ exists, by replacing the expectation in (2.1) by the empirical sum, we have an estimator $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ by solving the estimating equation

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

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. It might be thought that the additive form (2.4) is too restrictive for obtaining good estimators. However, we shall prove that the Fisher efficient estimator is included in this class in the m-curvature free models.

The asymptotic behavior of an M-estimator $\widehat{\boldsymbol{\theta}}$ is known by the following theorem (Godambe (1976), McLeish and Small (1988) for example).

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

$$\mathrm{AV}[\widehat{\boldsymbol{\theta}};\boldsymbol{y}] = A^{-1}\mathrm{E}_{\boldsymbol{\theta},\varphi}[\boldsymbol{y}\boldsymbol{y}^{\mathrm{T}}](A^{\mathrm{T}})^{-1}, \tag{2.5}$$

where the asymptotic covariance matrix is defined by

$$\mathrm{AV}[\widehat{\boldsymbol{\theta}};\boldsymbol{y}] = \lim_{n\to\infty} n\mathrm{E}_{\boldsymbol{\theta},\varphi}[(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{\mathrm{T}}], \tag{2.6}$$

A is the matrix defined by

$$A = \mathrm{E}_{\boldsymbol{\theta},\varphi}[\partial_{\boldsymbol{\theta}}\boldsymbol{y}(x,\boldsymbol{\theta})],$$

and the superfix T *denotes the transposition of a vector or a matrix.*

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

The present paper aims at obtaining the minimum value of $\mathrm{AV}[\widehat{\boldsymbol{\theta}};\boldsymbol{y}]$. Before that, we give two examples of semiparametric statistical models.

1. Neyman-Scott problem and mixture models :

   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 $q(x_i,\boldsymbol{\theta},\boldsymbol{\xi}_i)$, where $\boldsymbol{\theta}$ is common but $\boldsymbol{\xi}_i$ takes a different value at each observation. Then, estimating $\boldsymbol{\theta}$ from observations $\boldsymbol{x}=(x_1,\cdots,x_n)$ is called the Neyman-Scott problem, where the underlying probability distribution

$$q(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\xi}_1,\cdots,\boldsymbol{\xi}_n)=\prod_{i=1}^{n} q(x_i,\boldsymbol{\theta},\boldsymbol{\xi}_i)$$

   includes the nuisance parameters $\boldsymbol{\xi}_1,\cdots,\boldsymbol{\xi}_n$ as large as the number of observations. This problem can be treated by the following semiparametric model. Let us assume that the unknown $\boldsymbol{\xi}_i$ are independently generated subject to a common but unknown probability distribution having a density function $\varphi(\boldsymbol{\xi})$. Then, the x_i are regarded as independent observations from the semiparametric model

$$p(x,\boldsymbol{\theta},\varphi)=\int q(x,\boldsymbol{\theta},\boldsymbol{\xi})\varphi(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}, \tag{2.7}$$

   where $\varphi(\boldsymbol{\xi})$ is the nuisance parameter of function-degrees of freedom. This model is called the mixture model.

   This type of problems was studied by Neyman and Scott (1948) and has attracted many researchers (Andersen (1970), Lindsay (1982), Kumon and Amari (1984), Amari and Kumon (1988), Pfanzagl (1990) etc.). There are a lot of interesting and important examples in this class. A typical example is the following class of distributions of the form,

$$q(x,\boldsymbol{\theta},\boldsymbol{\xi})=\exp\{\boldsymbol{\xi}\cdot\boldsymbol{s}(x,\boldsymbol{\theta})+r(x,\boldsymbol{\theta})-\psi(\boldsymbol{\theta},\boldsymbol{\xi})\}, \tag{2.8}$$

   where $\boldsymbol{s}(x,\boldsymbol{\theta})$ 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 $\boldsymbol{\theta}$ is fixed.

2. Blind separation of mixture signals :

   Let S_a, $a=1,2,\cdots,r$, be r signal sources which produce r time serieses $s_a(t)$, $t=1,2,\ldots$. We assume that each $s_a(t)$ is an ergodic time series having the probability density $q_a(s_a)$ at any t. Moreover, $s_1,\cdots,s_r$ are assumed to be independent. Then, their joint probability is written as

$$\varphi(\boldsymbol{s})=\prod_{a=1}^{r} q_a(s_a) \tag{2.9}$$

   at any t where $\boldsymbol{s}=(s_1,\cdots,s_r)$.

   We assume that we cannot directly observe the r signals $s_a(t)$ but we can observe their mixtures,

$$x_i(t) = \sum_{a=1}^{r} M_a^i s_a(t), \qquad i = 1, \cdots, r, \tag{2.10}$$

where $M = (M_a^i)$ is an $r \times r$ non-singular matrix consisting of fixed mixing coefficients M_a^i. Then, the joint probability density function of $\boldsymbol{x}^{\mathrm{T}} = (x_1, \cdots, x_r)$ is given by

$$p(\boldsymbol{x}) = |W| \varphi(W\boldsymbol{x}), \tag{2.11}$$

where

$$W = M^{-1}. \tag{2.12}$$

If we know M or W, the original source signals $\boldsymbol{s}(t)$ are recovered from the observed $\boldsymbol{x}(t)$ by

$$\boldsymbol{s}(t) = W\boldsymbol{x}(t). \tag{2.13}$$

When we do not know M or W, we should estimate W from the observed $\boldsymbol{x}(t)$, $t = 1, 2, \cdots$, where the density functions $q_1(s_1), \cdots, q_r(s_r)$ are usually unknown. Such a problem often occurs in medical or communication signal processing, and is called the blind separation of sources. See Amari et al. (1996).

This gives a typical semiparametric statistical model,

$$p(\boldsymbol{x}, W, \varphi) = |W| \varphi(W\boldsymbol{x}), \tag{2.14}$$

where W is the parameter of interest and

$$\varphi(\boldsymbol{s}) = \prod_{a=1}^{r} q_a(s_a)$$

is the nuisance functions.

3 Hilbert Tangent Spaces and Score Functions

Given a probability density function $p(x)$, let us consider a one-parameter statistical model

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

where t $(0 \le t < \varepsilon)$ is the parameter. The constraint

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

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

$$\int p(x,t)\mathrm{d}\mu(x) \;=\; \int p(x)\{1+ta(x)\}\,\mathrm{d}\mu(x) \;=\; 1. \qquad (3.2)$$

When t is small, $p(x,t)$ is a small deviation in the direction of $a(x)$ from $p(x)$. The model (3.1) is a curve parameterized by t in the set of all the probability density functions. Let us consider the linear space of functions $a(x)$ which satisfy

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

The set of all such $a(x)$ is a Hilbert space H_p with the inner product of $a(x)$ and $b(x)$ defined by

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

The Hilbert space H_p consists of all the deviations $a(x)$ of probability distribution from $p(x)$.

The random variable

$$a(x) = \left.\frac{\mathrm{d}}{\mathrm{d}t}\log p(x,t)\right|_{t=0} \qquad (3.5)$$

is the tangent vector of the curve (3.1) at $p(x)$. This is the score function for the one-dimensional statistical model (3.1) parameterized by t.

Given a semiparametric model $S = \{p(x, \boldsymbol{\theta}, \varphi)\}$, we construct the Hilbert space $H_{\boldsymbol{\theta},\varphi}$ denoting the set of all the deviations from $p(x) = p(x, \boldsymbol{\theta}, \varphi)$. Since we have interest in estimating functions, we define it by

$$H_{\boldsymbol{\theta},\varphi} = \left\{a(x) \left| \mathrm{E}_{\boldsymbol{\theta},\varphi}[a(x)] = 0,\ \mathrm{E}_{\boldsymbol{\theta},\varphi'}\left[\{a(x)\}^2\right] < \infty \text{ for all } \varphi' \right.\right\}$$

restricting the space such that it consists of functions $a(x)$ which are square integrable at all $p(x, \boldsymbol{\theta}, \varphi')$ even when it is defined at $(\boldsymbol{\theta}, \varphi)$.

The tangent directions along the parameter of interest are the score functions

$$u_i(x, \boldsymbol{\theta}, \varphi) = \frac{\partial}{\partial\theta^i}\log p(x, \boldsymbol{\theta}, \varphi). \qquad (3.6)$$

Obviously,

$$\mathrm{E}_{\boldsymbol{\theta},\varphi}[u_i] = 0 \qquad (3.7)$$

and we further assume that u_i is square-integrable at any $(\boldsymbol{\theta}, \varphi')$. Then it belongs to $H_{\boldsymbol{\theta},\varphi}$. We call the subspace spanned by these u_i's the tangent subspace $T_{\boldsymbol{\theta},\varphi}$ along the parameter of interest. The vector score function is $\boldsymbol{u} = (u_1, \cdots, u_m)$.

We next define the tangent directions along the nuisance parameter. Let us consider a curve $c(t)$ connecting functions φ and φ' such that $c(0) = \varphi$ and

$c(t_0) = \varphi'$. We then have the one-dimensional statistical model $p\{x, \boldsymbol{\theta}, c(t)\}$ parameterized by t. Its score function is given by

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

which we assume to belong to the $H_{\boldsymbol{\theta},\varphi}$. This v is the tangent vector along $c(t)$ of the nuisance parameter. There are infinitely many curves $c(t)$ and the corresponding v's, when φ is a function. Let $T^N_{\boldsymbol{\theta},\varphi}$ be the smallest closed subspace including all such v's. We call it the nuisance tangent space. This is a closed subspace of $H_{\boldsymbol{\theta},\varphi}$.

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

Let $T^A_{\boldsymbol{\theta},\varphi}$ be the orthogonal complement of $T^N_{\boldsymbol{\theta},\varphi} \oplus T^E_{\boldsymbol{\theta},\varphi}$. It 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 space (see Amari (1987), Amari and Kumon (1988), see also Small and McLeish (1988)),

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

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

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

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 $\widehat{\boldsymbol{\theta}}$,

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

for any asymptotically normally distributed unbiased estimators in a semiparametric model. There is, however, no guarantee that this bound is asymptotically attainable by choosing an estimating function even when φ is finite-dimensional. (This bound is attainable when φ is finite dimensional by taking the joint m.l.e. $(\widehat{\boldsymbol{\theta}}, \widehat{\varphi})$ of $\boldsymbol{\theta}$ and φ.) So we need to search for a new bound, called the Godambe information bound, attainable by an estimating function explicitly.

4 Global Decomposition of Hilbert Spaces

Let us temporarily fix a φ_0. When φ_0 is the true nuisance parameter,

$$\sum_{i=1}^{n} \boldsymbol{u}^E(x_i, \boldsymbol{\theta}, \varphi_0) = 0$$

gives a good estimator. However, $\boldsymbol{u}^E(x_i, \boldsymbol{\theta}, \varphi_0)$ is not an estimating function in general, because it does not satisfy the condition

$$\mathrm{E}_{\boldsymbol{\theta},\varphi}[\boldsymbol{u}^E(x_i, \boldsymbol{\theta}, \varphi_0)] = 0.$$

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

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

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

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

are called the e-parallel transport and the m-parallel transport of $a(x)$ from $(\boldsymbol{\theta}, \varphi)$ to $(\boldsymbol{\theta}, \varphi')$, respectively.

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 do not go into mathematical details of differential geometry.

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

Lemma 1 *The two parallel transports are dual in the sense that, for any two $a(x), b(x) \in H_{\boldsymbol{\theta},\varphi}$, the inner product is kept invariant when one is e-transported and the other is m-transported to $H_{\boldsymbol{\theta},\varphi'}$,*

$$\langle a, b \rangle_{\boldsymbol{\theta},\varphi} = \left\langle \overset{(e)}{\prod}{}_{\varphi}^{\varphi'} a, \overset{(m)}{\prod}{}_{\varphi}^{\varphi'} b \right\rangle_{\boldsymbol{\theta},\varphi'}, \tag{4.3}$$

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

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

$$\overset{(e)}{\prod}{}_{\varphi}^{\varphi'} \boldsymbol{y}(x,\boldsymbol{\theta}) = \boldsymbol{y}(x,\boldsymbol{\theta})$$

because of (2.1) where $\overset{(e)}{\prod}$ operates componentwise. Now we rewrite the unbiased condition by using the parallel transport. Let us consider a curve $\varphi = \varphi(t)$, $\varphi_0 = \varphi(0)$, in the nuisance space. By differentiating (2.1) with respect to t along the curve $\varphi = \varphi(t)$, we have

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \int p\{x,\boldsymbol{\theta},\varphi(t)\}\boldsymbol{y}(x,\boldsymbol{\theta})\,\mathrm{d}\mu(x)\Big|_{t=0} & \\ = & \int v\{x,\boldsymbol{\theta},\varphi_0\}p\{x,\boldsymbol{\theta},\varphi_0\}\boldsymbol{y}(x,\boldsymbol{\theta})\,\mathrm{d}\mu(x) \\ = & \langle v, \boldsymbol{y}(x,\boldsymbol{\theta})\rangle_{\boldsymbol{\theta},\varphi_0} = 0 \end{aligned}$$

where

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

is the nuisance tangent direction at φ_0 along the curve $\varphi(t)$. This holds for any φ_0 so that any estimating function $\boldsymbol{y}(x,\boldsymbol{\theta})$ is orthogonal to $v(\boldsymbol{\theta},\varphi)$ at any point $(\boldsymbol{\theta},\varphi)$. However, from

$$\begin{aligned} \langle v, \boldsymbol{y}(x,\boldsymbol{\theta})\rangle_{\boldsymbol{\theta},\varphi'} &= \left\langle \overset{(m)}{\prod}{}_{\varphi'}^{\varphi} v,\ \overset{(e)}{\prod}{}_{\varphi'}^{\varphi} \boldsymbol{y} \right\rangle_{\boldsymbol{\theta},\varphi} \\ &= \left\langle \overset{(m)}{\prod}{}_{\varphi'}^{\varphi} v,\ \boldsymbol{y} \right\rangle_{\boldsymbol{\theta},\varphi}, \qquad (4.4) \end{aligned}$$

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

To incorporate with this global structure, we define the enlarged nuisance tangent space $F^N_{\boldsymbol{\theta},\varphi}$ by

$$F^N_{\boldsymbol{\theta},\varphi} = \operatorname{span}\left\{ \overset{(m)}{\prod}{}_{\varphi'}^{\varphi} T^N_{\boldsymbol{\theta},\varphi'} \text{ for all } \varphi' \in \Phi \right\},$$

that is, $F^N_{\boldsymbol{\theta},\varphi}$ is the subspace of $H_{\boldsymbol{\theta},\varphi}$ spanned by the m-parallel transports of $T^N_{\boldsymbol{\theta},\varphi'}$ from φ' to φ for all φ'. It might occur that the m-parallel transport of $a(x) \in T^N_{\boldsymbol{\theta},\varphi'}$ does not belong to $H_{\boldsymbol{\theta},\varphi}$ because of

$$\mathrm{E}_{\boldsymbol{\theta},\varphi}\left[\left\{\overset{(m)}{\prod}{}^{\varphi}_{\varphi'} a(x)\right\}^2\right] = \infty.$$

In this case, just ignore it.

We next project $T_{\boldsymbol{\theta},\varphi}$ or $T^E_{\boldsymbol{\theta},\varphi}$ to the subspace orthogonal to $F^N_{\boldsymbol{\theta},\varphi}$. The resultant subspace is called the information subspace and is denoted by $F^I_{\boldsymbol{\theta},\varphi}$. The subspace orthogonal to $F^N_{\boldsymbol{\theta},\varphi}$ and $F^I_{\boldsymbol{\theta},\varphi}$ is called the shrinked ancillary space $F^A_{\boldsymbol{\theta},\varphi}$. We then have the following orthogonal decomposition of $H_{\boldsymbol{\theta},\varphi}$:

$$H_{\boldsymbol{\theta},\varphi} = F^I_{\boldsymbol{\theta},\varphi} \oplus F^N_{\boldsymbol{\theta},\varphi} \oplus F^A_{\boldsymbol{\theta},\varphi}. \tag{4.5}$$

Obviously,

$$T^N_{\boldsymbol{\theta},\varphi} \subset F^N_{\boldsymbol{\theta},\varphi}, \qquad T^A_{\boldsymbol{\theta},\varphi} \supset F^A_{\boldsymbol{\theta},\varphi}.$$

5 Estimating Functions and Godambe Information

The decomposition (4.5) of the Hilbert space $H_{\boldsymbol{\theta},\varphi}$ makes it possible to characterize the set of all the estimating functions. We first show an important lemma (see Amari and Kawanabe (1996) for the details of the proof).

Lemma 2 *A necessary and sufficient condition for a function $w(x,\boldsymbol{\theta})$ to be e-invariant is that it belongs to $F^I_{\boldsymbol{\theta},\varphi} \oplus F^A_{\boldsymbol{\theta},\varphi}$ for some φ.*

Proof When $w(x,\boldsymbol{\theta})$ is e-invariant, we have

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

We have already shown that w belongs to $F^I_{\boldsymbol{\theta},\varphi} \oplus F^A_{\boldsymbol{\theta},\varphi}$ for any φ in this case. On the other hand, let $w(x,\boldsymbol{\theta})$ be a function belonging to $F^I_{\boldsymbol{\theta},\varphi_0} \oplus F^A_{\boldsymbol{\theta},\varphi_0}$. In order to show that it is e-invariant, we consider a path $\varphi = \varphi(t)$ in the nuisance space and put

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

Obviously, $f(0) = 0$ where $\varphi(0) = \varphi_0$. By differentiating this with respect to t, we can prove that

$$\frac{\mathrm{d}}{\mathrm{d}t} f(t) = 0$$

for any t (see Amari and Kawanabe (1996)), showing that $f(t) = 0$ for any t. It is also proved that, when w belongs to $F^I_{\boldsymbol{\theta},\varphi} \oplus F^A_{\boldsymbol{\theta},\varphi}$ for a φ, it automatically belongs to $F^I_{\boldsymbol{\theta},\varphi'} \oplus F^A_{\boldsymbol{\theta},\varphi'}$ at all φ'. □

Lemma 3 *Let $\boldsymbol{y}(x, \boldsymbol{\theta})$ be an estimating function and let $y_i^I(x, \boldsymbol{\theta})$ be the projection of the i-th component of $\boldsymbol{y}(x, \boldsymbol{\theta})$ to $F^I_{\boldsymbol{\theta},\varphi}$. Then $y_i^I(x, \boldsymbol{\theta})$, $i = 1, \cdots, m$, span $F^I_{\boldsymbol{\theta},\varphi}$ at any φ.*

Proof By differentiating (2.1) with respect to $\boldsymbol{\theta}$, we have

$$\mathrm{E}_{\boldsymbol{\theta},\varphi}[\partial_{\boldsymbol{\theta}} \boldsymbol{y}(x, \boldsymbol{\theta})] + \langle \boldsymbol{u}, \boldsymbol{y}(x, \boldsymbol{\theta}) \rangle_{\boldsymbol{\theta},\varphi} = 0$$

at all φ, where $\langle \boldsymbol{u}, \boldsymbol{y} \rangle$ is a matrix whose elements are $\langle u_i, y_j \rangle$. From this, we have

$$\mathrm{E}_{\boldsymbol{\theta},\varphi}[\partial_{\boldsymbol{\theta}} \boldsymbol{y}(x, \boldsymbol{\theta})] = -\langle \boldsymbol{u}, \boldsymbol{y} \rangle = -\langle \boldsymbol{u}^I, \boldsymbol{y} \rangle,$$

where the components u_i^I of the information score $\boldsymbol{u}^I$ are the projections of the components of the score $\boldsymbol{u}$ to $F^I_{\boldsymbol{\theta},\varphi}$. When $\boldsymbol{y}$ is an estimating function, this is non-degenerate from (2.2), proving the lemma. □

Combining the above lemmas, we have the following fundamental theorem, which gives the set of all estimating functions.

Theorem 2 *Any estimating function $\boldsymbol{y}(x, \boldsymbol{\theta}) = \{y_i(x, \boldsymbol{\theta})\}$ can be decomposed at any φ as a sum*

$$\boldsymbol{y}(x, \boldsymbol{\theta}) = T(\boldsymbol{\theta}, \varphi)\boldsymbol{u}^I(x, \boldsymbol{\theta}, \varphi) + \boldsymbol{a}(x, \boldsymbol{\theta}, \varphi), \tag{5.1}$$

where the component $a_i(x, \boldsymbol{\theta}, \varphi)$ of $\boldsymbol{a}$ belongs to $F^A_{\boldsymbol{\theta},\varphi}$ and $T(\boldsymbol{\theta}, \varphi)$ is a non-singular matrix. Conversely, any function $\boldsymbol{y}(x, \boldsymbol{\theta})$ defined in the form of (5.1) at a fixed φ_0 gives an estimating function provided the projections of the components $y_i(x, \boldsymbol{\theta})$ to $F^I_{\boldsymbol{\theta},\varphi'}$ span $F^I_{\boldsymbol{\theta},\varphi'}$ at every φ'.

It is possible to choose a basis for the information scores such that $T(\boldsymbol{\theta}, \varphi)$ becomes the identity at a φ. The theorem also shows a condition for the existence of an estimating function.

Theorem 3 *A local estimating function at φ_0 exists when and only when $F^I_{\boldsymbol{\theta},\varphi_0}$ is non-degenerate, that is, m-dimensional. A necessary condition for the existence of a global estimating function is that $F^I_{\boldsymbol{\theta},\varphi}$ is non-degenerate at all φ.*

Proof The necessary condition follows immediately from (2.2) and (5.1). When $F^I_{\boldsymbol{\theta},\varphi_0}$ is non-degenerate, $\boldsymbol{u}^I(x, \boldsymbol{\theta}, \varphi_0)$ is a local estimating function in a neighborhood of φ_0. □

Remark When we treat local estimating functions, the definition of $F^N_{\boldsymbol{\theta},\varphi}$ and hence the decomposition (4.5) should be defined locally.

We now derive the optimal estimating function and the amount of information (Godambe information) derived thereby. Let us define

$$G^I(\boldsymbol{\theta},\varphi) = \mathrm{E}_{\boldsymbol{\theta},\varphi}[\boldsymbol{u}^I(x,\boldsymbol{\theta},\varphi)\{\boldsymbol{u}^I(x,\boldsymbol{\theta},\varphi)\}^{\mathrm{T}}]. \tag{5.2}$$

Given $\boldsymbol{y}$, we also define

$$G^A(\boldsymbol{\theta},\varphi;\boldsymbol{a}) = \mathrm{E}_{\boldsymbol{\theta},\varphi}[\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}], \tag{5.3}$$

where any estimating function can be decomposed as

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

where $\boldsymbol{a} \in F^A_{\boldsymbol{\theta},\varphi}$. It is immediate to show

$$\begin{aligned}\mathrm{E}_{\boldsymbol{\theta},\varphi}[\partial_{\boldsymbol{\theta}}\boldsymbol{a}] &= -\langle \boldsymbol{u}, \boldsymbol{a}^{\mathrm{T}}\rangle = 0,\\ \mathrm{E}_{\boldsymbol{\theta},\varphi}[\partial_{\boldsymbol{\theta}}\boldsymbol{y}] &= -\langle \boldsymbol{u}^I, (\boldsymbol{u}^I)^{\mathrm{T}}\rangle.\end{aligned}$$

Hence, we have the following result.

Theorem 4 *The asymptotic covariance matrix derived from an estimating function is*

$$\mathrm{AV}[\widehat{\boldsymbol{\theta}};\boldsymbol{y}] = (G^I)^{-1} + (G^I)^{-1}G^A(G^I)^{-1}. \tag{5.4}$$

The estimating function $\boldsymbol{u}^I(x,\boldsymbol{\theta},\varphi_0)$ where φ_0 is fixed is the optimal estimating function at φ_0 and the Godambe information is given by G^I.

6 Curvature-freeness

The information score $\boldsymbol{u}^I$ is different from the efficient score $\boldsymbol{u}^E$ in general, and

$$G^E \geq G^I. \tag{6.1}$$

The quantity

$$G^E - G^I = \mathrm{E}_{\boldsymbol{\theta},\varphi}[(\boldsymbol{u}^E - \boldsymbol{u}^I)(\boldsymbol{u}^E - \boldsymbol{u}^I)^{\mathrm{T}}] \tag{6.2}$$

is the loss of information caused by using estimating functions. However, in many cases, $\boldsymbol{u}^E = \boldsymbol{u}^I$ and $G^E = G^I$ for any $\boldsymbol{\theta}$ and φ. In this case, the estimating function method is fully efficient, if the optimal $\boldsymbol{y}$ is chosen. When does this happen?

To answer this question, we consider the statistical submodel $S_{\boldsymbol{\theta}} = \{p(x,\boldsymbol{\theta},\varphi)\}$ by fixing $\boldsymbol{\theta}$ where $\varphi \in \Phi$ is the only free parameter. The tangent vectors of $S_{\boldsymbol{\theta}}$ compose the nuisance tangent space $T^N_{\boldsymbol{\theta},\varphi}$. Let us consider the m-parallel transports of $T^N_{\boldsymbol{\theta},\varphi'}$ from $(\boldsymbol{\theta},\varphi')$ to $(\boldsymbol{\theta},\varphi)$ and see how it is different from $T^N_{\boldsymbol{\theta},\varphi}$. 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 $T^N_{\boldsymbol{\theta},\varphi}$ and $T^N_{\boldsymbol{\theta},\varphi'}$ 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 semiparametric statistical model S is said to be m-flat or m-convex, when the $T^N_{\boldsymbol{\theta},\varphi'}$ are invariant under the m-parallel transports, that is,*

$$\overset{(m)}{\prod}{}^{\varphi}_{\varphi'} T^N_{\boldsymbol{\theta},\varphi'} \subset T^N_{\boldsymbol{\theta},\varphi} \tag{6.3}$$

for any φ, φ' and $\boldsymbol{\theta}$. When the m-parallel transports of $T^N_{\boldsymbol{\theta},\varphi'}$ from $(\boldsymbol{\theta},\varphi')$ to $(\boldsymbol{\theta},\varphi)$ does not include the $T^E_{\boldsymbol{\theta},\varphi}$ components for any φ, φ' and $\boldsymbol{\theta}$, that is,

$$\overset{(m)}{\prod}{}^{\varphi}_{\varphi'} T^N_{\boldsymbol{\theta},\varphi'} \subset T^N_{\boldsymbol{\theta},\varphi} \oplus T^A_{\boldsymbol{\theta},\varphi'}, \tag{6.4}$$

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

It is easy to see that, when S is m-flat, it is m-information-curvature free. When $S_{\boldsymbol{\theta}}$ is not m-flat, $S_{\boldsymbol{\theta}}$ is curved in general, because its tangent directions change as φ changes.

Theorem 5 *When S is m-information-curvature free, $G^E = G^I$ for any $\boldsymbol{\theta}$ and φ. Moreover, $\boldsymbol{y}(x,\boldsymbol{\theta}) = \boldsymbol{u}^I(x,\boldsymbol{\theta},\varphi_0) = \boldsymbol{u}^E(x,\boldsymbol{\theta},\varphi_0)$ is the optimal estimating function at φ_0 and is efficient at φ_0.*

It should be noted that most semiparametric models so far treated by many researchers are m-flat. The important role of the m-flatness in the estimation function method is noted by Amari and Kumon (1988), Amari (1987), 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 that the estimating function method is fully efficient. However, the optimal estimating function depends on the true φ so that there is still a serious problem of choosing a good estimate φ_0 from observed data to derive a good estimating function. It is a merit of estimating functions that, even if we misspecify the true φ and choose a wrong φ_0, the estimator is still $\sqrt{n}$-consistent. A practical method of choosing a good estimating function is given by Amari and Kawanabe (1996).

7 Examples

Example 1. Mixture model

A mixture model is m-flat and is hence m-information-curvature free. In particular, when a model is given by (2.7) and (2.8), the $\boldsymbol{\theta}$-score, the information score $\boldsymbol{u}^I$, and the nuisance score in direction $a(\boldsymbol{\xi})$ can be calculated explicitly.

The $\boldsymbol{\theta}$-score $\boldsymbol{u}$ is given by

$$\boldsymbol{u} = \frac{1}{p(x, \boldsymbol{\theta}, \varphi)} \int (\partial_{\boldsymbol{\theta}} s \cdot \boldsymbol{\xi} + \partial_{\boldsymbol{\theta}} r - \partial_{\boldsymbol{\theta}} \psi) \varphi(\boldsymbol{\xi}) \exp\{\boldsymbol{\xi} \cdot s + r - \psi\} d\boldsymbol{\xi}. \quad (7.1)$$

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

$$p(\boldsymbol{\xi}|s) = \frac{\varphi(\boldsymbol{\xi}) \exp\{\boldsymbol{\xi} \cdot s - \psi\}}{\int \varphi(\boldsymbol{\xi}) \exp\{\boldsymbol{\xi} \cdot s - \psi\} d\boldsymbol{\xi}}, \quad (7.2)$$

the $\boldsymbol{\theta}$-score may be written as

$$\boldsymbol{u} = \partial_{\boldsymbol{\theta}} s \cdot \mathrm{E}[\boldsymbol{\xi}|s] + \partial_{\boldsymbol{\theta}} r - \mathrm{E}[\partial_{\boldsymbol{\theta}} \psi | s], \quad (7.3)$$

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

$$v[a] = \mathrm{E}\left[\left. \frac{a(\boldsymbol{\xi})}{\varphi(\boldsymbol{\xi})} \right| s \right]. \quad (7.4)$$

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

It is known that the projection of a random variable t to the space generated by s_i is given by the conditional expectation $\mathrm{E}[t|s_i]$ and the projection to 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

$$\begin{aligned} \boldsymbol{u}^I = \boldsymbol{u}^E &= \boldsymbol{u} - \mathrm{E}[\boldsymbol{u}|s] \\ &= \{\partial_{\boldsymbol{\theta}} s - \mathrm{E}[\partial_{\boldsymbol{\theta}} s|s]\} \cdot \mathrm{E}[\boldsymbol{\xi}|s] + \{\partial_{\boldsymbol{\theta}} r - \mathrm{E}[\partial_{\boldsymbol{\theta}} r|s]\}, \quad (7.5) \end{aligned}$$

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

Example 2. Blind separation of mixture signals

In order to assure the identifiability, we put further restrictions

$$\mathrm{E}[s_a] = 0 \quad (7.6)$$

$$\mathrm{E}[(s_a)^2] = 1 \quad (7.7)$$

for the source distributions ($a = 1, \cdots, r$). Then, the score functions (matrix) are

$$\begin{aligned} U = \frac{\partial}{\partial W} \log p(\boldsymbol{x}, W, \varphi) &= \sum_a \frac{\partial}{\partial W} \log q_a(s_a) + (W^{-1})^{\mathrm{T}} \\ &= \sum_a l'_a(s_a) \frac{\partial}{\partial W} s_a + M^{\mathrm{T}}, \end{aligned}$$

where

$$s_a = \sum_i W_i^a x_i, \qquad l'_a(s) = \frac{\mathrm{d}}{\mathrm{d}s} \log q_a(s).$$

In the component form, the score functions at the distribution (W, φ) are

$$u_a^i(\boldsymbol{x}, W, \varphi) = l'_a(s_a) x_i + M_a^i, \tag{7.8}$$

where M_a^i are the components of the mixing matrix M.

We next search for the nuisance scores. Let us consider a small change in the form of φ. We can write it as

$$q_a(s, \tau) = q_a(s)\{1 + \tau \alpha_a(s)\}, \tag{7.9}$$

where τ is the parameter to denote the changes of functions q_a in the direction of $\alpha_a(s)$.

The score function in the direction of the nuisance parameter φ in the direction of $\boldsymbol{\alpha} = (\alpha_a)$ is given by the score function

$$\begin{aligned} v(\boldsymbol{\alpha}) &= \left. \frac{\mathrm{d}}{\mathrm{d}\tau} \log p\{\boldsymbol{x}, W, \varphi(s, \tau)\} \right|_{\tau=0} \\ &= \sum_a \alpha_a(s_a). \end{aligned} \tag{7.10}$$

The linear space spanned by the functions $v(\boldsymbol{\alpha})$ is called the nuisance tangent space,

$$T_{W,\varphi}^N = \mathrm{span}\{\, v(\boldsymbol{\alpha})\}. \tag{7.11}$$

We then have the important observation that this model is m-information-curvature free. So the information score is given by the efficient score. The information score is given by

$$F_{W,\varphi}^I = \mathrm{span}\{l'_a(s_a)s_b,\ (s_a)^2 - c_a s_a - 1\}, \tag{7.12}$$

where $c_a = \mathrm{E}[(s_a)^3]$. The above consideration gives a very effective learning algorithm to this problem of blind signal separation. See Amari and Cardoso (1997) for details.

References

[1] Amari, S. (1985). *Differential-Geometrical Method in Statistics.* Lecture Notes in Statistics Vol.28, Springer, New York.

[2] Amari, S. (1987). Dual connections on the Hilbert bundles of statistical models. in Dodson, C.T.J. (ed.) *Geometrization of Statistical Theory.* 123 - 152. ULDM, Lancaster.

[3] Amari, S. and Cardoso, J. -P. (1997). Blind source separation - Semiparametric statistical approach. *IEEE Trans. on SIgnal Processing*, accepted.

[4] Amari, S., Cichocki, A., and Yang, M. (1996). A new learning algorithm for blind signal separation, in *NIPS'95* Vol.8, MIT Press, 757-763.

[5] Amari, S. and Kawanabe, M. (1996). Information geometry of estimating functions in semiparametric statistical models. *Bernoulli*, 3, 29-54.

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

[7] Andersen, E. B. (1970). Asymptotic properties of conditional maximum likelihood estimators. *J. R. Statist. Soc.* B 32, 283 - 301.

[8] 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.

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

[10] Chan, S. and Ghosh, M. (1996). Estimating functions: a linear space approach. Symposium on Estimating Functions, Univ. of Georgia.

[11] Durairajan, T.M. (1996). Optimal estimating function for estimation and predictions in semiparametric models. Symposium on Estimating Functions, Univ. of Georgia.

[12] Godambe, V. P. (1960). An optimum property of regular maximum likelihood estimation. *Ann. Math. Statist.* 31, 1208 - 1212.

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

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

[15] 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.

[16] Li, B. (1996). A minimax approach to consistency and efficiency for estimating equations. Symposium on Estimating Functions, Univ. of Georgia.

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

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

[19] McLeish, D. L. and Small, C. G. (1988). *The theory and applications of statistical inference functions*, Lecture Notes in Statistics 44. Springer, New York.

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

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

[22] Pfanzagl, J. (1990). *Estimation in Semiparametric Models: Some Recent Developments.* Lecture Notes in Statistics Vol.63, Springer, New York.

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

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

[25] Waterman, R. P. and Lindsay, B. G. (1996). Projection score methods for approximating conditional scores. *Biometrika* 83, 1 - 13.