Bayesian Estimating Equation Based on Hilbert Space [1]

Lu LIN

School of Mathematics and System Sciences, Shandong University,
Jinan, Shandong Province, 250100, P. R. China

E-mail: linlu@sdu.edu.cn

Abstract. This paper introduces and investigates the validity of Bayesian estimating equation derived from the Hilbert space method. A validity for Hilbert-based Bayesian estimating function is established via the Hilbert-based unbiasedness and information unbiasedness. As an application, the newly proposed method is used to construct an estimating equation for nonlinear regression model. Furthermore, the new notion is employed to lay a theoretical foundation for the penalty-based methods such as penalized likelihood and penalized least squares.

Keywords. Bayesian inference, Hilbert space method, estimating equation, valid inference, penalized likelihood, penalized squares.

AMS 2000 Subject Classification. 62F15, 62G99.

[1] This paper is supported by NNSF project (10371059) of China.

1. Introduction

The use of estimating equation has received much attention in the literature, see for example Godambe and Thompson (1984), Li and McCullagh (1994), Small and Mcleish (1994), Heyde (1997) and so on. Some comprehensive treatments may be found in Godambe (1991) and the viewpoint of numerical computation may be found in Small (2003). The estimating function plays the role of the score function whether or not the form of distribution of data is known. The main purpose of an estimating equation is to produce an estimator of parameter, the estimator being obtained as a root of this estimating equation. So the basis requirement of an estimating function is the unbiasedness for zero.

Estimating function can be derived from the moment conditions. A widely-used estimating function is so-called quasi score function, which is based only on the first and two moments of data or of a function of data and parameters, see for example Wedderburn (1974), Godambe and Heyde (1987), Heyde (1997), McCullagh (1983), McCullagh and Nelder (1989) and Lin (2004). A purely theoretical way to define an estimating function is the Hilbert space method, the estimating function being derived basically from on a function space endowed with a class of inner products, for comprehensive knowledge about this topic see Small and Mcleish (1994).

The derivation of Bayesian estimating function, namely, the posterior score, is due to Ferreira (1982) and Ghosh (1993). According to the point of Bayesian inference, an estimating function is designed to be unbiased in two senses of conditional unbiasedness and average unbiasedness for zero. A recent study by Lin (2006) introduced a theoretical framework of quasi Bayesian likelihood, in which the quasi posterior score and likelihood are based on the Hilbert space method. Although the general framework was established in this paper, the validity about Hilbert-based Bayesian inference has been unexplored. The essential difficulty with this topic may be that the true (posterior) likelihood function is in general unknown and then there is no an explicit foundation to define and compute the validity. So the validity of the Hilbert-based Bayesian estimating function is still a challenge.

A recent work by Monahan and Boos (1992) provided criterion for calculating whether or not an alternative likelihood is proper for Bayesian inference. They introduced a definition of validity based on the coverage property of posterior sets, along with a numerical technique that may be used to invalidate a certain likelihood. Lazar (2003) and Schennach (2005) used this notion to define and examine the validity of Bayesian empirical likelihood.

Following the above works, this paper introduces and investigates systemically the validity of Bayesian estimating equation. The estimating function is derived from the Hilbert space method and then the estimating function depends only on a space of estimating functions endowed with a class of inner products. The Hilbert-based posterior likelihood is suppositional and therefore can not be in general expressed explicitly. So some new and significant problems arise such as: How shall we define the validity? How shall we calculate and check the validity? And, how shall we use it in practice?

The remainder of this paper is organized as follows. In Section 2 the Bayesian likelihood is outlined via the Hilbert space method. A validity for Hilbert-based Bayesian estimating function is defined in Section 3 via a Hilbert-based (conditional) unbiasedness and information unbiasedness. In Section 4, some examples are presented to illustrate the theoretical results. A theoretical foundation for penalty-based methods, such as penalized likelihood and penalized least squares, is established in Section 5.

2. Hilbert-based likelihood

We first assume the distribution $P_\theta(y)$ of observations $y_i, \cdots, y_n$ belongs to a class of underlying distributions $\mathcal{P} = \{P_\theta(y), \theta \in \Theta\}$ and a dominating measure ν exists on R^n such that $dP_\theta(y) = p_\theta(y)d\nu$, where $p_\theta(y)$ is a density function with respect to the measure ν and the parameter space Θ is a subset of R^p. Then the true score function can be expressed as $s(\theta, y) = \partial \log p_\theta(y)/\partial\theta$. On the other hand, suppose that a prior distribution is defined on the parameter space Θ by $\Pi(\theta)$ with density $\pi(\theta) > 0$ for $\theta \in \Theta$. In this case the true posterior density of θ is $\pi(\theta|y) = \pi(\theta)p_\theta(y)/\int_\Theta \pi(\theta)p_\theta(y)d\theta$ and the true posterior score function can be expressed as

$$s(\theta|y) = \partial \log \pi(\theta|y)/\partial\theta = s(\theta, y) + \pi^{-1}(\theta)\dot{\pi}(\theta), \tag{2.1}$$

where $\dot{\pi}(\theta)$ stands for the derivative of $\pi(\theta)$.

However the forms of distribution of data and / or the prior distribution of parameter may be unknown. In this case a general theoretical framework for inference is desired. We now outline the theoretical framework based on Hilbert space method. For some similar notions see Small and Mcleish (1994) and Lin (2006).

Let a p-dimensional function space $\mathcal{G}$ be endowed with a family of inner products $\langle\cdot,\cdot\rangle_\theta$ indexed by $\theta \in \Theta$. According to the existing Hilbert space version (Small and Mcleish 1994), the mean of $g(\theta, y) \in \mathcal{G}$ is defined by $E_\theta(g(\theta, y)) = \langle g(\theta, y), \mathbf{1}_\mathcal{G}\rangle_\theta$, where $\mathbf{1}_\mathcal{G}$ is an unitary element of $\mathcal{G}$. If this mean is regarded as a linear functional E_θ: $\mathcal{G} \to R$, then, the Riesz representation theorem ensures that, under some regularity conditions, there exists a function $L(y|\theta)$ such that

$$E_\theta(g(\theta, y)) = \int_\mathcal{Y} g(\theta, y)L(y|\theta)dy, \text{ for any } g(\theta, y) \in \mathcal{G} \text{ and } \theta \in \Theta. \tag{2.2}$$

In this case we call $L(y|\theta)$ the Hilbert-based likelihood function. From this definition we can see that the Hilbert-based likelihood function depends on the defined inner $\langle\cdot,\cdot\rangle_\theta$ but not on the form of the distribution of data and then it is still available when the form of the distribution of data is unknown.

In this section, we suppose that the prior $\pi(\theta)$ of θ on Θ may be unknown and then define another function space $\mathcal{F}$: $\Theta \to R$ and an inner product $\langle\cdot,\cdot\rangle_0$ on $\mathcal{F}$. The mean of $f(\theta) \in \mathcal{F}$ is defined by $E_0(f(\theta)) = \langle f(\theta), \mathbf{1}_\mathcal{F}\rangle_0$, where $\mathbf{1}_\mathcal{F}$ is the unitary element of $\mathcal{F}$. The Riesz representation theorem ensures that, under some regularity conditions, there exists a function $\gamma(\theta)$ such that

$$E_0(f(\theta)) = \int_\Theta f(\theta)\gamma(\theta)d\theta. \tag{2.3}$$

We call $\gamma(\theta)$ the Hilbert-based prior density of θ. By inner product $\langle\cdot,\cdot\rangle_0$, together with inner product $\langle\cdot,\cdot\rangle_\theta$, we define a new inner on $\mathcal{G}$ as

$$\langle g_1(\theta, y), g_2(\theta, y)\rangle_* = \langle\langle g_1(\theta, y), g_2(\theta, y)\rangle_\theta, \mathbf{1}_\mathcal{F}\rangle_0, \text{ for any } g_1(\theta, y), g_2(\theta, y) \in \mathcal{G}.$$

Under this inner product, the mean of $g(\theta, y) \in \mathcal{G}$ is defined by $E_0E_\theta(g(\theta, y)) = \langle g(\theta, y), \mathbf{1}_\mathcal{G}\rangle_*$ and then, by Riesz representation theorem, there exists a function $L(\theta, y)$ such that

$$E_0E_\theta(g(\theta, y)) = \int_\Theta\int_\mathcal{Y} g(\theta, y)L(\theta, y)dyd\theta \text{ for any } g(\theta, y) \in \mathcal{G} \text{ and } \theta \in \Theta. \tag{2.4}$$

In this case we call $L(\theta, y)$ the Hilbert-based joint density function of θ and y. Combining (2.2), (2.3) and (2.4) leads to

$$L(\theta, y) = \gamma(\theta)L(y|\theta). \tag{2.5}$$

Finally, we define

$$L(\theta|y) = \frac{\gamma(\theta)L(y|\theta)}{p(y)} \tag{2.6}$$

as the Hilbert-based posterior density function of θ given y, and

$$h(\theta|y) = \partial \log L(\theta|y)/\partial\theta \tag{2.7}$$

as the Hilbert-based posterior score function of θ given y, where

$$p(y) = \int_{\Theta} \gamma(\theta)L(y|\theta)d\theta. \tag{2.8}$$

The new theoretical framework is similar to the classical one. The basic characteristic of the Hilbert-based method is that it depends on the defined inner products but not on the distributions of data and parameter θ, and then is still available when these distributions are unknown.

3. Hilbert-based Bayesian estimating equation

As aforementioned, the basis requirement for constructing a proper estimating function is the unbiasedness for zero. In Bayesian context (Ferreira 1982 and Ghosh 1993), the the unbiasedness can be defined as follows. A Bayesian estimating function $g(\theta, y)$ is said to be conditionally unbiased, if

$$E(g(\theta, y)|y) = 0 \tag{3.1}$$

holds with probability one. And a function $g(\theta, y)$ is said to be average unbiased if

$$E(g(\theta, y)) = 0. \tag{3.2}$$

The further requirement for constructing a proper estimating function is information unbiased. A Bayesian estimating function $g(\theta, y)$ is said to be conditionally information unbiased, if

$$E(g(\theta, y)g'(\theta, y)|y) = -E(\dot{g}(\theta, y)|y) \tag{3.3}$$

holds with probability one, where $\dot{g}(\theta, y)$ is the derivative of $g(\theta, y)$ with respect to θ. Similarly, a Bayesian estimating function $g(\theta, y)$ is said to be average information unbiased, if

$$E(g(\theta, y)g'(\theta, y)) = -E(\dot{g}(\theta, y)). \tag{3.4}$$

We need the unbiasedness to get a consistent estimator of θ. According to the Bayesian point of view, an estimator $\hat{\theta}_n$ is said to be consistent if $\hat{\theta}_n - \theta$ converges in probability to zero as $n \to \infty$, where the probability is calculated suing the posterior distribution of θ given y. That is the following holds

$$P\{|\hat{\theta}_n - \theta| < \varepsilon | y\} > 1 - \eta, \quad n > N,$$

for some N, given any $\varepsilon, \eta > 0$. Unlike the frequentist definition of consistency, where the parameter θ is fixed and the estimator $\hat{\theta}_n$ is a random vector, the roles are reversed here. The estimator $\hat{\theta}_n$ is a function of the data y and then is constant given y, while θ is regarded as random vector in the definition of Bayesian consistency above.

Under both the unbiasedness and the information unbiasedness, furthermore, the estimating function may share some of the properties that are typically associated with log-likelihoods. For example, in this case the common iterative algorithms result in a convergent iterative solution. For details see for example Small (2003). More precisely, for the existence of log-likelihood, we require that the estimating function is conservative. A common possibility to this goal is the matrix $\dot{g}(\theta, y)$ is symmetric for all θ and all y. In this case the vector field is conservative so that there exists a real-valued function $Q(\theta)$ such that $\dot{Q}(\theta) = g(\theta, y)$. The function $Q(\theta)$ could be a log-likelihood.

It is worth noting that, in the Hilbert space context, the expectation in (3.1) is taken under the Hilbert-based posterior likelihood function $L(\theta|y)$ as defined in (2.6). We then extend the definitions above to the Hilbert space context as follows. An estimating function $g(\theta, y)$ is said to be the Hilbert-based conditionally unbiased, if

$$E_L(g(\theta, y)|y) \equiv \int_\Theta g(\theta, y)L(\theta|y)d\theta = 0 \tag{3.5}$$

holds with probability one, where the probability is based on the probability density $p(y)$ defined by (2.8). Similarly, a function $g(\theta, y)$ is said to be the

Hilbert-based average unbiased if

$$E_L(g(\theta, y)) \equiv \int_{\Theta} \int_{\mathcal{Y}} g(\theta, y) L(\theta, y) dy d\theta = 0. \tag{3.6}$$

Using the relation between the expectation with respect to the Hilbert-based joint likelihood and that with respect to the Hilbert-based posterior likelihood, we can see that (3.5) implies (3.6), in general.

Similar to (3.3) and (3.4), in the Hilbert space context, we say that a Bayesian estimating function $g(\theta, y)$ is Hilbert-based conditionally information unbiased, if

$$E_L(g(\theta, y) g'(\theta, y) | y) = -E_L(\dot{g}(\theta, y) | y) \tag{3.7}$$

holds with probability one. And a function $g(\theta, y)$ is said to be Hilbert-based average information unbiased, if

$$E_L(g(\theta, y) g'(\theta, y)) = -E_L(\dot{g}(\theta, y)). \tag{3.8}$$

Note that here the expectations $E_L(\cdot|y)$ and $E_L(\cdot)$ are defined respectively as in (3.5) and (3.6). Again, the Hilbert-based conditional information unbiasedness implies in general the Hilbert-based average information unbiasedness.

4. Examples.

In what follows we always suppose that it is permitted to interchange the differentiation with respect to θ and the integration over the parameter space Θ or the sample space $\mathcal{Y}$. To illustrate the theory above, we consider following examples. For the sake of convenience, we assume that, in the examples below, θ is supposed to be a real-valued parameter and belongs to (a, b), where both a and b may be infinite.

Example 1. We here consider the validity of the Hilbert-based posterior score function defined by (2.7). By the assumptions above and the definition (2.7) we have

$$E_L(h(\theta|y)|y) = \int_a^b \dot{L}(\theta|y) d\theta = \lim_{\theta \to b^-} L(\theta|y) - \lim_{\theta \to a^+} L(\theta|y),$$

where $\dot{L}(\theta|y)$ stands for the derivative of $L(\theta|y)$ with respect to θ. This means that the Hilbert-based posterior score function $h(\theta|y)$ is Hilbert-based conditionally unbiased if and only if

$$\lim_{\theta \to b^-} L(\theta|y) - \lim_{\theta \to a^+} L(\theta|y) = 0 \tag{4.1}$$

holds with probability one. Furthermore, from the definition (2.7) it follows that

$$\begin{aligned} E_L(\dot{h}(\theta|y)) &= \int_a^b [L^{-1}(\theta|y)\ddot{L}(\theta|y) - L^{-2}(\theta|y)\dot{L}(\theta|y)\dot{L}'(\theta|y)]L(\theta|y)d\theta \\ &= \lim_{\theta\to b^-}\dot{L}(\theta|y) - \lim_{\theta\to a^+}\dot{L}(\theta|y) - E_L(h(\theta|y)h'(\theta|y)). \end{aligned}$$

Then

$$E_L(h(\theta|y)h'(\theta|y)|y) = -E_L(\dot{h}(\theta|y)|y)$$

if and only if

$$\lim_{\theta\to b^-}\dot{L}(\theta|y) - \lim_{\theta\to a^+}\dot{L}(\theta|y) = 0 \tag{4.2}$$

holds with probability one. Thus $h(\theta|y)$ is Hilbert-based conditionally information unbiased if and only if the condition (4.2) holds with probability one.

From (2.6) we can see that some common distributions, such as normal distribution, Poisson distribution, exponential distribution and uniform distribution, satisfy the conditions (4.1) and (4.2). However, we can not guarantee that the conditions (4.1) and (4.2) is always satisfied. This means that, like the classical posterior score function $s(\theta|y)$, $h(\theta|y)$ is sometimes Hilbert-based conditionally biased and conditionally information biased.

We now turn to the average unbiasedness. From (2.6) and (2.7) we can see that

$$h(\theta|y) = \frac{\partial \log L(y|\theta)}{\partial\theta} + \frac{\partial \log\gamma(\theta)}{\partial\theta}.$$

Note that $\int_{\mathcal{Y}} \frac{\partial L(y|\theta)}{\partial\theta}dy = 0$. Consequently,

$$\begin{aligned} E_L(h(\theta|y)) &= \int_a^b\int_{\mathcal{Y}} h(\theta|y)\gamma(\theta)L(y|\theta)dyd\theta \\ &= \int_a^b \gamma(\theta)d\theta\int_{\mathcal{Y}}\frac{\partial L(y|\theta)}{\partial\theta}dy + \int_a^b\frac{\partial\gamma(\theta)}{\partial\theta}d\theta\int_{\mathcal{Y}}L(y|\theta)dy \\ &= \lim_{\theta\to b^-}\gamma(\theta) - \lim_{\theta\to a^+}\gamma(\theta). \end{aligned}$$

Therefore, $h(\theta|y)$ is Hilbert-based average unbiased if and only if

$$\lim_{\theta\to b^-}\gamma(\theta) - \lim_{\theta\to a^+}\gamma(\theta) = 0. \tag{4.3}$$

To get the Hilbert-based average information unbiasedness, we suppose

$$\lim_{\theta\to b^-}\dot{\gamma}(\theta) - \lim_{\theta\to a^+}\dot{\gamma}(\theta) = 0. \tag{4.4}$$

Denote by $\dot{L}(y|\theta)$ the derivative of $L(y|\theta)$ with respect to θ. Note that

$$\int_a^b \int_{\mathcal{Y}} \dot{r}(\theta)\dot{L}(y|\theta)dyd\theta = 0$$

and

$$\int_{\mathcal{Y}} \int_a^b \ddot{r}(\theta)L(y|\theta)d\theta dy = \int_{\mathcal{Y}} \Big[\dot{r}(\theta)L(y|\theta)\Big|_a^b - \int_a^b \dot{r}(\theta)\dot{L}(y|\theta)d\theta\Big]dy = 0.$$

Then

$$\begin{aligned} E_L(\dot{h}(\theta|y)) &= \int_a^b \int_{\mathcal{Y}} \Big[L^{-1}(y|\theta)\ddot{L}(y|\theta) - L^{-2}(y|\theta)\dot{L}^2(y|\theta) \\ &\qquad +r^{-1}(\theta)\ddot{r}(\theta) - r^{-2}(\theta)\dot{r}^2(\theta)\Big] r(\theta)L(y|\theta)dyd\theta \\ &= -E_L(h^2(\theta|y)). \end{aligned}$$

Therefore, $h(\theta|y)$ is Hilbert-based average information unbiased if (4.4) holds.

The condition (4.3) is very common. So $h(\theta|y)$ is in general Hilbert-based average unbiased. Comparing the condition (4.4) with (4.1) and (4.2), the condition (4.4) is mild and, consequently, $h(\theta|y)$ is sometimes Hilbert-based average information unbiased. □

Example 1 shows that, under some conditions, the Hilbert-based posterior score function $h(\theta|y)$ is a proper estimating function. In general, however, it does not lend us a hand for the Hilbert-based Bayesian inference because it depends on the Hilbert-based posterior likelihood function $L(\theta|y)$, which is sometimes unknown. But the following example gives us an useable Hilbert-based method to construct a proper estimating function for general nonlinear regression models.

Example 2. Suppose that the n-dimensional random variable $y = (y_1, \cdots, y_n)'$ has mean $\mu(\theta) = (\mu_1(\theta), \cdots, \mu_n(\theta))'$ and covariance matrix $\sigma^2 V(\theta) \equiv \sigma^2(v_{ij}(\theta))$, the estimating function space is chosen as

$$\mathcal{G} = \left\{ g(\theta, y) : g(\theta, y) = \sum_{i=1}^n a_i(\theta)y_i + c(\theta) \right\},$$

where $a_i(\theta)$ and $c(\theta)$ are arbitrary functions depending only on θ. The main goal of this example is to find an optimal estimating function in $\mathcal{G}$.

Assume that an inner product on $\mathcal{G}$ is defined by

$$\langle y_i, 1\rangle_\theta = \mu_i(\theta) \quad \text{and} \quad \langle y_i, y_j\rangle_\theta = \sigma^2 v_{ij}(\theta) + \mu_i(\theta)\mu_j(\theta) \quad \text{for all } i, j = 1, \cdots, n.$$

In this case the usual notion of covariance is

$$Cov_\theta(y_i, y_j) = \langle y_i - E_\theta(y_i), y_j - E_\theta(y_j)\rangle_\theta.$$

Then we can verify by definition that the estimating function $g(\theta, y) \in \mathcal{G}$ is Hilbert-based conditionally unbiased if and only if

$$\sum_{i=1}^{n} a_i(\theta)\mu_i(\theta) + c(\theta) = 0.$$

Furthermore, define an inner product on $\mathcal{F}$: $\Theta \to R$ by

$$\langle f_1(\theta), f_2(\theta)\rangle_0 = \int_\Theta f_1(\theta)f_2(\theta)\pi(\theta)d\theta \text{ for any } f_1(\theta), f_2(\theta) \in \mathcal{F},$$

where $\pi(\theta)$ is the prior density of θ on Θ and is supposed to be known. As a result, we get an inner product defined by

$$\langle y_i, 1\rangle_* = \int_\Theta \mu_i(\theta)\pi(\theta)d\theta$$

and

$$\langle y_i, y_j\rangle_* = \int_\Theta [\sigma^2 v_{ij}(\theta) + \mu_i(\theta)\mu_j(\theta)]\pi(\theta)d\theta \text{ for all } i, j = 1, \cdots, n.$$

Let $L^2(P_\theta, \Pi)$ denote the class of estimating functions $g(\theta, y)$: $\Theta \times \mathcal{Y} \to R$ such that $E_\pi E_\theta(g(\theta, y)g(\theta, y)) = \int_\Theta \int_\mathcal{Y} g(\theta, y)g(\theta, y)dP_\theta(y)d\Pi(\theta) < \infty$, where $P_\theta(y)$ and $\Pi(\theta)$ stand for the distribution functions of data and θ, respectively. Then we can verify that the analytic continuation of the inner product $\langle \cdot, \cdot\rangle_*$ to $L^2(P_\theta, \Pi)$ is

$$\langle g_1(\theta, y), g_2(\theta, y)\rangle = \int_\Theta \int_\mathcal{Y} g_1(\theta, y)g_2(\theta, y)dP_\theta(y)d\Pi(\theta),$$

for any $g_1(\theta, y), g_2(\theta, y) \in L^2(P_\theta, \Pi)$. It follows from Corollary 1 of Lin (2006) that the projection of true posterior score function $s(\theta|y)$ onto $\mathcal{G}$ is

$$q(\theta|y) = -q(\theta, y) + \pi^{-1}(\theta)\dot{\pi}(\theta), \tag{4.5}$$

where $q(\theta, y)$ is the quasi score function as defined by

$$q(\theta, y) = \sigma^{-2}\{\dot{\mu}(\theta)\}'\{V(\theta)\}^{-1}e(y, \theta),$$

$\dot{\mu}(\theta)$ is an n-dimensional column vector with components $\partial\mu_i(\theta)/\partial\theta$ and $e(y, \theta) = y - \mu(\theta)$. We call $q(\theta|y)$ the quasi posterior score function in $\mathcal{G}$.

Note that

$$E_L(q(\theta,y)) = 0 \;\text{ and }\; E_L(\pi^{-1}(\theta)\dot{\pi}(\theta)) = \lim_{\theta\to b^-}\pi(\theta) - \lim_{\theta\to a^+}\pi(\theta).$$

Then $q(\theta|y)$ is Hilbert-based average unbiased if and only if

$$\lim_{\theta\to b^-}\pi(\theta) - \lim_{\theta\to a^+}\pi(\theta) = 0. \tag{4.6}$$

Note that $q(\theta,y)$ is Hilbert-based conditionally information unbiased. Then $q(\theta|y)$ is Hilbert-based average information unbiased if and only if

$$\lim_{\theta\to b^-}\dot{\pi}(\theta) - \lim_{\theta\to a^+}\dot{\pi}(\theta) = 0. \tag{4.7}$$

The conditions(4.6) and (4.7) are mild in general. The results above mean that, under the condition (4.6) and (4.7), the optimal estimating function in $\mathcal{G}$ is $q(\theta|y)$ defined by (4.5). Here the optimality means that $q(\theta|y)$ is the projection of the true posterior sore function onto $\mathcal{G}$ and, at the same time, is Hilbert-based average unbiased and average information unbiased. It is also worth pointing out that the inference here depends only on the form of inner product and thus is free of the form of distribution of data.

Particularly, consider the following linear regression model:

$$E_\theta(y) = X\theta, \quad Var(y) = \sigma^2 I, \tag{4.8}$$

where X is an $n\times p$ design matrix and $\theta \sim N(0, k^{-1}\sigma^2 I)$ for some $k > 0$. This normal prior satisfies (4.6) and (4.7). By (4.5), the quasi posterior score function can be expressed as

$$q(\theta|y) = \sigma^{-2}X'(y - X\theta) - k\sigma^{-2}\theta. \tag{4.9}$$

This estimating equation is Hilbert-based average unbiased and average information unbiased. Solving the equation $q(\theta|y) = 0$ for θ leads to a quasi posterior estimator of θ as

$$\hat{\theta}(k) = (X'X + kI)^{-1}X'y, \tag{4.10}$$

which is just the Ridge estimator or the Bayesian estimator if the distribution of data is normal, as suggested in literature. □

5. A principle for penalty-based methods

The penalty-based method has become an important topic in both parametric and nonparametric statistics inferences. To review its basic framework, let us look at the following two examples.

The first example is the penalized least squares. Consider linear regression model (4.8) and, for the sake of convenience, σ^2 is supposed to be 1. It is well-known that, when $X'X$ is nearly singular, the least squares method will have a bad property of mean error. To improve the least squares, the penalized least squares is introduced by minimizing

$$Q(\theta) = \frac{1}{2}||y - X\theta||^2 + \frac{k}{2}\theta'\theta,$$

where $k > 0$ is a parameter. Obviously, the objective above is designed to be a length penalty of parameter vector θ. The corresponding estimating equation is

$$g(\theta, y) = X'(y - X\theta) - k\theta, \tag{5.1}$$

and the resulting estimator is the Ridge estimator as defined as in (4.10). Note that $X'(y-X\theta)$ is Hilbert-based average unbiased and average information unbiased by the results of the example 2 above. Then $g(\theta, y)$ is Hilbert-based average unbiased if and only if

$$\int_\Theta \theta\gamma(\theta)d\theta = 0, \tag{5.2}$$

where $\gamma(\theta)$ is the Hilbert-based prior density. A common sufficient condition for (5.2) is that the prior density $\gamma(\theta)$ is a even function. Further, we can verify that $g(\theta, y)$ is Hilbert-based average information unbiased if and only if

$$\int_\Theta \theta\theta'\gamma(\theta)d\theta = \frac{1}{k}I. \tag{5.3}$$

A common sufficient condition for (5.3) is that the elements $\theta_1, \cdots, \theta_p$ of θ are independent with a mean $E(\theta_i) = 0$ variance $Var(\theta_i) = k^{-1}$, where the mean and variance are taken under the Hilbert-based prior $\gamma(\theta)$. This means that in the Bayesian context, the above penalized least squares is proper if (5.2) and (5.3) hold.

The second example is the penalized likelihood for variable selection. Consider a generalized linear model: $\{\mathbf{x}_i, y_i\}$ are collected independently with density $f_i(h(\mathbf{x}_i'\theta), y_i)$, where $h(\cdot)$ is a known link function. Let $l_i = \log f_i$ denote the conditional log-likelihood of y_i. A form of penalized likelihood is defined by

$$Q(\theta) = \sum_{i=1}^{n} l_i(h(\mathbf{x}_i'\theta), y_i) - n\sum_{j=1}^{p} p_i(\theta_i).$$

This objective is designed as a dimensionality penalty of parameter vector θ. The corresponding estimating function is

$$g(\theta, y) = \dot{Q}(\theta) = \sum_{i=1}^{n} \mathbf{x}_i \dot{l}(h(\mathbf{x}_i'\theta), y_i) - n(\dot{p}_1(\theta_1), \cdots, \dot{p}_p(\theta_p))'. \tag{5.4}$$

As a true score function, $\sum_{i=1}^{n} \mathbf{x}_i \dot{l}(h(\mathbf{x}_i'\theta), y_i)$ is average unbiased and average information unbiased. Then, like the first example, here $g(\theta, y)$ is average unbiased if and only if

$$\int_{\Theta} (\dot{p}_1(\theta_1), \cdots, \dot{p}_p(\theta_p))'\gamma(\theta) d\theta = 0. \tag{5.5}$$

A common sufficient condition for (5.5) is that $\dot{p}_i(\theta_i)$ are odd functions and the Hilbert-based prior density $\gamma(\theta)$ is a even function. The widely-used penalty functions for variable selection satisfy this condition. A common example, proposed by Fan and Li (2001), is defined by satisfying

$$\dot{p}_i(\theta_i) = \lambda \left\{ I(\theta_i \leq \lambda) + \frac{(a\lambda - \theta_i)_+}{(a-1)\lambda} I(\theta_i > \lambda) \right\} \tag{5.6}$$

for some $a > 2$ and $\theta_i > 0$, and $\dot{p}_i(-\theta_i) = -\dot{p}_i(\theta_i)$. Here the penalty parameter λ satisfies $\lambda \to 0$ as $n \to \infty$. The penalty function defined by (5.6) satisfy condition (5.5) if $\gamma(\theta)$ is even function. Further, $g(\theta, y)$ is average information unbiased if and only if

$$\int_{\Theta} \ddot{p}_i(\theta_i)\gamma(\theta) d\theta = -n \int_{\Theta} \dot{p}_i^2(\theta_i)\gamma(\theta) d\theta, \int_{\Theta} \dot{p}_i(\theta_i)\dot{p}_j(\theta_j)\gamma(\theta) d\theta = 0 \text{ for } i \neq j. \tag{5.7}$$

Note that the penalty function defined by (5.6) does not satisfy (5.7), in general. However, when $\theta_1, \cdots, \theta_p$ are independent and there exists a element, say θ_1, satisfying $\theta_1 \equiv 0$, the condition (5.7) holds. This means that, for variable selection, the above penalty function is proper in the sense of that the resulting estimating function is both average unbiased and average information unbiased.

By summarizing the framework of estimating functions given in (5.1) and (5.4), we can see that, in general, the estimating functions in the penalty-based methods have the following decomposable structure:

$$g(\theta, y) = g_1(\theta, y) + g_2(\theta), \tag{5.8}$$

where $g_1(\theta, y)$ is Hilbert-based average unbiased and average information unbiased, and $g_2(\theta)$ is a penalty function to be specified. In this case $g(\theta, y)$ is

Hilbert-based average unbiased or average information unbiased if and only $g_2(\theta)$ is unbiased or information unbiased, respectively.

The above discussion proposes a principle that can be used to select penalty function for the penalty-based methods in the Bayesian context. The principle can be stated as follows: In the decomposable form (5.8), when $g_1(\theta, y)$ is Hilbert-based average unbiased and besides, is sometimes Hilbert-based average information unbiased, the penalty function $g_2(\theta)$ should be chosen to be at least unbiased and, if possible, be information unbiased.

References

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its Oracle properties, *JASA*. **96**,1348-1360.

Ferreira, P. E. (1982). Estimating equation in the presence of prior knowledge, Biometrika, **69**, 667-669.

Ghosh, M. (1993). On a Bayesian analog of the theory of estimating function, *C. G. Khatri Memorial Volume Gujarat Statistical Review*, **17A**, 47-52.

Godambe, V. P. (1991). *Estimating functions*, Clarendon Press, Oxford, New York.

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

Godambe, V. P. and Thompson, M. E. (1984). Robust estimation through estimating equation, Biometrika **71**, 115-125.

Heyde, C. C. (1997). *Quasi-likelihood and its application*, Springer-Verlag New York, Inc.

Lazar, N .A. (2003). Bayesian empirical likelihood, Biometrika, **90**, 319-326.

Li, B. and McCulagh, P. (1994). Potential function and conservative estimating functions, Ann. Statist. **22**, 340-356.

Lin, L. (2004). Generalized quasi-likelihood, Statistics Papers **45**, 529-544.

Lin, L. (2006). Quasi Bayesian likelihood, Statistical Methodology (preprint).

McCullagh, P. (1983). Quasi-Likelihood functions, Ann. Statist. **11**, 59–67.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, Second Edition, Chmpman and Hall, London.

Monahan, J. F. and Boos, D. D. (1992). Proper likelihood for Bayesian analysis, Biometrika **79**, 271-278.

Schennach, S. M. (2005). Bayesian exponential tilted empirical likelihood, Biometrika **92**, 31-46.

Small, C. G. (2003). *Numerical methods for nonlinear estimating equations,* Clarendon Press, Oxford, New York.

Small, C. G. and Mcleish, D. L. (1994). *Hilbert space methods in probability and statistical inference*, John Wiley and Sons, Inc.

Wedderburn, R. W. M. (1974). Quasi-Likelihood functions, generalized Linear models and the Guess-Newton method, Biometrika **61**, 439-447.