Bayesian Estimating Equation Based on Hilbert Space Method

Lu LIN
(Shandong University)

Nankai University, Tianjin, China, 2006.7.12

Outline

- Hilbert-based likelihood
- Hilbert-based Bayesian estimating equation
- An example

1. Hilbert-based likelihood

- Posterior density and score functions.

$p_\theta(y)$ is the density function of data Y.

$\pi(\theta)$ is the prior density function of parameter θ.

The posterior density of θ is defined by

$$\pi(\theta|y) = \pi(\theta)p_\theta(y)/\int_\Theta \pi(\theta)p_\theta(y)d\theta.$$

The posterior score function can be expressed as

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

If the function forms of density $p_\theta(y)$ and / or the prior density $\pi(\theta)$ are unknown, then the function forms of the posterior density $\pi(\theta|y)$ and posterior score function $s(\theta|y)$ are unknown as well.

In this case a new theoretical framework for Bayesian inference is desired.

• Hilbert-based likelihood function.

To find a replacer of the density function of data Y, we define a function space $\mathcal{G}$ endowed with a family of inner products $\langle \cdot, \cdot \rangle_\theta$ indexed by $\theta \in \Theta$. According to the existing Hilbert space version, 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: $\mathcal{G} \longrightarrow R$, then, the Riesz representation theorem ensures that, under some regularity conditions, there exists a function $L(y|\theta)$ such that

$$E_\theta(g(\theta, y)) = \int_\mathcal{Y} g(\theta, y) L(y|\theta) dy. \tag{1.1}$$

In this case we call $L(y|\theta)$ the Hilbert-based likelihood function.

The Hilbert-based likelihood function depends on the defined inner product $\langle \cdot, \cdot \rangle_\theta$ 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.

• Hilbert-based prior density.

To get a replacer of the prior density of θ, we 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

$$E_0(f(\theta)) = \int_{\Theta} f(\theta)\gamma(\theta) d\theta. \tag{1.2}$$

We call $\gamma(\theta)$ the Hilbert-based prior density of θ.

• Hilbert-based joint density function

By inner product $\langle\cdot,\cdot\rangle_0$, together with inner product $\langle\cdot,\cdot\rangle_\theta$, we define a derived inner product on $\mathcal{G}$ as

$$\langle g_1(\theta,y), g_2(\theta,y)\rangle_* = \langle\langle g_1(\theta,y), g_2(\theta,y)\rangle_\theta, \mathbf{1}_\mathcal{F}\rangle_0$$

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

$$E_0E_\theta(g(\theta,y)) = \int_\Theta\int_\mathcal{Y} g(\theta,y)L(\theta,y)dyd\theta. \tag{1.3}$$

In this case we call $L(\theta,y)$ the Hilbert-based joint density function of θ and Y.

- Hilbert-based posterior score function

  Combining (1.1), (1.2) and (1.3) leads to

$$L(\theta, y) = \gamma(\theta) L(y|\theta). \tag{1.4}$$

Finally, we define

$$L(\theta|y) = \frac{\gamma(\theta) L(y|\theta)}{p(y)} \tag{1.5}$$

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

$$h(\theta|y) = \partial \log L(\theta|y)/\partial\theta \tag{1.6}$$

as the Hilbert-based posterior score function of θ given y, where $p(y) = \int_\Theta \gamma(\theta) L(y|\theta) d\theta$.

The new theoretical framework 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.

- **2. Hilbert-based Bayesian estimating equation**
- Hilbert-based unbiasedness.

Let $\hat{\theta}$ be the root of an estimating equation $g(\theta, y) = 0$. We need the unbiasedness to get a consistent estimator $\hat{\theta}$ of θ.

An estimating function $g(\theta, y)$ is said to be the Hilbert-based conditionally unbiased, if

$$E_L(g(\theta, y)|y) \equiv \int_\Theta g(\theta, y) L(\theta|y) d\theta = 0 \tag{2.1}$$

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

$$E_L(g(\theta, y)) \equiv \int_\Theta \int_\mathcal{Y} g(\theta, y) L(\theta, y) dy d\theta = 0. \tag{2.2}$$

- Hilbert-based information unbiasedness.

We say that a Bayesian estimating function $g(\theta, y)$ is Hilbert-based conditionally information unbiased, if

$$E_L(g(\theta, y)g'(\theta, y)|y) = -E_L(\dot{g}(\theta, y)|y) \tag{2.3}$$

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

$$E_L(g(\theta, y)g'(\theta, y)) = -E_L(\dot{g}(\theta, y)). \tag{2.4}$$

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.

3. An example

• Function space. 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 of $\theta \in (a, b)$. The main goal of this example is to find an optimal estimating function in $\mathcal{G}$.

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

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

In this case the usual notion of covariance is

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

• Inner product defined on $\mathcal{F}$. Furthermore, define an inner product on $\mathcal{F}$: $\Theta \rightarrow R$ 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.

• Derived product defined on $\mathcal{G}$. We get an inner product defined by

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

and

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

• Projection of true posterior score function. We can verify that the projection of true posterior score function $s(\theta|y)$ onto $\mathcal{G}$ is

$$q(\theta|y) = -q(\theta, y) + \pi^{-1}(\theta)\dot{\pi}(\theta), \tag{3.1}$$

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 $e(y, \theta) = y - \mu(\theta)$.

• Quasi posterior score. We call $q(\theta|y)$ the quasi posterior score function in $\mathcal{G}$.

• Unbiasedness. An estimating function $g(\theta, y) \in \mathcal{G}$ is Hilbert-based conditionally unbiased if and only if

$$\sum_{i=1}^{n} a_i(\theta)\mu_i(\theta) + c(\theta) = 0. \tag{3.2}$$

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

$$\lim_{\theta \to b^-} \pi(\theta) - \lim_{\theta \to a^+} \pi(\theta) = 0. \tag{3.3}$$

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

$$\lim_{\theta \to b^-} \dot{\pi}(\theta) - \lim_{\theta \to a^+} \dot{\pi}(\theta) = 0. \tag{3.4}$$

The conditions(3.2)-(3.4) are mild. These conditions give a principle about how to choose prior density $\pi(\theta)$.

• Optimal estimating function.

The optimal estimating function in $\mathcal{G}$ is $q(\theta|y)$, the quasi score function, defined by (3.1), i.e,

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

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.

• Linear regression. Particularly, consider the following linear regression model:

$$E_\theta(y) = X\theta, \quad Var(y) = \sigma^2 I, \tag{3.5}$$

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

$$q(\theta|y) = \sigma^{-2}X'(y - X\theta) - k\sigma^{-2}\theta. \tag{3.6}$$

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

$$\hat{\theta}(k) = (X'X + kI)^{-1}X'y, \tag{3.7}$$

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

• Extensions.

Hilbert-based Bayesian estimating equation is a general theoretical framework. It can be employed to investigate many statistical problems including penalized least squares, penalized likelihood and variable selection.

Thank you very much !