Seminar on Statistics, FS 2008

Bayesian Statistics: ASYMPTOTICS

Francesco Croci and Christian Veit

Monday, March 31, 2008

1 Introduction and basic definitions

In this seminar class, we consider the properties of the posterior that hold in the limit as the sample size becomes large.

We now collect some important definitions and facts.

Definition 1.1 (consistent)**.** *A sequence of estimators $(T_n)_{n\in\mathbb{N}}$ is called* consistent *for $g(\theta)$ if for all θ*

$$T_n \to g(\theta) \text{ in } P_\theta\text{-probability,}$$

that is

$$\forall\varepsilon > 0 \text{ and } \forall\theta \in \Theta : \quad \lim_{n\to\infty} P_\theta(\|T_n - g(\theta)\| > \varepsilon) = 0.$$

Theorem 1.2 (Cramér Rao Lower Bound)**.** *Suppose that $\Theta \subset \mathbb{R}$ is an open interval, $A := \{x : p_\theta(x) > 0\}$ does not depend on θ (so $P_\theta(A^c) = 0$), $\psi_\theta(x) := \frac{d}{d\theta}\log(p_\theta(x))$ exists in $L^2(P_\theta)$, i.e.*

$$\lim_{\Delta\to 0}\int \left|\frac{p_{\theta+\Delta}(x) - p_\theta(x)}{p_\theta(x)\Delta} - \psi_\theta(x)\right|^2 P_\theta(dx) = 0,$$

and suppose that T is an unbiased estimator of $g(\theta)$. Then
$E_\theta[\psi_\theta(x)] = \int \psi_\theta(x)P_\theta(dx) = 0 \;\; \forall\theta$, $g'(\theta) = \frac{d}{d\theta}g(\theta)$ exists, and

$$Var_\theta[T] \geq \frac{[g'(\theta)]^2}{I(\theta)}$$

where $I(\theta)$ is the Fisher-Information:

$$I(\theta) = E_\theta[\psi_\theta^2(x)] \;\; \forall\theta.$$

Definition 1.3 (asymptotically efficient)**.** *Assume regularity: if*

$$\sqrt{n}(T_n - \theta) \to^d \mathcal{N}(0, 1/I(\theta)) \;\; \forall\theta,$$

then: T_n is called asymptotically efficient*.*

2 Normal approximation to the posterior distribution

Definition 2.1 (posterior mode or maximum a posteriori estimator)**.** *The* posterior mode or maximum a posteriori estimator *is defined by*

$$\widehat{\theta} = \arg\max_{\theta} p(x|\theta)\pi(\theta),$$

where $\pi(.)$ *is the prior density.*

The results in this section are given under some regularity conditions (notably that the likelihood is a continuous function of θ and that θ_0, the true parameter value, is not on the boundary of the parameter space).

If the posterior distribution $p(\theta|y)$ is unimodal and roughly symmetric, it is often convenient to approximate it by a normal distribution centered at the mode. A Taylor series expansion of $\log p(\theta|y)$ centered at the posterior mode, $\widehat{\theta}$ gives

$$\log p(\theta|y) = \log p(\widehat{\theta}|y) + \frac{1}{2}(\theta - \widehat{\theta})^T \left[\frac{d^2}{d\theta^2} \log p(\theta|y)\right]_{\theta=\widehat{\theta}} (\theta - \widehat{\theta}) + \dots \quad (1)$$

where the linear term in the expansion is zero because the log-posterior density has zero derivative at its mode, the remainder terms of higher order fade in importance relative to the quadratic term when θ is close to $\widehat{\theta}$ and n is large.

This yields the approximation

$$p(\theta|y) \approx \mathcal{N}(\widehat{\theta}, [J(\widehat{\theta})]^{-1}), \quad (2)$$

where

$$J(\theta) = -\frac{d^2}{d\theta^2} \log p(\theta|y).$$

It can be shown that the posterior mode is consistent for θ_0, so that as $n \to \infty$, the mass of the posterior distribution $p(\theta|y)$ becomes concentrated in smaller and smaller neighborhoods of θ_0 and the distance $|\widehat{\theta} - \theta_0|$ approaches zero.

Furthermore, we can rewrite the coefficient of the quadratic term in (1) as

$$\left[\frac{d^2}{d\theta^2} \log p(\theta|y)\right]_{\theta=\widehat{\theta}} = \left[\frac{d^2}{d\theta^2} \log \pi(\theta)\right]_{\theta=\widehat{\theta}} + \sum_{i=1}^{n} \left[\frac{d^2}{d\theta^2} \log p(y_i|\theta)\right]_{\theta=\widehat{\theta}}.$$

Considered as a function of θ, this coefficient is a constant plus the sum of n terms, each of whose expected value under the true sampling distribution of y_i, $p(y|\theta_0)$, is approximately $-I(\theta_0)$, as long as $\widehat{\theta}$ is close to θ_0 (we are assuming now that $f(y) = p(y|\theta_0)$ for some θ_0). Therefore, for large n, the curvature of the log posterior density can be approximated by the Fisher information, evaluated at either $\widehat{\theta}$ or θ_0 (where of course only the former is available in practice).

Example 2.2 (normal distribution with unknown mean and variance)**.** *We illustrate the approximate normal distribution with a simple theoretical example. Let $y_1, \dots, y_n$ be iid observations from a $\mathcal{N}(\mu, \sigma^2)$ distribution, and, for simplicity, we assume a uniform prior density for $(\mu, \log\sigma)$. We set up a normal approximation to the posterior distribution of $(\mu, \log\sigma)$, which has the virtue of restricting σ to positive values. To construct the approximation, we need the second derivatives of the log posterior density,*

$$\log p(\mu, \log\sigma|y) = constant - n\log\sigma - \frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\overline{y}-\mu)^2\right].$$

The first derivatives are

$$\begin{aligned}\frac{d}{d\mu}\log p(\mu, \log\sigma|y) &= \frac{n(\overline{y}-\mu)}{\sigma^2} \quad \textit{and} \\ \frac{d}{d(\log\sigma)}\log p(\mu, \log\sigma|y) &= -n + \frac{(n-1)s^2 + n(\overline{y}-\mu)^2}{\sigma^2},\end{aligned}$$

from which the posterior mode is readily obtained as

$$(\widehat{\mu}, \log\widehat{\sigma}) = \left(\overline{y}, \frac{1}{2}\log\left(\frac{n-1}{n}s^2\right)\right).$$

The second derivatives of the log posterior density are

$$\begin{aligned}\frac{d^2}{d\mu^2}\log p(\mu, \log\sigma|y) &= -\frac{n}{\sigma^2}, \\ \frac{d^2}{d\mu d(\log\sigma)}\log p(\mu, \log\sigma|y) &= -2n\frac{\overline{y}-\mu}{\sigma^2} \quad \textit{and} \\ \frac{d^2}{d(\log\sigma)^2}\log p(\mu, \log\sigma|y) &= -\frac{2}{\sigma^2}((n-1)s^2 + n(\overline{y}-\mu)^2).\end{aligned}$$

The matrix of second derivatives at the mode is then $\begin{pmatrix} -n/\widehat{\sigma}^2 & 0 \\ 0 & -2n \end{pmatrix}$. *From* (2), *the posterior distribution can be approximated as*

$$p(\mu, \log\sigma|y) \approx \mathcal{N}\left(\begin{pmatrix}\mu \\ \log\sigma\end{pmatrix}\middle|\begin{pmatrix}\overline{y} \\ \log\widehat{\sigma}\end{pmatrix}, \begin{pmatrix}\widehat{\sigma}^2/n & 0 \\ 0 & 1/(2n)\end{pmatrix}\right).$$

3 Asymptotic efficiency of Bayes estimators

Example 3.1. *Consider X binomial distributed $Bin(n,p)$. Then the Bayes estimator, with respect to a quadratic loss, of p corresponding to the beta prior $Beta(a,b)$ is $T_n(X) = (a+X)/(a+b+n)$.*
Thus

$$\sqrt{n}(T_n(X) - p) = \sqrt{n}\left(\frac{n}{a+b+n}\cdot\left(\frac{X}{n} - p\right) + \frac{1}{a+b+n}\cdot(a - p(a+b))\right).$$

This implies: $\sqrt{n}(T_n(X) - p)$ has the same limit distribution as $\sqrt{n}(X/n - p)$, namely, the normal distribution $\mathcal{N}(0, p(1-p))$.

Remarks 3.2. *The Bayes estimator is asymptotically efficient.* $\mathcal{N}(0, p(1-p))$ *is independent of the parameters of the prior distribution.*
This raises the question whether the same limit distribution is obtained when more general prior distributions are used and what happens in more general situations.

4 The principal result

Let $X_1, \ldots, X_n$ be iid with density $f(x_i|\theta)$ (with respect to μ), where θ is real valued and the parameter space Θ is an open interval. The true value of θ will be denoted by θ_0. Moreover, we write $P = P_{\theta_0}$.

The suitable conditions for our principal result (Theorem 4.3) are exactly the following (B1)-(B5) assumptions.

> (B1) The log likelihood function $l(\theta) = l(\theta|\mathbf{x}) = \sum_{i=1}^{n} \log f(x_i|\theta)$ satisfies the assumptions of [2, Theorem 2.6].

To motivate the next assumption, note that under the assumptions of [2, Theorem 2.6], if $\theta = \widetilde{\theta}_n$ is any sequence for which $\widetilde{\theta}_n \xrightarrow{P} \theta_0$ then

$$l(\theta) = l(\theta_0) + (\theta - \theta_0)l'(\theta_0) - \frac{1}{2}(\theta - \theta_0)^2[nI(\theta_0) + R_n(\theta)] \tag{3}$$

where

$$\frac{1}{n}R_n(\theta) \to^{P} 0 \text{ as } n \to \infty.$$

We require here the following stronger assumption.

> (B2) Given any $\varepsilon > 0$, there exists $\delta > 0$ such that in the expansion (3), the probability under θ_0 of the event
>
> $$\sup\left\{\left|\frac{1}{n}R_n(\theta)\right| : |\theta - \theta_0| \leq \delta\right\} \geq \varepsilon$$
>
> tends to zero as $n \to \infty$.

In the present case it is not enough to impose conditions on $l(\theta)$ in the neighborhood of θ_0, as is typically the case in asymptotic results. Since the Bayes estimators involve integration over the whole range of θ values, it is also necessary to control the behavior of $l(\theta)$ at a distance from θ_0.

> (B3) For any $\delta > 0$, there exists $\varepsilon > 0$ such that the probability under θ_0 of the event
>
> $$\sup\left\{\frac{1}{n}[l(\theta) - l(\theta_0)] : |\theta - \theta_0| \geq \delta\right\} \leq -\varepsilon$$
>
> tends to 1 as $n \to \infty$.

(B4) The prior density π of θ is continuous and positive for all $\theta \in \Theta$.

(B5) The expectation of θ under π exists, that is,

$$\int |\theta| \pi(\theta) d\theta < \infty.$$

To establish the asymptotic efficiency of Bayes estimators under these assumptions, we shall first prove that for large values of n, the posterior distribution of θ given the X's is approximately normal with

$$\text{mean} = \theta_0 + \frac{1}{nI(\theta_0)} l'(\theta_0) \quad \text{and variance } = \frac{1}{nI(\theta_0)}.$$

Theorem 4.1. *If $\pi^*(t|\boldsymbol{x})$ is the posterior density of $\sqrt{n}(\theta - T_n)$ where*

$$T_n = \theta_0 + \frac{1}{nI(\theta_0)} l'(\theta_0),$$

(i) then if (B1)-(B4) hold,

$$\int \left| \pi^*(t|\boldsymbol{x}) - \sqrt{I(\theta_0)}\phi[t\sqrt{I(\theta_0)}] \right| dt \to^P 0. \tag{4}$$

(ii) If, in addition, (B5) holds, then

$$\int (1 + |t|) \left| \pi^*(t|\boldsymbol{x}) - \sqrt{I(\theta_0)}\phi[t\sqrt{I(\theta_0)}] \right| dt \to^P 0. \tag{5}$$

Proof. (i) By the definition of T_n,

$$\begin{aligned} \pi^*(t|\mathbf{x}) &= \frac{\pi\left(T_n + \frac{t}{\sqrt{n}}\right) \exp\left[l\left(T_n + \frac{t}{\sqrt{n}}\right)\right]}{\int \pi\left(T_n + \frac{u}{\sqrt{n}}\right) \exp\left[l\left(T_n + \frac{u}{\sqrt{n}}\right)\right] du} & (6) \\ &= e^{\omega(t)} \pi\left(T_n + \frac{t}{\sqrt{n}}\right) / C_n \end{aligned}$$

where

$$\omega(t) = l\left(T_n + \frac{t}{\sqrt{n}}\right) - l(\theta_0) - \frac{1}{2nI(\theta_0)}[l'(\theta_0)]^2 \tag{7}$$

and

$$C_n = \int e^{\omega(u)} \pi\left(T_n + \frac{u}{\sqrt{n}}\right) du.$$

Using assumptions (B1)-(B4), and the following characterization of $\omega(t)$,

Lemma 4.2. *The quantity $\omega(t)$, defined by (7), is equal to*

$$\omega(t) = -I(\theta_0)\frac{t^2}{2} - \frac{1}{2n} R_n\left(T_n + \frac{t}{\sqrt{n}}\right)\left[t + \frac{1}{I(\theta_0)\sqrt{n}} l'(\theta_0)\right]^2$$

where R_n is the function defined in (3), one can show that

$$J_1 = \int \left| e^{\omega(t)} \pi \left(T_n + \frac{t}{\sqrt{n}} \right) - e^{-t^2 I(\theta_0)/2} \pi(\theta_0) \right| dt \to^P 0, \tag{8}$$

and therefore that

$$C_n \to^P \int e^{-t^2 I(\theta_0)/2} \pi(\theta_0) dt = \pi(\theta_0) \sqrt{2\pi / I(\theta_0)}. \tag{9}$$

The left side of (4) is equal to J/C_n, where

$$J = \int \left| e^{\omega(t)} \pi \left(T_n + \frac{t}{\sqrt{n}} \right) - C_n \sqrt{I(\theta_0)} \phi[t\sqrt{I(\theta_0)}] \right| dt$$

and, by (9), it is enough to show that $J \to^P 0$.

Now, $J \leq J_1 + J_2$ where J_1 is given by (8) and

$$\begin{aligned} J_2 &= \int \left| C_n \sqrt{I(\theta_0)} \phi \left[t\sqrt{I(\theta_0)} \right] - \exp\left[-\frac{t^2}{2} I(\theta_0) \right] \pi(\theta_0) \right| dt \\ &= \left| \frac{C_n \sqrt{I(\theta_0)}}{\sqrt{2\pi}} - \pi(\theta_0) \right| \int \exp\left[-\frac{t^2}{2} I(\theta_0) \right] dt. \end{aligned}$$

By (8) and (9), J_1 and J_2 tend to zero in probability, and this completes the proof of part (i).

(ii) The left side of (5) is equal to

$$\frac{1}{C_n} J' \leq \frac{1}{C_n} (J_1' + J_2')$$

where J', J_1', and J_2' are obtained from J , J_1, and J_2, respectively, by inserting the factor $(1+|t|)$ under the integral signs. It is therefore enough to prove that J_1' and J_2' both tend to zero in probability. The proof for J_2' is the same as that for J_2; the proof for J_1' follows from (8) and assumption (B5).

□

On the basis of Theorem 4.1, we are now able to prove the principal result of this section.

Theorem 4.3. *If (B1)-(B5) hold, and if $\widetilde{\theta}_n$ is the Bayes estimator when the prior density is π and the loss is squared error, then*

$$\sqrt{n}(\widetilde{\theta}_n - \theta_0) \to^d \mathcal{N}(0, 1/I(\theta_0)),$$

so that $\widetilde{\theta}_n$ is consistent for θ_0 and asymptotically efficient.

Proof. We have

$$\sqrt{n}(\widetilde{\theta_n} - \theta_0) = \sqrt{n}(\widetilde{\theta}_n - T_n) + \sqrt{n}(T_n - \theta_0).$$

By the CLT, the second term has the limit distribution $\mathcal{N}(0, 1/I(\theta_0))$, so that it only remains to show that

$$\sqrt{n}(\widetilde{\theta}_n - T_n) \rightarrow^P 0.$$

Note that Equation (6) says that $\pi^*(t|\mathbf{x}) = \frac{1}{\sqrt{n}}\pi(T_n + \frac{t}{\sqrt{n}}|\mathbf{x})$, and, hence, by a change of variable, we have

$$\begin{aligned} \widetilde{\theta}_n &= \int \theta\pi(\theta|\mathbf{x})d\theta \\ &= \int \left(\frac{t}{\sqrt{n}} + T_n\right)\pi^*(t|\mathbf{x})dt \\ &= \frac{1}{\sqrt{n}}\int t\pi^*(t|\mathbf{x})dt + T_n \end{aligned}$$

and hence

$$\sqrt{n}(\widetilde{\theta}_n - T_n) = \int t\pi^*(t|\mathbf{x})dt.$$

Now, since $\int t\sqrt{I(\theta_0)}\phi[t\sqrt{I(\theta_0)}]dt = 0$,

$$\begin{aligned} \sqrt{n}|\widetilde{\theta}_n - T_n| &= \left|\int t\pi^*(t|\mathbf{x})dt - \int t\sqrt{I(\theta_0)}\phi\left[t\sqrt{I(\theta_0)}\right]dt\right| \\ &\leq \int |t|\left|\pi^*(t|\mathbf{x}) - \sqrt{I(\theta_0)}\phi\left[t\sqrt{I(\theta_0)}\right]\right|dt. \end{aligned}$$

which tends to zero in probability by Theorem 4.1. □

Observation 4.4. *Assumptions (B1)-(B5) are satisfied in exponential families.*

5 Bibliography

References

[1] Andrew Gelman, John B. Carlin, Hal S. Stern & Donald B. Rubin, *Bayesian Data Analysis*, Second Edition, Texts in Statistial Science, Chapman & Hall/CRC, 2003.

[2] E. L. Lehmann & George Casella, *Theory of Point Estimation*, Second Edition, New York, Springer-Verlag, 1998.

[3] Van de Geer Sara, Lecture notes on ”Mathematische Grundlagen der Statistik”, autumn semester 2007, [`http://stat.ethz.ch/teaching/lectures/HS_2007/math_stat/Sara_lecture_note_v2.pdf`].