

Stat 710: Mathematical Statistics

Lecture 5

Jun Shao

Department of Statistics
University of Wisconsin
Madison, WI 53706, USA

Lecture 5: Bayes estimators in normal models and MCMC

We first consider two examples of Bayes estimators in normal models

Example 4.8

Let X_1, \dots, X_n be i.i.d. from $N(\mu, \sigma^2)$ with unknown $\mu \in \mathcal{R}$ and $\sigma^2 > 0$. Let the prior for $\omega = (2\sigma^2)^{-1}$ be the gamma distribution $\Gamma(\alpha, \gamma)$ with known α and γ and the prior for μ be $N(\mu_0, \sigma_0^2/\omega)$ (conditional on ω). Then the posterior p.d.f. of (μ, ω) is proportional to

$$\omega^{(n+1)/2+\alpha-1} \exp \left\{ - \left[\gamma^{-1} + Y + n(\bar{X} - \mu)^2 + \frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right] \omega \right\},$$

where $Y = \sum_{i=1}^n (X_i - \bar{X})^2$ and \bar{X} is the sample mean.

Note that

$$n(\bar{X} - \mu)^2 + \frac{(\mu - \mu_0)^2}{2\sigma_0^2} = \left(n + \frac{1}{2\sigma_0^2} \right) \mu^2 - 2 \left(n\bar{X} + \frac{\mu_0}{2\sigma_0^2} \right) \mu + n\bar{X}^2 + \frac{\mu_0^2}{2\sigma_0^2}.$$

Lecture 5: Bayes estimators in normal models and MCMC

We first consider two examples of Bayes estimators in normal models

Example 4.8

Let X_1, \dots, X_n be i.i.d. from $N(\mu, \sigma^2)$ with unknown $\mu \in \mathcal{R}$ and $\sigma^2 > 0$. Let the prior for $\omega = (2\sigma^2)^{-1}$ be the gamma distribution $\Gamma(\alpha, \gamma)$ with known α and γ and the prior for μ be $N(\mu_0, \sigma_0^2/\omega)$ (conditional on ω). Then the posterior p.d.f. of (μ, ω) is proportional to

$$\omega^{(n+1)/2+\alpha-1} \exp \left\{ - \left[\gamma^{-1} + Y + n(\bar{X} - \mu)^2 + \frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right] \omega \right\},$$

where $Y = \sum_{i=1}^n (X_i - \bar{X})^2$ and \bar{X} is the sample mean.

Note that

$$n(\bar{X} - \mu)^2 + \frac{(\mu - \mu_0)^2}{2\sigma_0^2} = \left(n + \frac{1}{2\sigma_0^2} \right) \mu^2 - 2 \left(n\bar{X} + \frac{\mu_0}{2\sigma_0^2} \right) \mu + n\bar{X}^2 + \frac{\mu_0^2}{2\sigma_0^2}.$$

Example 4.8 (continued)

Hence, the posterior p.d.f. of (μ, ω) is proportional to

$$\omega^{(n+1)/2+\alpha-1} \exp \left\{ - \left[\gamma^{-1} + W + \left(n + \frac{1}{2\sigma_0^2} \right) (\mu - \zeta(X))^2 \right] \omega \right\},$$

where

$$\zeta(X) = \frac{n\bar{X} + \frac{\mu_0}{2\sigma_0^2}}{n + \frac{1}{2\sigma_0^2}} \quad \text{and} \quad W = Y + n\bar{X}^2 + \frac{\mu_0^2}{2\sigma_0^2} - \left(n + \frac{1}{2\sigma_0^2} \right) [\zeta(X)]^2.$$

Thus, the posterior of ω is the gamma distribution

$\Gamma(n/2 + \alpha, (\gamma^{-1} + W)^{-1})$ and the posterior of μ (given ω and X) is $N(\zeta(X), [(2n + \sigma_0^{-2})\omega]^{-1})$.

Under the squared error loss, the Bayes estimator of μ is $\zeta(X)$ and the Bayes estimator of $\sigma^2 = (2\omega)^{-1}$ is $(\gamma^{-1} + W)/(n + 2\alpha - 2)$, provided that $n + 2\alpha > 2$.

Apparently, these Bayes estimators are biased but the biases are of the order n^{-1} ; and they are consistent as $n \rightarrow \infty$.

To consider the next example, we need the following useful lemma.

Lemma 4.1

Suppose that X has a p.d.f. $f_\theta(x)$ w.r.t. a σ -finite measure ν .
Suppose that $\theta = (\theta_1, \theta_2)$, $\theta_j \in \Theta_j$, and that the prior has a p.d.f.

$$\pi(\theta) = \pi_{\theta_1|\theta_2}(\theta_1)\pi_{\theta_2}(\theta_2),$$

where $\pi_{\theta_2}(\theta_2)$ is a p.d.f. w.r.t. a σ -finite measure ν_2 on Θ_2 and for any given θ_2 , $\pi_{\theta_1|\theta_2}(\theta_1)$ is a p.d.f. w.r.t. a σ -finite measure ν_1 on Θ_1 .

Suppose further that if θ_2 is given, the Bayes estimator of $h(\theta_1) = g(\theta_1, \theta_2)$ under the squared error loss is $\delta(X, \theta_2)$.

Then the Bayes estimator of $g(\theta_1, \theta_2)$ under the squared error loss is $\delta(X)$ with

$$\delta(x) = \int_{\Theta_2} \delta(x, \theta_2) p_{\theta_2|x}(\theta_2) d\nu_2,$$

where $p_{\theta_2|x}(\theta_2)$ is the posterior p.d.f. of $\vec{\theta}_2$ given $X = x$.

To consider the next example, we need the following useful lemma.

Lemma 4.1

Suppose that X has a p.d.f. $f_\theta(x)$ w.r.t. a σ -finite measure ν .
Suppose that $\theta = (\theta_1, \theta_2)$, $\theta_j \in \Theta_j$, and that the prior has a p.d.f.

$$\pi(\theta) = \pi_{\theta_1|\theta_2}(\theta_1)\pi_{\theta_2}(\theta_2),$$

where $\pi_{\theta_2}(\theta_2)$ is a p.d.f. w.r.t. a σ -finite measure ν_2 on Θ_2 and for any given θ_2 , $\pi_{\theta_1|\theta_2}(\theta_1)$ is a p.d.f. w.r.t. a σ -finite measure ν_1 on Θ_1 .

Suppose further that if θ_2 is given, the Bayes estimator of $h(\theta_1) = g(\theta_1, \theta_2)$ under the squared error loss is $\delta(X, \theta_2)$.

Then the Bayes estimator of $g(\theta_1, \theta_2)$ under the squared error loss is $\delta(X)$ with

$$\delta(x) = \int_{\Theta_2} \delta(x, \theta_2) p_{\theta_2|x}(\theta_2) d\nu_2,$$

where $p_{\theta_2|x}(\theta_2)$ is the posterior p.d.f. of $\vec{\theta}_2$ given $X = x$.

Example 4.9

Consider a linear model

$$X_{ij} = \beta^\tau Z_i + \varepsilon_{ij}, \quad j = 1, \dots, n_i, \quad i = 1, \dots, k,$$

where $\beta \in \mathcal{R}^p$ is unknown, Z_i 's are known vectors, ε_{ij} 's are independent, and ε_{ij} is $N(0, \sigma_i^2)$, $j = 1, \dots, n_i$, $i = 1, \dots, k$.

Let X be the sample vector containing all X_{ij} 's.

The parameter vector is $\theta = (\beta, \omega)$, $\omega = (\omega_1, \dots, \omega_k)$ and $\omega_i = (2\sigma_i^2)^{-1}$. Assume the prior for θ has the Lebesgue p.d.f.

$$c \pi(\beta) \prod_{i=1}^k \omega_i^\alpha e^{-\omega_i/\gamma},$$

where $\alpha > 0$, $\gamma > 0$, and $c > 0$ are known constants and $\pi(\beta)$ is a known Lebesgue p.d.f. on \mathcal{R}^p .

The posterior p.d.f. of θ is then proportional to

$$h(X, \theta) = \pi(\beta) \prod_{i=1}^k \omega_i^{n_i/2 + \alpha} e^{-[\gamma^{-1} + v_i(\beta)]\omega_i},$$

where $v_i(\beta) = \sum_{j=1}^{n_i} (X_{ij} - \beta^\tau Z_i)^2$.

Example 4.9 (continued)

If β is known, the Bayes estimator of σ_i^2 under the squared error loss is

$$\int \frac{1}{2\omega_i} \frac{h(X, \theta)}{\int h(X, \theta) d\omega} d\omega = \frac{\gamma^{-1} + v_i(\beta)}{2\alpha + n_i}.$$

By Lemma 4.1, the Bayes estimator of σ_i^2 is

$$\hat{\sigma}_i^2 = \int \frac{\gamma^{-1} + v_i(\beta)}{2\alpha + n_i} f_{\beta|X}(\beta) d\beta,$$

where

$$\begin{aligned} f_{\beta|X}(\beta) &\propto \int h(X, \theta) d\omega \\ &\propto \pi(\beta) \prod_{i=1}^k \int \omega_i^{\alpha+n_i/2} e^{-[\gamma^{-1}+v_i(\beta)]\omega_i} d\omega_i \\ &\propto \pi(\beta) \prod_{i=1}^k [\gamma^{-1} + v_i(\beta)]^{-(\alpha+1+n_i/2)} \end{aligned}$$

is the posterior p.d.f. of β .

Example 4.9 (continued)

The Bayes estimator of $l^\tau \beta$ for any $l \in \mathcal{R}^p$ is then the posterior mean of $l^\tau \beta$ w.r.t. the p.d.f. $f_{\beta|X}(\beta)$.

In this problem, Bayes estimators do not have explicit forms.

A numerical method has to be used to evaluate Bayes estimators (see Example 4.10).

Let \bar{X}_i and S_i^2 be the sample mean and variance of X_{ij} , $j = 1, \dots, n_i$ (S_i^2 is defined to be 0 if $n_i = 1$)

Let $\sigma_0^2 = (2\alpha\gamma)^{-1}$ (the prior mean of σ_i^2).

Then the Bayes estimator $\hat{\sigma}_i^2$ can be written as

$$\frac{2\alpha}{2\alpha + n_i} \sigma_0^2 + \frac{n_i - 1}{2\alpha + n_i} S_i^2 + \frac{n_i}{2\alpha + n_i} \int (\bar{X}_i - \beta^\tau Z_i)^2 f_{\beta|X}(\beta) d\beta.$$

This Bayes estimator is a weighted average of prior information, "within group" variation, and averaged squared "residuals".

Markov chain Monte Carlo (MCMC)

Often, Bayes actions or estimators have to be computed numerically. Typically we need to compute

$$E_p(g) = \int_{\Theta} g(\theta) p(\theta) d\nu$$

with some function g , where $p(\theta)$ is a p.d.f. w.r.t. a σ -finite measure ν on $(\Theta, \mathcal{B}_{\Theta})$ and $\Theta \subset \mathcal{R}^k$.

If g is an indicator function of $A \in \mathcal{B}_{\Theta}$ and $p(\theta)$ is the posterior p.d.f. of θ given $X = x$, then $E_p(g)$ is the posterior probability of A .

There are many numerical methods for computing integrals $E_p(g)$.

The simple Monte Carlo method

Generate i.i.d. $\theta^{(1)}, \dots, \theta^{(m)}$ from a p.d.f. $h(\theta) > 0$ w.r.t. ν .

By the SLLN, as $m \rightarrow \infty$,

$$\hat{E}_p(g) = \frac{1}{m} \sum_{j=1}^m \frac{g(\theta^{(j)}) p(\theta^{(j)})}{h(\theta^{(j)})} \rightarrow_{a.s.} \int_{\Theta} \frac{g(\theta) p(\theta)}{h(\theta)} h(\theta) d\nu = E_p(g).$$

Hence $\hat{E}_p(g)$ is a numerical approximation to $E_p(g)$.

Markov chain Monte Carlo (MCMC)

Often, Bayes actions or estimators have to be computed numerically. Typically we need to compute

$$E_p(g) = \int_{\Theta} g(\theta) p(\theta) d\nu$$

with some function g , where $p(\theta)$ is a p.d.f. w.r.t. a σ -finite measure ν on $(\Theta, \mathcal{B}_{\Theta})$ and $\Theta \subset \mathcal{R}^k$.

If g is an indicator function of $A \in \mathcal{B}_{\Theta}$ and $p(\theta)$ is the posterior p.d.f. of θ given $X = x$, then $E_p(g)$ is the posterior probability of A .

There are many numerical methods for computing integrals $E_p(g)$.

The simple Monte Carlo method

Generate i.i.d. $\theta^{(1)}, \dots, \theta^{(m)}$ from a p.d.f. $h(\theta) > 0$ w.r.t. ν .

By the SLLN, as $m \rightarrow \infty$,

$$\hat{E}_p(g) = \frac{1}{m} \sum_{j=1}^m \frac{g(\theta^{(j)}) p(\theta^{(j)})}{h(\theta^{(j)})} \rightarrow_{a.s.} \int_{\Theta} \frac{g(\theta) p(\theta)}{h(\theta)} h(\theta) d\nu = E_p(g).$$

Hence $\hat{E}_p(g)$ is a numerical approximation to $E_p(g)$.

The simple Monte Carlo method may not work well because

- the convergence of $\hat{E}_p(g)$ is very slow when k (the dimension of Θ) is large
- generating a random vector from some k -dimensional distribution may be difficult, if not impossible.

More sophisticated MCMC methods

Different from the simple Monte Carlo in two aspects:

- generating random vectors can be done using distributions whose dimensions are much lower than k
- $\theta^{(1)}, \dots, \theta^{(m)}$ are not independent, but form a homogeneous Markov chain.

Many MCMC methods were developed in the last 20 years

We only consider one of them as an example

The simple Monte Carlo method may not work well because

- the convergence of $\hat{E}_p(g)$ is very slow when k (the dimension of Θ) is large
- generating a random vector from some k -dimensional distribution may be difficult, if not impossible.

More sophisticated MCMC methods

Different from the simple Monte Carlo in two aspects:

- generating random vectors can be done using distributions whose dimensions are much lower than k
- $\theta^{(1)}, \dots, \theta^{(m)}$ are not independent, but form a homogeneous Markov chain.

Many MCMC methods were developed in the last 20 years
We only consider one of them as an example

The simple Monte Carlo method may not work well because

- the convergence of $\hat{E}_p(g)$ is very slow when k (the dimension of Θ) is large
- generating a random vector from some k -dimensional distribution may be difficult, if not impossible.

More sophisticated MCMC methods

Different from the simple Monte Carlo in two aspects:

- generating random vectors can be done using distributions whose dimensions are much lower than k
- $\theta^{(1)}, \dots, \theta^{(m)}$ are not independent, but form a homogeneous Markov chain.

Many MCMC methods were developed in the last 20 years

We only consider one of them as an example

Gibbs sampler

Let $y = (y_1, y_2, \dots, y_d)$. (y_j 's may be vectors with different dimensions)

At step $t = 1, 2, \dots$, given $y^{(t-1)}$, generate

$y_1^{(t)}$ from $P(y_2^{(t-1)}, \dots, y_d^{(t-1)} | y_1^{(t-1)})$, \dots ,

$y_j^{(t)}$ from $P(y_1^{(t)}, \dots, y_{j-1}^{(t)}, y_{j+1}^{(t-1)}, \dots, y_k^{(t-1)} | y_j^{(t-1)})$, \dots ,

$y_k^{(t)}$ from $P_k(y_1^{(t)}, \dots, y_{k-1}^{(t)} | y_k^{(t-1)})$.

Example 4.10

Consider Example 4.9 (normal linear model).

Under the given prior for $\theta = (\beta, \omega)$, it is difficult to generate random vectors directly from the posterior p.d.f.

$$p(\theta) \propto \pi(\beta) \prod_{i=1}^k \omega_i^{n_i/2 + \alpha} e^{-[\gamma^{-1} + v_i(\beta)]\omega_i},$$

which does not have a familiar form.

To apply a Gibbs sampler with $y = \theta$, $y_1 = \beta$, and $y_2 = \omega$, we need to generate random vectors from the posterior of β , given x and ω , and the posterior of ω , given x and β .

Gibbs sampler

Let $y = (y_1, y_2, \dots, y_d)$. (y_j 's may be vectors with different dimensions)

At step $t = 1, 2, \dots$, given $y^{(t-1)}$, generate

$y_1^{(t)}$ from $P(y_2^{(t-1)}, \dots, y_d^{(t-1)} | y_1^{(t-1)})$, \dots ,

$y_j^{(t)}$ from $P(y_1^{(t)}, \dots, y_{j-1}^{(t)}, y_{j+1}^{(t-1)}, \dots, y_k^{(t-1)} | y_j^{(t-1)})$, \dots ,

$y_k^{(t)}$ from $P_k(y_1^{(t)}, \dots, y_{k-1}^{(t)} | y_k^{(t-1)})$.

Example 4.10

Consider Example 4.9 (normal linear model).

Under the given prior for $\theta = (\beta, \omega)$, it is difficult to generate random vectors directly from the posterior p.d.f.

$$p(\theta) \propto \pi(\beta) \prod_{i=1}^k \omega_i^{n_i/2 + \alpha} e^{-[\gamma^{-1} + v_i(\beta)]\omega_i},$$

which does not have a familiar form.

To apply a Gibbs sampler with $y = \theta$, $y_1 = \beta$, and $y_2 = \omega$, we need to generate random vectors from the posterior of β , given x and ω , and the posterior of ω , given x and β .

Example 4.10 (continued)

Since

$$p(\theta) \propto \pi(\beta) \prod_{i=1}^k \omega_i^{n_i/2 + \alpha} e^{-[\gamma^{-1} + v_i(\beta)]\omega_i},$$

the posterior of $\omega = (\omega_1, \dots, \omega_k)$, given x and β , is a product of marginals of ω_i 's that are the gamma distributions $\Gamma(\alpha + 1 + n_i/2, [\gamma^{-1} + v_i(\beta)]^{-1})$, $i = 1, \dots, k$.

Assume now that $\pi(\beta) \equiv 1$ (noninformative prior for β).

The posterior p.d.f. of β , given x and ω , is proportional to

$$\prod_{i=1}^k e^{-\omega_i v_i(\beta)} \propto e^{-\|W^{1/2}Z\beta - W^{1/2}X\|^2},$$

where W is the diagonal block matrix whose i th block is $\omega_i I_{n_i}$.

Let $n = \sum_{i=1}^k n_i$.

The posterior of $W^{1/2}Z\beta$, given X and ω , is $N_n(W^{1/2}X, 2^{-1}I_n)$ and the posterior of β , given X and ω , is $N_p((Z^\tau WZ)^{-1}Z^\tau WX, 2^{-1}(Z^\tau WZ)^{-1})$ ($Z^\tau WZ$ is assumed of full rank for simplicity), since

$$\beta = [(Z^\tau WZ)^{-1}Z^\tau W^{1/2}]W^{1/2}Z\beta.$$

Random generation using these two posterior distributions is easy.