Bayesian Inference for Surveys

Roderick Little and Trivellore Raghunathan

Module 6: Computational Methods

- A Bayesian analysis uses the entire posterior distribution of the parameter of interest.
- Summaries of the posterior distribution are used for statistical inferences
  - Means, Median, Modes or measures of central tendency
  - Standard deviation, mean absolute deviation or measures of spread
  - Percentiles or intervals
- Conceptually, all these quantities can be expressed analytically in terms of integrals of functions of parameter with respect to its posterior distribution
- Computations
  - Numerical integration routines
  - Simulation techniques

Numerical Integration

- Mean $\int_a^b \theta \pi(\theta \mid Data)\, d\theta$
- Variance $\int_a^b \theta^2 \pi(\theta \mid Data)\, d\theta - \left[\int_a^b \theta \pi(\theta \mid Data)\, d\theta \right]^2$
- Probability

$$\Pr(\theta \le c \mid Data) = \int_a^b I_{[\theta \le c]} \pi(\theta \mid Data)\, d\theta$$

$$I_{[x \le y]} = 1 \text{ if } x \le y \text{ and } 0 \text{ otherwise}$$

- Gaussian quadrature approximates

$$\int_a^b w(x) f(x) dx = \sum_{i=1}^{n} w_i f(x_i);$$

$$\int_a^b f(x) dx = \int_a^b w(x) \times (f(x) / w(x)) dx$$

x_i = Roots of the polynomials of order n

w_i = Weight function evaluated at the roots

$a = 0, b = \infty$: Polynomial=Laguerre, $w(x) = x^{\alpha} e^{-x}, \alpha > -1$

$a = -\infty, b = \infty$: Polynomial=Hermite, $w(x) = e^{-x^2}$

$a = -1, b = 1$: Polynomial=Jacobi,

$w(x) = (1-x)^{\alpha} + (1+x)^{\beta}, \alpha, \beta > -1$

$a = -1, b = 1$: Polynomial=Legendre, $w(x) = 1$

- Abramovitz and Stegun give a table of values of the weight and abscissa.
- These can be computed in R. Download and install package “statmod” from the r-project web site.
- After installation, use the command
  - library(“statmod”)
  - gauss.quad(n,polynomial=,a=,b=)
  - See R manual for more help
- One can use simple SAS macro or even an Excel spread sheet to do these computations.

Example

$$\int_0^{\infty} x^2 e^{-x^2}\,dx = \frac{1}{2}\int_{-\infty}^{\infty} x^2 e^{-x^2}\,dx = \int_0^{\infty} x^2 e^{-x} e^{-x^2+x}\,dx$$

Substitute $u = x^2$,

$$\frac{1}{2}\int_0^{\infty} u^{1/2} e^{-u}\,du$$

```
y=gauss.quad(10,kind="hermite")
> w=y$weight
> x=y$nodes
> a=sum(w*x*x)/2
> a
[1] 0.4431135
> sqrt(pi)/4
[1] 0.4431135
```

```
> y=gauss.quad(10,kind="laguerre",alpha=0.5)
> w=y$weight
> a=sum(w)/2
> a
> [1] 0.4431135
```

Types of Simulation

- Direct simulation (Binomial and normal examples)
- Approximate direct simulation
  - Discrete approximation of the posterior density
  - Rejection sampling
  - Sampling Importance Resampling
- Iterative simulation techniques
  - Gibbs sampler
  - Metropolis Algorithm

Simulation Techniques

- Numerical integration though can be extended to multidimensional integrals but can be quite time consuming.
- Error in approximation can be large
- Alternative is to draw samples from the posterior distribution and use the sample to characterize the features of the posterior distribution

$X \sim F$

Density : $f(x)$

- Objective:Compute $E(t(X))$

$x_1, x_2, \cdots x_K \sim$ draws from $f(x)$

$$E(t(X)) \approx \bar{t} = \frac{1}{K}\sum_{i=1}^{K} t_i, t_i = t(x_i)$$

Monte-Carlo Error

(in the approximation)

$$e = \sqrt{\frac{1}{K(K-1)}\sum_{i=1}^{K}(t_i - \bar{t})^2}$$

$$\Pr(t(X) \geq t_o) \approx p_o = \frac{1}{K}\sum_{i=1}^{K} I_{[t_i \geq t_o]}$$

$I_A = 1$ if A is true

$= 0$ otherwise

$$MCSE = \sqrt{p_o(1-p_o)/K}$$

Estimation of distribution function

Estimation of percentiles

$$\Pr(t(X) \leq ?) = p_o$$

Order statistics : $t_{(1)} \leq t_{(2)} \leq \mathrm{L} \leq t_{(K)}$

$$[Kp_o] \leq Kp_o \leq [Kp_o]+1$$

$$? \approx t_{([Kp_o])}(Kp_o - [Kp_o]) + t_{([Kp_o]+1)}([Kp_o]+1-Kp_o)$$

- Equal tail probability interval

$$\Pr(t(X) \leq ?_L) = \alpha / 2$$

$$\Pr(t(X) \leq ?_U) = 1 - \alpha / 2$$

- Highest posterior density interval (approximation)
  - Smooth density estimates and then compute highest HPD interval
  - Numerical approximation (assuming that K is large)

Unimodal

- Order the values and construct intervals

$$t_{(1)} \le t_{(2)} \le \cdots \le t_{(K)}$$

$$R_j : (t_{(j)}, t_{([j+(1-\alpha)K])})$$

- Each R_j is an posterior interval
- Choose the interval that is shortest
- Need a more general approach for multimodal situation. See R manual

Simulation for the Normal Example

- Revisit normal example

$$\sigma^2 \mid y_{\text{inc}} \sim (n-1)s^2 / \chi^2_{n-1}$$

$$\mu \mid \sigma^2, y_{\text{inc}} \sim N(\bar{y}, \sigma^2 / n)$$

$$\bar{Y}_k \mid \mu, \sigma^2, y_{\text{inc}} \sim N(\mu, \sigma^2 / k)$$

```
# Draws for the normal case
sampsize=20
k=5
ybar=10
ssquare=5
nsimul=1000
result=matrix(0,nsimul,3)
for (i in 1:nsimul){
tmp=rnorm(sampsize-1)
chisq=sum(tmp*tmp)
sigmasq=(sampsize-1)*ssquare/chisq;
mu=ybar+sqrt(sigmasq/
sampsize)*rnorm(1)
ybark=mu+sqrt(sigmasq/k)*rnorm(1)
result[i,1]=sigmasq
result[i,2]=mu
result[i,3]=ybark}
```

Multivariate Example

- In an investigation several versions of a question asking about an outcome *Y* were to be investigated. The true values of *Y* were known for a sample of subjects.
- The *m* versions of the questions were administered to the same sample resulting in measurements $x_1, x_2, x_3 \ldots, x_{\mathrm{m}}$
- Objective is to infer about the largest of the *m* correlation coefficients

$$\rho_{y,x_j}; j = 1, 2, \ldots, m$$

Example: Model

- Suppose that these measures are continuous and a multivariate normal model is posited:

$$U = (Y, X_1, X_2, ..., X_m) \sim MVN_{m+1}(\mu, \Sigma)$$

$$\pi(\mu, \Sigma) \propto |\Sigma^{-1}|^{-(m+1)/2}$$

- It is analytically difficult to derive the posterior distribution of $\theta = \underset{1 \le j \le m}{\text{Max}}(\rho_{y,x_j})$
- Even more interesting is to find the posterior mean of

$$\lambda_j = \Pr(\rho_{y,x_j} \ge \rho_{y,x_i} \forall i \ne j)$$

- Likelihood

$$\prod_{i=1}^{n} |\Sigma|^{-1/2} \exp[-(U_i - \mu)^t \Sigma^{-1} (U_i - \mu)/2]$$

$$= |\Sigma|^{-n/2} \exp\left[-\sum_i (U_i - \bar{U})^t \Sigma^{-1} (U_i - \bar{U})/2\right] \times$$

$$\exp\left[-n(\mu - \bar{U})^t \Sigma^{-1} (\mu - \bar{U})/2\right]$$

- Posterior distribution

$$\left[|\Sigma^{-1}|^{(n-m-2)/2} \exp\left[-Tr(S\Sigma^{-1})/2\right]\right] \times$$

$$\left[|\Sigma/n|^{-1/2} \exp\left[-(\mu - \bar{U})^t (\Sigma/n)^{-1} (\mu - \bar{U})/2\right]\right]$$

Wishart and Inverse-Wishart Distributions

$Z =$ Positive definite symmetric random matrix of dimension p with $p(p+1)/2$ distinct random variables.

Z has a Wishart distribution if

$$pdf(Z) = C\,|B|^{-\nu/2}|Z|^{(\nu-p-1)/2}\exp[-Tr(B^{-1}Z)/2]$$

$$C^{-1} = 2^{\nu p/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma((\nu+1-i)/2)$$

$$Z \sim Wishart(B,\nu)$$

$$U \sim Inv-Wishart(B,\nu) \text{ if } U^{-1} \sim Wishart(B,\nu)$$

Example:Simulation

- It is easy to simulate from the posterior distribution of m and S.

$$\Sigma^{-1} \mid \text{Data} \sim \text{Wishart}(S^{-1}, n-1)$$

$$S = \sum_{i=1}^{n} (u_i - \bar{u})(u_i - \bar{u})^t$$

$$\bar{u} = \sum_i u_i / n$$

Generate $z_j \sim N(0, S^{-1}); j = 1, 2, \ldots, n-1$

Define $\Sigma_*^{-1} = \sum_j z_j z_j^t$

Generate $\mu_* \sim N(\bar{u}, \Sigma_* / n)$

- Also $\mu \mid \text{Data}, \Sigma \sim N(\bar{u}, \Sigma / n)$
- Compute the desired function of (μ_*, Σ_*)
- Repeat the above steps to simulate several draws from the posterior distribution.

Approximate Direct Simulation

- Approximating the posterior distribution by a normal distribution by matching the posterior mean and variance.
  - Posterior mean and variance computed using numerical integration techniques
- An alternative is to use the mode and a measure of curvature at the mode
  - Mode and the curvature can be computed using many different methods
- Approximate the posterior distribution using a grid of values of the parameter and compute the posterior density at each grid and then draw values from the grid with probability proportional to the posterior density

Normal Approximation

Posterior density : $\pi(\theta \mid x)$

Easy to work with log-posterior density

$l(\theta) = \log(\pi(\theta \mid x))$

At the mode, $f(\theta) = l'(\theta) = 0$

Curvature : $f'(\theta) = l''(\theta)$

For logarithm of the normal density

Mode is the mean and

the curvature at the mode

is negative of the precision

(Precision:reciprocal of variance)

Rejection Sampling

- Actual Density from which to draw from
- Candidate density from which it is easy to draw
- The importance ratio is bounded
- Sample q from g, accept q with probability p otherwise redraw from g

$$\pi(\theta \mid \text{data})$$

$$g(\theta), \text{ with } g(\theta) > 0 \text{ for all } \theta$$

$$\text{with } \pi(\theta \mid \text{data}) > 0$$

$$\frac{\pi(\theta \mid \text{data})}{g(\theta)} \leq M$$

$$p = \frac{\pi(\theta \mid \text{data})}{M \times g(\theta)}$$

Sampling Importance Resampling

- Target density from which to draw

$$\pi(\theta \mid \text{data})$$

- Candidate density from which it is easy to draw

$$g(\theta), \text{ such that } g(\theta) > 0$$

$$\text{for all } \theta \text{ with } \pi(\theta \mid \text{data}) > 0$$

- The importance ratio

$$w(\theta) \propto \frac{\pi(\theta \mid \text{data})}{g(\theta)}$$

- Sample M values of q from g

$$\theta_1^*, \theta_2^*, \ldots, \theta_M^*$$

- Compute the M importance ratios and resample with probability proportional to the importance ratios.

$$w(\theta_i^*); i = 1,2,\ldots, M$$

Markov Chain Simulation

- In real problems it may be hard to apply direct or approximate direct simulation techniques.
- The Markov chain methods involve a random walk in the parameter space which converges to a stationary distribution that is the target posterior distribution.
  - Metropolis-Hastings algorithms
  - Gibbs sampling

Gibbs sampling

- Gibbs sampling a particular case of Markov Chain Monte Carlo method suitable for multivariate problems

$$\underline{x} = (x_1, x_2, \ldots, x_p) \sim f(\underline{x})$$

$$f(x_i \mid x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p)$$

Gibbs sequence:

$$x_1^{(t+1)} \sim f(x_1 \mid x_2^{(t)}, x_3^{(t)}, \ldots, x_p^{(t)})$$

$$x_2^{(t+1)} \sim f(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)})$$

M

$$x_i^{(t+1)} \sim f(x_i \mid x_1^{(t+1)}, \ldots, x_{i-1}^{(t+1)}, x_{i+1}^{(t)}, \ldots, x_p^{(t)})$$

M

$$x_p^{(t+1)} \sim f(x_p \mid x_1^{(t+1)}, \ldots, x_{p-1}^{(t+1)})$$

1. This is also a Markov Chain whose stationary Distribution is $f(\underline{x})$
2. This is an easier Algorithm, if the conditional densities are easy to work with
3. If the conditionals are harder to sample from, then use MH or Rejection technique within the Gibbs sequence

Metropolis-Hastings Approach

- A Markov Chain can be constructed whose stationary distribution is the desired posterior distribution
- Metropolis et al (1953) showed how and the procedure was later generalized by Hastings (1970). This is called Metropolis-Hastings algorithm.
- Algorithm:
  - Step 1 At iteration t, draw

$$y \sim p(y \mid x^{(t)})$$

$$y: \text{ Candidate Point}$$

$$p: \text{ Candidate Density}$$

- Step 2: Compute the ratio

$$w = Min\left\{1, \frac{f(y) / p(y \mid x^{(t)})}{f(x^{(t)}) / p(x^{(t)} \mid y)}\right\}$$

- Step 3: Generate a uniform random number, u

$$X^{(t+1)} = y \text{ if } u \leq w$$
$$X^{(t+1)} = X^{(t)} \text{ otherwise}$$

- This Markov Chain has stationary distribution $f(x)$.
- Any $p(y|x)$ that has the same support as $f(x)$ will work
- If $p(y|x)=f(x)$ then we have independent samples
- Closer the proposal density $p(y|x)$ to the actual density $f(x)$, faster will be the convergence.

Remarks

- To reduce the impact of starting point usually a few draws are ignored (“Burn-in period”)
- Each successive draws are dependent. Sample every k^{th} observation to get approximate draws.
- Assessing whether or not the chain has converged is difficult.
- Several sequences starting at different points in the parameter space may be useful to assess convergence.

Convergence

- Suppose that there are J parallel sequences each with n iterations. Let q be the scalar parameter of interest.
- Between variance

$$B = \frac{n}{J-1}\sum_{j=1}^{J}(\bar{\theta}_{+j} - \bar{\theta}_{++})^2$$

$$\bar{\theta}_{+j} = \sum_i \theta_{ij}/n; \qquad \bar{\theta}_{++} = \sum_j \bar{\theta}_{+j}/J$$

- Within variance

$$W = \frac{\sum_i \sum_j (\theta_{ij} - \bar{\theta}_{+j})^2}{J(n-1)}$$

Convergence

- At convergence the statistic

$$R = \sqrt{\frac{n-1}{n} + \frac{B}{nW}}$$

should be approximately equal 1. Note that "Between" variance cannot be computed without multiple sequences.