Bayesian Inference and Sampling Techniques

Chase Joyner

Department of Mathematical Sciences
Clemson University

Outline

- Motivate and introduce Bayesian Statistics
- Gibbs Sampling
- Metropolis–Hastings
- Generalized Linear Models
- Bayesian Iterative Re–weighted Least Squares
- Simulations

Motivation

- Suppose you flip a fair coin 100 times and recorded 64 heads and 36 tails.
- The sample percentage of heads is 0.64, but $P(\text{heads}) = 0.5$.
- *A priori* of flipping the coin, we believe it to be fair. We can use this.

Motivation cont.

Nate Silver used Bayesian statistics to

- predict the results of the 2008 presidential election and got 49 out of the 50 states correct.
- predict the results of the 2012 presidential election and got 50 out of the 50 states correct.

Bayesian Inference

Bayesian inference uses Bayes rule to obtain a posterior distribution.

- *A priori* information specified through a prior distribution, denoted $\pi(\boldsymbol{\theta})$.
- Likelihood function, denoted $f(\mathbf{y}|\boldsymbol{\theta})$, specified by the data.

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\mathbf{y})} = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\Theta} f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

- $f(\boldsymbol{\theta}|\mathbf{y})$ is the posterior distribution. It is an update of $\pi(\boldsymbol{\theta})$ after seeing $\mathbf{y}$.

Recognizable Posteriors

- Conjugate priors ensure known posteriors.
- Beta–Binomial, Normal–Normal, and Gamma–Poisson are examples of conjugate priors.
- Typically, known posteriors are not obtainable and so we discuss what to do about this.

Gibbs Sampling

- Posterior distribution is recognizable, but can not sample directly from it due to reliance on other parameters.
- If $\boldsymbol{\theta}$ is our parameter of interest with length r, i.e. we have r parameters of interest, then the Gibbs sampler algorithm is as follows:
  - Given initial values $\boldsymbol{\theta}^{(0)}$, set $t = 1$.
  - Sample $\theta_i^{(t)}$ from $f(\theta_i|\boldsymbol{\theta}_{(-i)}, \mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\theta_i|\boldsymbol{\theta}_{(-i)})$ for $i = 1, ..., r$ and increment t by 1.
  - Repeat s times and obtain dependent sequence of samples $\{\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(s)}\}$.
- This sample acts as draws from true posterior distribution. By weak law of large numbers,

$$\frac{1}{s}\sum_{i=1}^{s}\boldsymbol{\theta}^{(i)} \to \mathrm{E}[\boldsymbol{\theta}|\mathbf{y}].$$

Metropolis–Hastings

- The posterior distribution $f(\boldsymbol{\theta}|\mathbf{y})$ not of any known form.
- So how to obtain the sequence of samples $\{\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(s)}\}$ like in Gibbs sampling?
- Intuitively, include new $\boldsymbol{\theta}^{\star}$ if its posterior density is greater than current $\boldsymbol{\theta}^{(t)}$, else accept it with some probability r.
  - $r = \frac{f(\boldsymbol{\theta}^{\star}|\mathbf{y})}{f(\boldsymbol{\theta}^{(t)}|\mathbf{y})} \frac{J(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{\star})}{J(\boldsymbol{\theta}^{\star}|\boldsymbol{\theta}^{(t)})} = \frac{f(\mathbf{y}|\boldsymbol{\theta}^{\star})\pi(\boldsymbol{\theta}^{\star})}{f(\mathbf{y}|\boldsymbol{\theta}^{(t)})\pi(\boldsymbol{\theta}^{(t)})} \frac{J(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{\star})}{J(\boldsymbol{\theta}^{\star}|\boldsymbol{\theta}^{(t)})}$
- Propose $\boldsymbol{\theta}^{\star}$ from some proposal distribution, denoted $J_{\boldsymbol{\theta}}$.
  - Use this proposal distribution to calculate $\frac{J(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{\star})}{J(\boldsymbol{\theta}^{\star}|\boldsymbol{\theta}^{(t)})}$ in r above. This is the correction factor, in case $\boldsymbol{\theta}^{\star}$ is more likely to be proposed than $\boldsymbol{\theta}^{(t)}$. Otherwise, $\boldsymbol{\theta}^{\star}$ will be over–represented in our sequence.

The Metropolis–Hastings algorithm is as follows:

1. Given initial values $\boldsymbol{\theta}^{(0)}$, set $t = 1$.
2. Propose $\boldsymbol{\theta}^{\star}$ from proposal distribution $J_{\boldsymbol{\theta}}$.
3. Compute acceptance ratio
$$r = \frac{f(\boldsymbol{\theta}^{\star}|\mathbf{y})}{f(\boldsymbol{\theta}^{(t)}|\mathbf{y})} \frac{J(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{\star})}{J(\boldsymbol{\theta}^{\star}|\boldsymbol{\theta}^{(t)})} = \frac{f(\mathbf{y}|\boldsymbol{\theta}^{\star})\pi(\boldsymbol{\theta}^{\star})}{f(\mathbf{y}|\boldsymbol{\theta}^{(t)})\pi(\boldsymbol{\theta}^{(t)})} \frac{J(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{\star})}{J(\boldsymbol{\theta}^{\star}|\boldsymbol{\theta}^{(t)})}.$$
4. Set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{\star}$ with probability $\min\{1, r\}$, $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$ otherwise.
5. Increment t by 1 and return to step 2.

The proposal distribution greatly affects the chain $\{\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(s)}\}$. What to do if a nice proposal distribution is hard to find?

Three major components of a GLM:

- Random component: conditional distribution of Y_i given covariates $\mathbf{X}_i$, which is a member of the exponential family, i.e.
$$f(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}$$
where θ_i depends on the covariates.
- Linear predictor: $\eta_i = \mathbf{X}_i^T\boldsymbol{\beta}$.
- Link function: $g(\mu_i) = \mathbf{X}_i^T\boldsymbol{\beta}$, where g is differentiable and invertible.

Bayesian Iterative Re–weighted Least Squares

- In the situation where covariates are included, $\boldsymbol{\beta}$ becomes an unknown parameter of interest. It can be difficult to find a good proposal distribution for $\boldsymbol{\beta}$.
- Placing a normal prior $N(\mathbf{a}, \mathbf{R})$ on $\boldsymbol{\beta}$, the posterior distribution of $\boldsymbol{\beta}$ takes form

$$f(\boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{a})'\mathbf{R}^{-1}(\boldsymbol{\beta}-\mathbf{a}) + \sum_i \frac{y_i\theta_i - b(\theta_i)}{\phi}\right\}.$$

- Approximating this posterior distribution would be a good choice for the proposal distribution.

Bayesian Iterative Re–weighted Least Squares cont.

- Consider a transformation of the data and weight matrix:

$$\widetilde{y}_i(\boldsymbol{\beta}) = \eta_i + (y_i - \mu_i)g'(\mu_i) \quad \text{and} \quad W_i(\boldsymbol{\beta}) = \frac{1}{b''(\theta_i)g'(\mu_i)^2}.$$

- Carrying out a second order Taylor expansion of the likelihood term

$$\sum_i \frac{y_i\theta_i - b(\theta_i)}{\phi}$$

about $\boldsymbol{\beta}^{(t-1)}$ results in an approximation of $f(\boldsymbol{\beta})$ to be a normal distribution with mean and covariance

$$\mathbf{m}^{(t)} = \mathbf{C}^{(t)} \times \left(\mathbf{R}^{-1}\mathbf{a} + \frac{1}{\phi}\mathbf{X}'\mathbf{W}(\boldsymbol{\beta}^{(t-1)})\widetilde{\mathbf{y}}(\boldsymbol{\beta}^{(t-1)})\right)$$

$$\mathbf{C}^{(t)} = \left(\mathbf{R}^{-1} + \frac{1}{\phi}\mathbf{X}'\mathbf{W}(\boldsymbol{\beta}^{(t-1)})\mathbf{X}\right)^{-1}.$$

- This means $J_{\boldsymbol{\beta}} \stackrel{d}{=} N(\mathbf{m}^{(t)}, \mathbf{C}^{(t)})$.

The analogous Bayesian derivation for this proposal distribution can be thought of as

- Specify the prior for $\boldsymbol{\beta}$ to be $N(\mathbf{a}, \mathbf{R})$.
- The likelihood function for the transformed observations is $\widetilde{\mathbf{y}}(\boldsymbol{\beta}^{(t-1)}) \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{W}^{-1}(\boldsymbol{\beta}^{(t-1)}))$.
- Combine this prior and likelihood to obtain an approximate 'posterior' distribution for $\boldsymbol{\beta}$ to be used as the proposal distribution $J_{\boldsymbol{\beta}}$.

Bayesian Iterative Re–weighted Least Squares cont.

Here we summarize Bayesian IRWLS:

1. Given initial values $\boldsymbol{\beta}^{(0)}$, set $t = 1$.
2. Propose $\boldsymbol{\beta}^{\star}$ from proposal distribution $J_{\boldsymbol{\beta}} \stackrel{d}{=} N(\mathbf{m}^{(t)}, \mathbf{C}^{(t)})$.
3. Compute acceptance ratio r.
4. Set $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{\star}$ with probability $\min\{1, r\}$, $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)}$ otherwise.
5. Increment t by 1 and return to step 2.

It should be noted that in step 3, the correction factor is necessary.

- We let $Y_1, ..., Y_n$ be a random sample from $N(\mu, \sigma^2)$.
- Specify the following two priors

$$\mu|\sigma^2 \sim N(\mu_0, \sigma^2/n_0) \quad \text{and} \quad \sigma^2 \sim IG(\alpha/2, \beta/2).$$

- Posterior distributions become

$$\mu|\sigma^2, \mathbf{Y} \sim N\left(\frac{n\bar{y} + n_0\mu_0}{n + n_0}, \frac{\sigma^2}{n + n_0}\right)$$
$$\sigma^2|\mathbf{Y} \sim IG\left(\frac{n+\alpha}{2}, \frac{\sum_{i=1}^n y_i^2 + n_0\mu_0^2 + \beta}{2} - \frac{(n\bar{y} + n_0\mu_0)^2}{2(n + n_0)}\right).$$

Simulation 1 – Results

$n = 250$, 1000 data sets, 10000 iterations each.

Gibbs Sampling			
Parameter	True values	Estimates	Std. Error
μ	2.3	2.2965	0.05684
σ^2	0.8	0.8117	0.07305

Table: Results of Gibbs sampling

Simulation 2 – Metropolis–Hastings

- We let $Y_1, ..., Y_n$ be a random sample from $N(\mu, \sigma^2)$.
- Specify the following two priors

$$\mu|\sigma^2 \sim N(\mu_0, \sigma^2/n_0) \quad \text{and} \quad \sigma^2 \sim IG(\alpha/2, \beta/2).$$

- Now assume that the posterior distributions are not obtainable (we saw in Gibbs sampling that they are).

Simulation 2 – Metropolis–Hastings cont.

It can easily be shown that the posterior distributions are

$$f(\mu|\sigma^2, \mathbf{Y}) \propto \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\mu)^2 + n_0(\mu-\mu_0)^2\right]\right\}$$

and

$$f(\sigma^2|\mu, \mathbf{Y}) \propto \left(\sigma^2\right)^{-\frac{n+\alpha+1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i-\mu)^2 + n_0(\mu-\mu_0)^2 + \beta\right]\right\}.$$

$n = 250$, 1000 data sets, 10000 iterations each.

Metropolis–Hastings			
Parameter	True values	Estimates	Std. Error
μ	2.3	2.2940	0.05817
σ^2	0.8	0.8148	0.07989

Table: Results of Metropolis–Hastings

Acceptance rate for μ and σ^2 both roughly 22%.

Simulation 3 – Bayesian IRWLS

- Observations are $\mathcal{C}_{ij} \sim \text{Gamma}(\alpha, \mu_{ij}/\alpha)$, ith person in jth group, $i = 1, ..., c_j$, $j = 1, ..., J$.
- Log link $\log \mu_{ij} = \mathbf{X}'_{ij}\boldsymbol{\beta}$.
- Independent prior distributions

$$\boldsymbol{\beta} \sim MVN(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}) \quad \text{and} \quad \alpha \sim \text{Exp}(\lambda).$$

The joint posterior distribution is

$$f(\alpha, \boldsymbol{\beta}|\boldsymbol{\mathcal{C}}) \propto \prod_{j=1}^{J}\prod_{i=1}^{c_j} \exp\left\{\frac{-e^{-\mathbf{X}_{ij}'\boldsymbol{\beta}}\mathcal{C}_{ij} - \mathbf{X}_{ij}'\boldsymbol{\beta}}{1/\alpha} + c(1/\alpha, \boldsymbol{\mathcal{C}})\right\} \cdot \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\right\} \cdot \exp\left\{-\frac{\alpha}{\lambda}\right\},$$

where $c(1/\alpha, \boldsymbol{\mathcal{C}}) = \alpha \log \alpha - \log \Gamma(\alpha) + (\alpha - 1) \log \mathcal{C}_{ij}$.

From the joint posterior distribution, we see that the posterior for α is

$$f(\alpha|\boldsymbol{\beta},\boldsymbol{\mathcal{C}}) \propto \exp\left\{\alpha\gamma + N(\alpha\log\alpha - \log\Gamma(\alpha))\right\},$$

where $N = \sum_{j=1}^{J} c_j$ and

$$\gamma = \sum_{j=1}^{J}\sum_{i=1}^{c_j} -e^{-\mathbf{X}'_{ij}\boldsymbol{\beta}}\mathcal{C}_{ij} - \sum_{j=1}^{J}\sum_{i=1}^{c_j}\mathbf{X}'_{ij}\boldsymbol{\beta} + \sum_{j=1}^{J}\sum_{i=1}^{c_j}\log\mathcal{C}_{ij} - \frac{1}{\lambda}.$$

The posterior distribution for $\boldsymbol{\beta}$ is

$$f(\boldsymbol{\beta}|\alpha,\boldsymbol{\mathcal{C}}) \propto \exp\left\{-\alpha\left(\sum_{j=1}^{J}\sum_{i=1}^{c_j} e^{-\mathbf{X}'_{ij}\boldsymbol{\beta}}\mathcal{C}_{ij} + \sum_{j=1}^{J}\sum_{i=1}^{c_j}\mathbf{X}'_{ij}\boldsymbol{\beta}\right) - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\}.$$

Simulation 3 – Bayesian IRWLS cont.

- We now have the posterior distributions. We must implement Metropolis–Hastings.
- The proposal distribution used for α was $J_\alpha \stackrel{d}{=} \exp\left\{N(\log \alpha^{(t-1)}, \sigma^2)\right\}$.
- We implemented Bayesian IRWLS to propose a new $\boldsymbol{\beta}$. It can be shown that

$$\mathbf{W}(\boldsymbol{\beta}) = I_{N\times N}$$

$$\widetilde{\mathcal{C}_{ij}}(\boldsymbol{\beta}) = \mathbf{X}_{ij}'\boldsymbol{\beta} + (\mathcal{C}_{ij} - \exp(\mathbf{X}_{ij}'\boldsymbol{\beta}))\frac{1}{\exp(\mathbf{X}_{ij}'\boldsymbol{\beta})}.$$

Then, the proposal distribution for $\boldsymbol{\beta}$ is $J_{\boldsymbol{\beta}} \stackrel{d}{=} N(\mathbf{m}^{(t)}, \mathbf{C}^{(t)})$, where

$$\mathbf{m}^{(t)} = \left(\boldsymbol{\Sigma}^{-1} + \alpha \mathbf{X}'\mathbf{X}\right)^{-1} \times \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0 + \alpha \mathbf{X}'\widetilde{\boldsymbol{\mathcal{C}}}(\boldsymbol{\beta}^{(t-1)})\right)$$

and

$$\mathbf{C}^{(t)} = \left(\boldsymbol{\Sigma}^{-1} + \alpha \mathbf{X}'\mathbf{X}\right)^{-1}.$$

Simulation 3 – Results

$n = 100$, 1000 data sets, 10000 iterations each.

Metropolis–Hastings		
Parameter	True values	Estimates
α	5	4.992035
$\boldsymbol{\beta}$	$(-3, 2, 1.1)$	$(-2.999, 2.0002, 1.0995)$

Table: Results of BIRWLS

Acceptance rate for α was 23.4% and the rate for $\boldsymbol{\beta}$ was 97.5%.

Conclusion

- We introduced Bayesian statistics and popular sampling techniques: Gibbs Sampling and Metropolis–Hastings.
- Bayesian Iterative Re–weighted Least Squares is an adaptive version of Metropolis–Hastings that improves the acceptance rates in a good way.
- High acceptance rates are not always good, which can be seen in regular Metropolis–Hastings where the target acceptance rate is between 20 and 50%.
- Great resource: *A First Course in Bayesian Statistical Methods* by Peter Hoff, 2010.
- Great resource: *Sampling from the posterior distribution in generalized linear mixed models* by Dani Gamerman, 1996.

Bibliography

- Hoff, Peter D. (2010). A First Course in Bayesian Statistical Methods. *New York: Springer*
- Gamerman, D. (1996). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing.*
- West, M., Harrison, P. J., and Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association.*

Bibliography cont.

- Grimmer, J. (2010). An Introduction to Bayesian Inference via Variational Approximations. *Political Analysis Advance.*
- Nelder, J. and Wedderburn, R. (1972). Generalized Linear Models. *Royal Statistical Society, Jstor.*

Thank you! Questions?