A clever proposal distribution for Metroplis-Hastings

Chase Joyner

MATH 802

Outline

- Motivate and introduce Bayesian Statistics
- Metropolis–Hastings
- Generalized Linear Models (brief)
- Bayesian Iteratively Weighted Least Squares (BIWLS)
- Discussion of BIWLS
- Small example

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.
- Looking for your phone.
- 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.

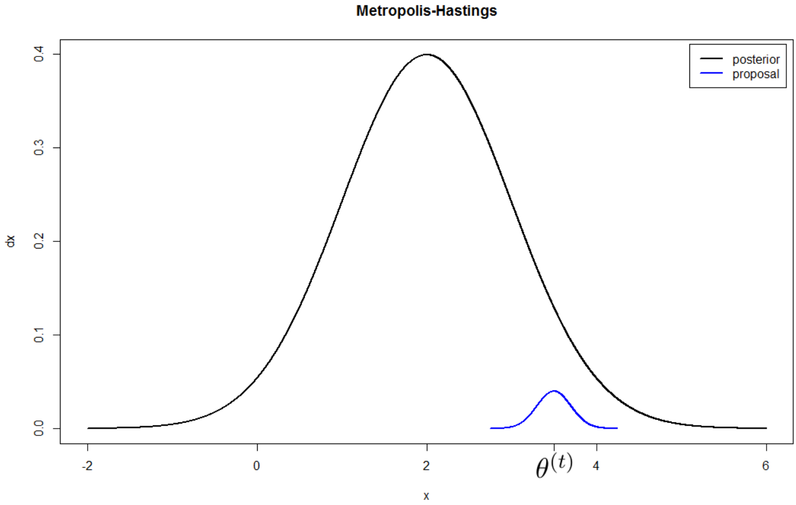
$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\mathbf{y})} = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_\Theta f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$$

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

Metropolis–Hastings

- The posterior distribution $f(\boldsymbol{\theta}|\mathbf{y})$ not of any known form.
- Want to obtain a sequence of samples $\{\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(s)}\}$ to empirically estimate $\boldsymbol{\theta}$.
- Intuitively, include new $\boldsymbol{\theta}^{\star}$ if its posterior density is greater than current $\boldsymbol{\theta}^{(t)}$, else accept it with 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.
  - 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.

Metropolis–Hastings cont.



Metropolis–Hastings cont.

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.
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.

$$f(y_i|\mathbf{x}_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\}$$

  where θ_i depends on the covariates and parameters.

- 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 Iteratively Weighted Least Squares (BIWLS)

- 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

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

- 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 $f(\boldsymbol{\beta} \mid \mathbf{y})$ 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 $J = N(\mathbf{m}^{(t)}, \mathbf{C}^{(t)})$.

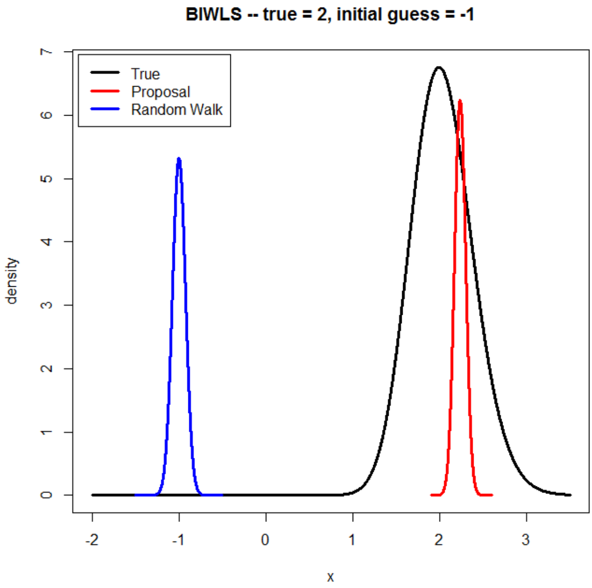
BIWLS cont.

Here we summarize Bayesian IRWLS:

1. Given initial values $\boldsymbol{\beta}^{(0)}$, set $t = 1$.
2. Propose $\boldsymbol{\beta}^{\star}$ from proposal distribution $J = N\left(\mathbf{m}^{(t)}, \mathbf{C}^{(t)}\right)$.
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.

NOTE: Correction factor in r is necessary! Numerator is density of $\boldsymbol{\beta}^{(t)}$ from $N(\mathbf{m}^{\star}, \mathbf{C}^{\star})$ and denominator is density of $\boldsymbol{\beta}^{\star}$ from $N\left(\mathbf{m}^{(t)}, \mathbf{C}^{(t)}\right)$.

BIWLS cont.


BIWLS -- true = 2, initial guess = -1
True
Proposal
Random Walk
density
0
1
2
3
4
5
6
7
-2
-1
0
1
2
3
x

Example

- Assume the independent data $y_i \sim \text{Bern}(p_i)$, where we impose the logistic link

$$g(p_i) = \log \frac{p_i}{1-p_i} = \mathbf{x}_i'\boldsymbol{\beta} \quad \Longrightarrow \quad p_i = \frac{\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}{1+\exp\{\mathbf{x}_i'\boldsymbol{\beta}\}}.$$

- Then the likelihood function is given by

$$\begin{aligned} f(\mathbf{y}) &= \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \exp\left\{\sum_{i=1}^{n}\left[y_i \log\frac{p_i}{1-p_i} + \log(1-p_i)\right]\right\} \\ &= \exp\left\{\sum_{i=1}^{n}\left[y_i\mathbf{x}_i'\boldsymbol{\beta} - \log\left(1+e^{\mathbf{x}_i'\boldsymbol{\beta}}\right)\right]\right\}. \end{aligned}$$

Example cont.

- Therefore the posterior distribution for $\boldsymbol{\beta}$ is given by

$$f(\boldsymbol{\beta} \mid \mathbf{y}) \propto \exp\Bigg\{ -\frac{1}{2}(\boldsymbol{\beta} - \mathbf{a})'\mathbf{R}^{-1}(\boldsymbol{\beta} - \mathbf{a}) + \sum_{i=1}^{n} \left[y_i \mathbf{x}_i' \boldsymbol{\beta} - \log\left(1 + e^{\mathbf{x}_i' \boldsymbol{\beta}}\right)\right] \Bigg\}.$$

- Here, $\theta_i = \mathbf{x}_i' \boldsymbol{\beta}$, $b(\theta_i) = \log\left(1 + e^{\theta_i}\right)$, $\phi = 1$.

Conclusion

- BIWLS improves the acceptance rate in a good way to speed up convergence.
- Could always accept proposed value, but usually not a good idea.
- Initial starting point can sometimes affect the BIWLS algorithm.
- Easily extended to mixed effects models, just affects terms associated with the linear predictor or link function.

Bibliography

- Gamerman, D. (1996). Sampling from the posterior distribution in generalized linear mixed models. *Statistics and Computing.*
- Hoff, Peter D. (2010). A First Course in Bayesian Statistical Methods. *New York: Springer*