A clever proposal distribution for Metroplis-Hastings

Chase Joyner

802 Final Project

Abstract

In this paper we discuss the content of the paper *Sampling from the posterior distribution in generalized linear mixed models* (1996) by Dani Gamerman. The paper outlines a clever way to obtain a proposal distribution to be used in a single Metropolis-Hastings iteration that is a good approximation to the true posterior distribution. As a result of this method, computation time can be greatly reduced and the framework allows for many situations, including models with random effects.

1 Introduction

Bayesian statistics has evolved into a widely used field of statistics. The holy grail of a Bayesian is the posterior distribution for a parameter of interest. However, in many situations this distribution is not recognizable and as a result, Bayesian statistics requires various sampling techniques to remedy this. One such technique is the Metroplis-Hastings algorithm in which a new value of the parameter is proposed from some proposal distribution, and then decided if it should be kept or discarded. While this algorithm can work very well and theoretically the proposal distribution does not matter up to a few assumptions, the convergence rate of the algorithm does rely on the proposal distribution. To this end, one should seek a good proposal distribution. Of course, there are many good proposal distributions that could be used, but we review the one proposed by Dani Gamerman in section 2.4.

2 Methods

2.1 Bayesian Inference

Bayesian techniques combine *a priori* information and data by the use of Baye's rule to obtain a posterior distribution. The *a priori* information specifies a prior distribution, denoted $\pi(\boldsymbol{\theta})$, that uses a person's belief of the true value of the parameters. The data specifies a likelihood function when given the unknown parameters, denoted $f(\mathbf{y}|\boldsymbol{\theta})$. Notice that these are both functions of the parameters of interest and we can apply Baye's rule to formulate the posterior distribution as follows

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

Generally, we can simplify the work by finding the distribution that is proportional to the posterior distribution, and then integrating over the entire parameter space while setting equal to one to find the normalizing constant. The posterior distribution is an update of the prior distribution after observing the data. Seldomly, the posterior distribution can be recognized as a known distribution, and in this case you can report posterior quantities such as the posterior mean, posterior mode, and posterior credible intervals of $\boldsymbol{\theta}$. Unfortunately in most situations the posterior distribution is not of any known form. Therefore, we turn to Markov chain Monte Carlo (MCMC), which is comprised of a set of sampling algorithms that empirically estimate the posterior distribution. One such algorithm is called the Metropolis-Hastings algorithm, which we discuss next.

2.2 Metroplis-Hastings

Metropolis–Hastings is an MCMC technique that can be used when the posterior distribution is not recognizable. Suppose that we have an initial value of our parameters, $\boldsymbol{\theta}^{(0)}$. If we propose a new value $\boldsymbol{\theta}^\star$ from some proposal distribution, say $J_{\boldsymbol{\theta}^\star}$, then an intuitive idea is to include this value in our sample if the posterior density of this proposed value is larger than the posterior density of the current parameter value. However, if this is not the case, then we should accept the proposed value $\boldsymbol{\theta}^\star$ with some probability. An instinctive way to achieve this is to calculate the ratio of these densities, which can be done by equation (1) and the

use of a correction factor. The correction factor is the ratio of the proposal distribution used to propose $\boldsymbol{\theta}^\star$, where the numerator is the proposal distribution centered at the proposed value and evaluated at the current parameter value, conversely for the denominator. This corrects for the jump in the stochastic process to unlikely values of $\boldsymbol{\theta}$. As a result, we obtain the acceptance ratio

$$r = \frac{f(\boldsymbol{\theta}^\star|\mathbf{y})}{f(\boldsymbol{\theta}^{(t)}|\mathbf{y})}\frac{J(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^\star)}{J(\boldsymbol{\theta}^\star|\boldsymbol{\theta}^{(t)})} = \frac{f(\mathbf{y}|\boldsymbol{\theta}^\star)\pi(\boldsymbol{\theta}^\star)}{f(\mathbf{y}|\boldsymbol{\theta}^{(t)})\pi(\boldsymbol{\theta}^{(t)})}\frac{J(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^\star)}{J(\boldsymbol{\theta}^\star|\boldsymbol{\theta}^{(t)})}. \tag{1}$$

After the acceptance ratio r has been computed, the next iteration parameter value is

$$\boldsymbol{\theta}^{(t+1)} = \begin{cases} \boldsymbol{\theta}^\star & \text{with probability } \alpha = \min\{r, 1\} \\ \boldsymbol{\theta}^{(t)} & \text{otherwise.} \end{cases} \tag{2}$$

An important feature of the acceptance ratio above is that if the proposal distribution is a symmetric distribution, then by definition we have that $J(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^\star) = J(\boldsymbol{\theta}^\star|\boldsymbol{\theta}^{(t)})$, and therefore the correction factor is not necessary. A desirable property of the proposal distribution is that the proposed value will get accepted between 20% and 50% of the time in order to have low correlation in the sequence of parameter estimates, but to still allow the chain to move around the parameter space to converge as efficiently as possible. Therefore, one must consider a proposal distribution that has this property, and this can, in certain situations, be difficult. With this said, we discuss in section 2.4 a smarter way to obtain a nice proposal distribution that will yield much higher acceptance rates while keeping the correlation of the sequence of parameters low.

2.3 Generalized Linear Models (GLMs)

A generalized linear model is a generalization of regular linear regression to response types other than normally distributed. GLMs are constructed with three primary components. The first, and most obvious, is the random variable. This will specify the conditional distribution of the response variable, Y_i, given the covariates in the model. It is assumed that this distribution is a member of the exponential family, that is the distribution can be written in the form

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

where the parameters of interest are captured in the canonical parameters θ_i and ϕ is the dispersion parameter, typically associated with the variance of the distribution. The second component is a linear predictor, which is most commonly a linear function of the regressors

$$\eta_i = \mathbf{x}_i'\boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip},$$

where $\mathbf{x}_i$ is a p-dimensional vector of covariates for the ith observation. The last component for a GLM is a smooth and invertible link function, $g(\cdot)$, which relates the mean response to the linear predictor. That is to say that if $\mu_i = \mathrm{E}(Y_i|\theta_i)$, then

$$g(\mu_i) = \eta_i = \mathbf{x}_i'\boldsymbol{\beta}.$$

A link function that one can consider in every situation is the canonical link $\theta_i = \eta_i$. However, there are many link functions that could be used, which depends on the situation or beliefs of how the true mean structure is related to the predictors.

2.4 Bayesian Interatively Weighted Least Squares (BIWLS)

Under the GLM framework, the parameters of interest include the regression coefficents vector $\boldsymbol{\beta}$. We begin by placing a normal prior distribution on $\boldsymbol{\beta}$, say $N(\mathbf{a}, \mathbf{R})$, to obtain a nice proposal distribution for $\boldsymbol{\beta}$. More specifically, the posterior distribution for $\boldsymbol{\beta}$ is considered to be of the form

$$f(\boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{a})'\mathbf{R}^{-1}(\boldsymbol{\beta}-\mathbf{a}) + \sum_i \frac{y_i\theta_i - b(\theta_i)}{\phi}\right\}, \tag{3}$$

where the first term in the exponential is from the prior distribution and the second term is the likelihood term, which depends on $\boldsymbol{\beta}$ through θ_i. The idea proposed by Gamerman is to approximate the posterior distribution with a normal distribution. By carrying out a second order Taylor expansion of the likelihood term

$$\ell(\boldsymbol{\beta}) = \sum_i \frac{y_i\theta_i - b(\theta_i)}{\phi}$$

around some value of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}^{(t-1)}$, and combining terms, we obtain a normal distribution with mean vector and covariance matrix

$$\mathbf{m}^{(t)} = \mathbf{C}^{(t)} \times \left(\mathbf{R}^{-1}\mathbf{a} + \frac{1}{\phi}\mathbf{X}'\mathbf{W}(\boldsymbol{\beta}^{(t-1)})\widetilde{\mathbf{y}}(\boldsymbol{\beta}^{(t-1)}) \right) \tag{4.1}$$

$$\mathbf{C}^{(t)} = \left(\mathbf{R}^{-1} + \frac{1}{\phi}\mathbf{X}'\mathbf{W}(\boldsymbol{\beta}^{(t-1)})\mathbf{X} \right)^{-1} \tag{4.2}$$

respectively, where we define $\widetilde{\mathbf{y}}_i(\boldsymbol{\beta}^{(t-1)})$ as a vector of transformed observations with entries

$$\widetilde{y}_i(\boldsymbol{\beta}^{(t-1)}) = \eta_i + (y_i - \mu_i)g'(\mu_i)$$

and $\mathbf{W}(\boldsymbol{\beta}^{(t-1)})$ as a diagonal matrix with entries

$$W_{ii}(\boldsymbol{\beta}^{(t-1)}) = \frac{1}{b''(\theta_i)g'(\mu_i)^2}.$$

This vector and matrix arise during the Taylor expansion and recombining with the prior terms to obtain the proposal distribution as a normal distribution. This is the distribution used as the proposal distribution, denoted by $J(\boldsymbol{\beta})$, since it is a good approximation to the posterior distribution in (3). The BIWLS method is summarized as follows:

1) start with $\boldsymbol{\beta}^{(0)}$ and set $t = 1$;
2) propose $\boldsymbol{\beta}^{\star}$ by sampling from the proposal distribution $J(\boldsymbol{\beta})$;
3) update $\boldsymbol{\beta}^{(t+1)}$ according to equation (2);
4) increase t by 1 and return to step 2.

An important note that should be made is that even though the proposal distribution $J(\boldsymbol{\beta})$, which is the approximate posterior distribution, is a normal distribution, its parameters $\mathbf{m}^{(t)}$ and $\mathbf{C}^{(t)}$ depend on the previous iterate $\boldsymbol{\beta}^{(t-1)}$, and therefore is not symmetric, i.e. $J(\boldsymbol{\beta}^{(t)} \mid \boldsymbol{\beta}^{(t-1)}) \neq J(\boldsymbol{\beta}^{(t-1)} \mid \boldsymbol{\beta}^{(t)})$. This will require the correction factor in computing the acceptance ratio r in step 3. Furthermore, for clarity, the ratio

$$\frac{J(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{\star})}{J(\boldsymbol{\theta}^{\star}|\boldsymbol{\theta}^{(t)})}$$

in (1) is as follows: The numerator is the density of $\boldsymbol{\beta}^{(t)}$ from $N(\mathbf{m}^{\star}, \mathbf{C}^{\star})$, where $\mathbf{m}^{\star}$ and $\mathbf{C}^{\star}$ are the equations in (4.1) and (4.2) using the proposed value $\boldsymbol{\beta}^{\star}$, and the denominator is

the density of $\boldsymbol{\beta}^{\star}$ from $N\left(\mathbf{m}^{(t)}, \mathbf{C}^{(t)}\right)$. The construction of $\mathbf{m}^{(t)}$ and $\mathbf{C}^{(t)}$ yields the posterior mode and an approximate posterior covariance matrix for $\boldsymbol{\beta}$. Because of this, the acceptance rate in this method is extremely high, usually 90% and higher, but also keeps the correlation low, and this is the benefit of using this method.

The above methodology is easily extended to models with random effects included. The key difference is that the link function is now given by

$$g'(\mu_i) = \eta_i = \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{z}_i'\boldsymbol{\gamma}$$

where $\mathbf{z}_i$ is a vector of additional covariates and $\boldsymbol{\gamma}$ is a vector of random effects with the prior distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$. Consequently, the proposal distribution for $\boldsymbol{\beta}$ is normal with moments

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

Similarly, the proposal distribution for the random effects $\boldsymbol{\gamma}$ is normal with moments

$$\begin{aligned}
\mathbf{m}^{(t)} &= \mathbf{C}^{(t)} \times \left(\frac{1}{\phi}\mathbf{Z}'\mathbf{W}(\boldsymbol{\gamma}^{(t-1)}) \left(\widetilde{\mathbf{y}}(\boldsymbol{\gamma}^{(t-1)}) - \mathbf{X}\boldsymbol{\beta}^{(t-1)} \right) \right) \\
\mathbf{C}^{(t)} &= \left(\boldsymbol{\Sigma}^{-1} + \frac{1}{\phi}\mathbf{Z}'\mathbf{W}(\boldsymbol{\gamma}^{(t-1)})\mathbf{Z} \right)^{-1}.
\end{aligned}$$

3 Discussion

In this paper, we have summarized the main points of Dani Gamerman's paper, *Sampling from the posterior distribution in generalized linear mixed models*. Bayesian statistics is a very useful field of statistics, but it comes with drawbacks as does any statistical method. One drawback is the computation time with the Metropolis-Hastings algorithm, i.e. time until convergence. Gamerman's proposed BIWLS method is a great solution as it allows for high acceptance rates. One can always accept the proposed value and have a perfect acceptance rate, but the values probably are not good. Therefore, high acceptance rates are not always desirable, but here it is. The proposal distribution $J(\boldsymbol{\beta})$ is a good approximation to the true posterior distribution, and hence draws from this are approximate draws from the true distribution, i.e. the distribution that we do want draws from. As a result, smaller

chains are necessary for convergence and so the computation time can be drastically reduced using the BIWLS method. Another point to make is that if one specifies a flat prior for $\boldsymbol{\beta}$, i.e. $\mathbf{R} \to \infty$, then the original iteratively weighted least squares method is recovered. Therefore, this method draws similarities to Frequentist techniques. A drawback of this method is that in some cases, the starting values can affect the algorithm. However, initial starting values are easily obtainable in many ways, such as fitting a simple linear model and using the estimates as the starting value.

References

Gamerman, D. (1996) Sampling from the posterior distribution in generalized linear mixed models.