Bayesian Inference and Sampling Techniques

Chase Joyner

Abstract

In this paper, we introduce Bayesian inference and sampling techniques necessary to estimate unknown parameters of interest. We begin by discussing terminology used in Bayesian methods and then introduce three common sampling techniques. The first is Gibbs Sampling and when it can be used. The second is the Metropolis–Hastings algorithm and how it can be implemented in many situations. The last method, which amends a particular aspect of Metropolis–Hastings, is the Bayesian version of the iterative re–weighted least squares algorithm under a generalized linear model setup. Finally, we simulate each of these methods and discuss the results.

1 Introduction

Bayesian inference is a method in which Baye's rule is primarily used in order to obtain a posterior distribution that can provide all information on unknown parameters of interest. One benefit of using Bayesian methods rather than Frequentist methods is that instead of obtaining just point estimates and confidence intervals for the parameters, a Bayesian obtains an entire distribution for the parameters of interest (Hoff, 2010). For example, suppose that you flip a fair coin 100 times and record 64 heads and 36 tails. Would you consider the coin to be bias? As you can see, a Frequentist approach requires a larger sample size to obtain a long–run frequency for a decent point–estimate. So why has Bayesian methods become more popular in the past few decades? The answer is simply the invention of computers which now enable us to run some highly expensive simulations necessary as a Bayesian (West, 1985). This leads us towards a Bayesian approach, where we include prior knowledge of the coin to assess its fairness.

The posterior distribution, if obtainable, provides all information on the unknown parameters. However, what if the posterior distribution is not of a recognizable form and so sampling from it is impossible? There are a few reasons that may cause this. The first main reason is that the posterior distribution cannot be completely specified due to reliance on other unknown parameters. In this case, Gibbs sampling is a technique that draws samples from full–conditional distributions. A second reason that commonly occurs is that the posterior distribution is not of recognizable form and so sampling cannot be done at all. To remedy this, Metropolis–Hastings or iterative re–weighted least squares are algorithms that can be used to acquire a sample. This method accounts for covariate information and a generalized linear model framework is used (Gamerman, 1996). The framework of this method is built upon the assumption that we have observations distributed according to some exponential family (Nelder, 1972). With these sampling techniques, often referred to as Markov Chain Monte Carlo, or MCMC, we are able to obtain a sample of our parameters as if we drew them directly from the posterior distribution, and hence giving us the desired information (Grimmer, 2010).

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 (Hoff, 2010)

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\mathbf{y})} = \frac{f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_\Theta f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{1}$$

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 1 to find the normalizing constant. The posterior distribution is an update of the prior distribution after observing the data. In most situations, this update may not yield a closed form and renders sampling practically impossible. However, conjugate prior distributions to the likelihood function will ensure a recognizable posterior distribution. A few examples of conjugate priors consist of Beta–Binomial, Normal–Normal, and Gamma–Poisson. For example, suppose that we specify a conjugate prior for a single unknown parameter θ and we have a random sample (independent and identically distributed) from a Poisson distribution. That is to say that

$$\begin{aligned} \text{prior:} \quad \pi(\theta) &= \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \\ \text{likelihood:}\ f(\mathbf{y}|\theta) &= \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!}e^{-\theta}. \end{aligned}$$

Then by equation (1) above, we have that the posterior distribution becomes

$$\begin{aligned} f(\theta|\mathbf{y}) &\propto \prod_{i=1}^{n} \frac{\theta^{y_i}}{y_i!}e^{-\theta} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta} \\ &\propto \theta^{\sum_{i=1}^n y_i}e^{-n\theta}\theta^{\alpha-1}e^{-\beta\theta} \\ &= \theta^{\left(\alpha+\sum_{i=1}^n y_i\right)-1}e^{-(n+\beta)\theta}. \end{aligned}$$

Here we see that the posterior is still Gamma distributed. There are many more conjugate prior examples that could potentially be useful. However, in practice, conjugate priors cannot typically be used.

Before diving into the MCMCs discussed in this paper, we introduce the idea of a full–conditional distribution. Notice that the posterior distribution in (1) is actually a joint distribution of the components of $\boldsymbol{\theta}$. Assume that $\boldsymbol{\theta}$ has length r, then we see that

$$f(\theta_i|\boldsymbol{\theta}_{(-i)},\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \propto f(\mathbf{y}|\boldsymbol{\theta})\pi(\theta_i|\boldsymbol{\theta}_{(-i)}), \tag{2}$$

where $\boldsymbol{\theta}_{(-i)} = (\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_r)$. Notice that here we do not need to include $\pi(\boldsymbol{\theta}_{(-i)})$ since it has nothing to do with the density of θ_i and is therefore just a part of the normalizing constant. The distribution in equation (2) above is referenced as the full–conditional distribution of θ_i, since it is a conditional distribution of θ_i given everything else. This allows the use of Gibbs Sampling, provided that all full–conditionals are analytically obtainable.

2.2 Gibbs Sampling

The idea behind Gibbs sampling is to generate a sequence of samples of the unknown parameters by using the full–conditional posterior distributions of each parameter of interest. To calculate the full–conditional posteriors, we simply use equation (2) above. With these conditional distributions, we are able to approximate the joint posterior distribution $f(\boldsymbol{\theta}|\mathbf{y})$ by generating a dependent sequence of parameters. Given an initial value for each component of $\boldsymbol{\theta}$, denoted $\theta_i^{(0)}$ for $i = 1, ..., r$, sample as follows

$$\theta_i^{(t+1)} \sim f(\theta_i|\boldsymbol{\theta}_{(-i)}^{(t)},\mathbf{y}),$$

where $\boldsymbol{\theta}_{(-i)}^{(t)}$ is a vector consisting of the current parameter estimates, excluding the ith component. After each iteration, we update our parameter vector with the new sample value and proceed. Each iteration consists of updating all r components of $\boldsymbol{\theta}$ and after s iterations, this gives a dependent sequence $\{\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, ..., \boldsymbol{\theta}^{(s)}\}$. With this sequence of samples, we can use the weak law of large numbers to induce properties such as

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

as $s \to \infty$. That is, as we increase the number of iterations without bound, the mean of the sample obtained from Gibbs sampling converges to the true mean (Hoff, 2010). Please see section 4 for a simulation of Gibbs sampling. Next, we introduce the Metropolis–Hastings algorithm, which should be used in the case where the full–conditionals are not analytically obtainable.

2.3 Metropolis–Hastings

Metropolis–Hastings is an MCMC technique that should be used when the posterior distribution is not recognizable and therefore cannot be sampled from. 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 density of this proposed value is larger than the density of the current parameter value. However, if the density is not greater than or equal to the density of the current parameter, 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 evaluated at the current parameter value and the denominator is the proposal distribution evaluated at the proposed value. As a result, we obtain the acceptance ratio (Gamerman, 1996)

$$r = \frac{f(\boldsymbol{\theta}^{\star}|\mathbf{y})}{f(\boldsymbol{\theta}^{(t)}|\mathbf{y})}\frac{J(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{\star})}{J(\boldsymbol{\theta}^{\star}|\boldsymbol{\theta}^{(t)})} = \frac{f(\mathbf{y}|\boldsymbol{\theta}^{\star})\pi(\boldsymbol{\theta}^{\star})}{f(\mathbf{y}|\boldsymbol{\theta}^{(t)})\pi(\boldsymbol{\theta}^{(t)})}\frac{J(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{\star})}{J(\boldsymbol{\theta}^{\star}|\boldsymbol{\theta}^{(t)})}. \tag{3}$$

After the acceptance ratio r has been computed, we set the next iteration parameter value to be

$$\boldsymbol{\theta}^{(t+1)} = \begin{cases} \boldsymbol{\theta}^{\star} & \text{if } r \geq 1 \\ \\ \boldsymbol{\theta}^{\star} & \text{with probability } r \text{ if } r < 1. \end{cases}$$

There are a number of ways to achieve the second line above if $r < 1$. One way would be to generate a uniform random variable between 0 and 1, and if this random variable has value less than r, set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{\star}$. Otherwise, we set $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$. 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 (Hoff, 2010) 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 2.5 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.

2.4 Generalized Linear Models (GLM)

The methods proposed in this paper thus far include only univariate models; however, in most realistic scenarios, it is ideal to include covariate information into a model. To this end, we discuss what a generalized linear model is and the three major components of a GLM. A generalized linear model is a generalization of regular linear regression to response types other than normally distributed. The first, and most obvious component, 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. The second component is a linear predictor, which is most commonly a linear function of the regressors, denoted

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

where $\mathbf{X}_i$ is a vector of covariates for the ith observation. The third requirement for a GLM is a smooth and invertible link function, $g(\cdot)$, which relates the mean of the response variable to the linear predictor. That is to say that if $\mu_i = \mathrm{E}(Y_i)$, then

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

Recall that a distribution is a member of the exponential family if it can be written in the form

$$f(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}. \tag{4}$$

A link function that one can consider in every situation is the canonical link $\theta_i = \eta_i$ (Nelder and Wedderburn, 1972). 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.

2.5 Bayesian Iterative Re–weighted Least Squares

The Bayesian version of the iterative re–weighted least squares was proposed by Mike West in 1985 for the special case of using a canonical link $\theta_i = \eta_i$; however, the method has a straight–forward extension to other links. This algorithm mimics the iterative re–weighted least squares used by Frequentists in order to obtain a nice proposal distribution to be used in a Metropolis–Hastings iteration. 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

$$f(\boldsymbol{\beta}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{a})'\mathbf{R}^{-1}(\boldsymbol{\beta}-\mathbf{a}) + \sum_i \frac{y_i\theta_i - b(\theta_i)}{\phi}\right\}, \tag{5}$$

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 (West, 1985). The idea is to approximate this posterior distribution, which is the true posterior, with a normal distribution to be used as the proposal 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 and covariance matrix

$$\mathbf{m}^{(t)} = \left(\mathbf{R}^{-1} + \frac{1}{\phi}\mathbf{X}'\mathbf{W}(\boldsymbol{\beta}^{(t-1)})\mathbf{X}\right)^{-1} \times \left(\mathbf{R}^{-1}\mathbf{a} + \frac{1}{\phi}\mathbf{X}'\mathbf{W}(\boldsymbol{\beta}^{(t-1)})\widetilde{\mathbf{y}}(\boldsymbol{\beta}^{(t-1)})\right) \tag{5.1}$$

and

$$\mathbf{C}^{(t)} = \left(\mathbf{R}^{-1} + \frac{1}{\phi}\mathbf{X}'\mathbf{W}(\boldsymbol{\beta}^{(t-1)})\mathbf{X}\right)^{-1} \tag{5.2}$$

respectively, that approximates the true posterior distribution in (5). This is the distribution used as the proposal distribution, denoted by $J(\boldsymbol{\beta})$. Notice that if we place a non–informative prior distribution on $\boldsymbol{\beta}$, i.e. $\mathbf{R} \to \infty$ and so we are not narrowing the parameter space of $\boldsymbol{\beta}$, then the original iterative re–weighted least squares algorithm is recovered. Although this method draws similarities from the Fisher scoring algorithm and iterative re–weighted least squares, there is an analogous construction of this approximate posterior distribution under the Bayesian framework. Following the work of McCullagh and Nelder in

1989, we define a vector of transformed observations $\widetilde{\mathbf{y}}(\boldsymbol{\beta})$ and an associated diagonal matrix of weights $\mathbf{W}(\boldsymbol{\beta})$ with respective components

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

This transformation of the data and diagonal weight matrix provide the posterior mode and an approximate posterior covariance matrix for $\boldsymbol{\beta}$ (Gamerman, 1996). Then, like before, combining the normal prior for $\boldsymbol{\beta}$ with a normal likelihood of the transformed observations $\widetilde{\mathbf{y}}(\boldsymbol{\beta}) \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{W}^{-1}(\boldsymbol{\beta}^{(t-1)}))$ results in the approximate posterior distribution with parameters given in (5.1) and (5.2). The iterative 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 – the approximate posterior distribution;
3) accept $\boldsymbol{\beta}^{\star}$ with probability r in equation (3). If accepted, set $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}^{\star}$;
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 rate r in step 3. 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.

3 Simulations

This section will contain examples to illustrate the ideas in section 2. We first look at the Gibbs sampler introduced in section 2.2 and display the results. Then, an example of the Metropolis–Hastings algorithm is simulated in section 3.2 with results. The last simulation in this paper will be the Bayesian version of the iterative re–weighted least squares under the GLM framework.

3.1 Gibbs Sampling

The first simulation is to demonstrate Gibbs Sampling. To do this, let $Y_1, ..., Y_n$ be a random sample from $N(\mu, \sigma^2)$. This will constitute our likelihood used in equation (1). Suppose we have prior belief on the unknown parameters μ and σ^2 that causes us to specify the following prior distributions:

$$\mu|\sigma^2 \sim N\left(\mu_0, \frac{\sigma^2}{n_0}\right) \qquad \text{and} \qquad \sigma^2 \sim IG\left(\frac{\alpha}{2}, \frac{\beta}{2}\right).$$

Under this formulation and using equation (1), the joint posterior distribution becomes

$$\begin{aligned}
f(\mu, \sigma^2|\mathbf{y}) &\propto f(\mathbf{y}|\mu, \sigma^2)\pi(\mu, \sigma^2) = f(\mathbf{y}|\mu, \sigma^2)\pi(\mu|\sigma^2)\pi(\sigma^2) \\
&\propto \prod_{i=1}^{n} \left(\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\} \cdot \left(\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{n_0}{2\sigma^2}(\mu - \mu_0)^2\right\} \cdot \left(\sigma^2\right)^{-\frac{\alpha}{2}-1} \exp\left\{-\frac{\beta}{2\sigma^2}\right\} \\
&= \left(\sigma^2\right)^{-\frac{n+\alpha+1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \mu)^2 + n_0(\mu - \mu_0)^2 + \beta\right]\right\} \\
&= \left(\sigma^2\right)^{-\frac{n+\alpha+1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} y_i^2 + n_0\mu_0^2 + \beta\right]\right\} \cdot \\
&\qquad\qquad \exp\left\{-\frac{n+n_0}{2\sigma^2}\left[\mu^2 - 2\mu\frac{n\overline{y} + n_0\mu_0}{n+n_0} + \left(\frac{n\overline{y} + n_0\mu_0}{n+n_0}\right)^2 - \left(\frac{n\overline{y} + n_0\mu_0}{n+n_0}\right)^2\right]\right\} \\
&= \left(\sigma^2\right)^{-\frac{n+\alpha+1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} y_i^2 + n_0\mu_0^2 + \beta - \frac{(n\overline{y} + n_0\mu_0)^2}{n+n_0}\right]\right\} \cdot \\
&\qquad\qquad \exp\left\{-\frac{n+n_0}{2\sigma^2}\left(\mu - \frac{n\overline{y} + n_0\mu_0}{n+n_0}\right)^2\right\}.
\end{aligned}$$

Then, we easily see that the posterior distribution for μ is

$$f(\mu|\sigma^2, \mathbf{y}) \propto \exp\left\{-\frac{n+n_0}{2\sigma^2}\left(\mu - \frac{n\overline{y}+n_0\mu_0}{n+n_0}\right)^2\right\},$$

and by integrating out μ from the joint posterior distribution above, we have

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

Therefore, the posterior distributions for our unknown parameters are

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

Notice that the posterior distribution's parameters have an intuitive meaning. The mean of the posterior for μ is a weighted average of the prior mean and the sample mean and its variance is σ^2 scaled by the sample size, where n_0 can be thought of as a prior sample size. The parameters for the posterior of σ^2 has similar meanings. We ran a simulation implementing the Gibbs Sampling algorithm discussed in section 2.2 of paper with a sample size of 250 observations, 1000 iterations of 10,000 samples of μ and σ^2 in each iteration. Figure 1 below displays the results:

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

Figure 1: Results of Gibbs sampling

Now, let us see what happens when the posteriors are not obtainable.

3.2 Metropolis–Hastings

The next simulation is to demonstrate the Metropolis–Hastings algorithm discussed in section 2.3. To do this, assume the situation as in the first simulation, i.e. $Y_1, ..., Y_n$ is a random sample from $N(\mu, \sigma^2)$ and μ and σ^2 are assumed to have the following prior distributions:

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

Then, the joint distribution for μ and σ^2 is

$$\begin{aligned} f(\mu, \sigma^2|\mathbf{y}) &\propto f(\mathbf{y}|\mu, \sigma^2)\pi(\mu, \sigma^2) = f(\mathbf{y}|\mu, \sigma^2)\pi(\mu|\sigma^2)\pi(\sigma^2) \\ &\propto \prod_{i=1}^{n} \left(\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\} \cdot \left(\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{n_0}{2\sigma^2}(\mu - \mu_0)^2\right\} \cdot \left(\sigma^2\right)^{-\frac{\alpha}{2}-1} \exp\left\{-\frac{\beta}{2\sigma^2}\right\}. \end{aligned}$$

We have shown in the first simulation that closed forms for the posterior distributions is obtainable, but for the sake of demonstration, assume not. Therefore, we have the posterior distributions of μ and σ^2 are

$$\begin{aligned} f(\mu|\sigma^2, \mathbf{y}) &\propto f(\mathbf{y}|\mu, \sigma^2)\pi(\mu|\sigma^2) \\ &= \prod_{i=1}^{n} \left(\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\} \cdot \left(\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{n_0}{2\sigma^2}(\mu - \mu_0)^2\right\} \\ &= \left(\sigma^2\right)^{-\frac{n+1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \mu)^2 + n_0(\mu - \mu_0)^2\right]\right\} \end{aligned}$$

and

$$\begin{aligned} f(\sigma^2|\mu, \mathbf{y}) &\propto f(\mathbf{y}|\mu, \sigma^2)\pi(\mu|\sigma^2)\pi(\sigma^2) \\ &= \prod_{i=1}^{n} \left(\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\} \cdot \left(\sigma^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{n_0}{2\sigma^2}(\mu - \mu_0)^2\right\} \cdot \left(\sigma^2\right)^{-\frac{\alpha}{2}-1} \exp\left\{-\frac{\beta}{2\sigma^2}\right\} \\ &= \left(\sigma^2\right)^{-\frac{n+\alpha+1}{2}-1} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i - \mu)^2 + n_0(\mu - \mu_0)^2 + \beta\right]\right\}. \end{aligned}$$

We ran a simulation with a sample size of 250 observations, 1000 iterations of 10,000 samples of μ and σ^2. Figure 2 below displays the results:

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

Figure 2: Results of Metropolis–Hastings

The acceptance rate for μ and σ^2 were both roughly 22%. Adjusting the parameters of the proposal distribution, which was chosen to be a normal distribution for each, gave this acceptance rate to allow for quick convergence and keeping correlation in the sequence low. In certain situations, a normal distribution is not a good proposal distribution, and in fact, finding a good proposal distribution can be very difficult at times. Now we simulate the Bayesian version of IRWLS discussed in section 2.5 to demonstrate a fix to this problem.

3.3 Bayesian IRWLS

In this simulation, we will use the methodology discussed in section 2.5. Let us assume that the observations are non-negative measurements of individuals who are placed into groups. Therefore a reasonable distributional assumption for individual i in group j of size c_j is $\mathcal{C}_{ij} \sim$ Gamma$(\alpha, \mu_{ij}/\alpha)$, where μ_{ij}/α is the scale parameter, $i = 1, ..., c_j$ and $j = 1, ..., J$. Notice that this distribution can be written as

$$
\begin{aligned}
f_{\mathcal{C}_{ij}} &= \frac{1}{\Gamma(\alpha)}\left(\frac{\mu_{ij}}{\alpha}\right)^{-\alpha}\mathcal{C}_{ij}^{\alpha-1}\exp\left\{-\frac{\alpha\mathcal{C}_{ij}}{\mu_{ij}}\right\} \\
&= \exp\left\{\frac{-\frac{1}{\mu_{ij}}\mathcal{C}_{ij}-\log\mu_{ij}}{1/\alpha}+\alpha\log\alpha-\log\Gamma(\alpha)+(\alpha-1)\log\mathcal{C}_{ij}\right\}. \qquad (6)
\end{aligned}
$$

Therefore, this distribution is a member of the exponential family. We also wish to include covariate information about each individual, and thus relate the mean of the distribution above to these covariates by a link function. Since the mean μ_{ij} must be positive valued, we use a log link to relate the mean to the covariate information, i.e.

$$\log\mu_{ij} = \mathbf{X}'_{ij}\boldsymbol{\beta}.$$

Then, our likelihood function above can be written in the form

$$f_{\mathcal{C}_{ij}} = \exp\left\{\frac{-e^{-\mathbf{X}'_{ij}\boldsymbol{\beta}}\mathcal{C}_{ij}-\mathbf{X}'_{ij}\boldsymbol{\beta}}{1/\alpha}+\alpha\log\alpha-\log\Gamma\left(\alpha\right)+(\alpha-1)\log\mathcal{C}_{ij}\right\}.$$

We specify the independent prior distributions

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

Notice here that the parameters of interest are α and $\boldsymbol{\beta}$, and so the joint posterior distribution is

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

From here, we see that the posterior distributions will not be of any known form. The posterior distribution 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

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

Also, the posterior distribution for $\boldsymbol{\beta}$ is

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

Now that we have the posterior distributions for the unknown parameters α and $\boldsymbol{\beta}$, we wish to implement the Metropolis–Hastings algorithm. The proposal distribution used for α is simply

$$\alpha^{\star} \sim \exp\left\{N(\log\alpha^{(t-1)}, \sigma^2_{(t-1)})\right\},$$

where we exponentiate a normal random variable centered around the log of the previous iterate in order to keep α positive valued. Also, $\sigma^2_{(t-1)}$ acts a tuning parameter to allow for decent acceptance rates as previously discussed. As for the proposal distribution for $\boldsymbol{\beta}$, we use the methodology discussed in section 2.5. By comparing equations (4) and (6), we see

$$\theta_{ij} = -\frac{1}{\mu_{ij}} \quad \text{and} \quad b(\theta_{ij}) = \log\mu_{ij} = -\log(-\theta_{ij}).$$

Therefore, we have that

$$b''(\theta_{ij}) = \frac{1}{\theta_{ij}^2} = \mu_{ij}^2 = \exp\{2\mathbf{X}_{ij}'\boldsymbol{\beta}\} \quad \text{and} \quad g'(\mu_{ij})^2 = \exp\{-2\mathbf{X}_{ij}'\boldsymbol{\beta}\},$$

and so the weight matrix is $\mathbf{W}(\boldsymbol{\beta}) = I_{N\times N}$. Lastly, we find the transformed observations to be

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

This gives the proposal distribution for $\boldsymbol{\beta}$ to be a normal distribution with parameters

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

and

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

From here, we use these proposal distributions for α and $\boldsymbol{\beta}$ to propose a new value of each in each iteration of the Metropolis–Hastings algorithm. We ran a simulation of 1000 data sets, 10000 iterations per data set, and a sample size of 100 observations. The results of this simulation can be seen below:

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

Figure 3: Results of BIRWLS

The acceptance rate for α was roughly 23.4% and $\boldsymbol{\beta}$ had a 97.5% acceptance rate.

4 Discussion

In this paper, we have introduced Bayesian statistics and, if successful, motivated its usefulness. We have also discussed three sampling techniques commonly used by Bayesians and ran simulations to verify their validity. While Bayesian inference is an extremely important, and quickly evolving, field of statistics, it comes with drawbacks as does any statistical method. The computation time used in these sampling techniques can often be highly expensive. Also, including incorrect prior knowledge into the model can greatly affect inference by drawing samples in the wrong part of the parameter space. However, with good prior knowledge and efficient code, Bayesian inference can be a great tool.

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.*

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.*