Metropolis-Hastings sampling

- Gibbs sampling requires that a sample from each full conditional distribution.
- In all the cases we have looked at so far the conditional distributions were conjugate so sampling was straight-forward.
- What if they are not conjugate?
- We could make draws from the conditional distributions using rejection sampling. This works well if there are only a few non-conjugate parameters but can be difficult to tune.
- Other methods have been proposed:

  1. Auxiliary variables
  2. Slice sampling
  3. Metropolis-Hastings sampling

- Metropolis-Hastings sampling is the most widely used.

Auxiliary Variables

- Sometimes the model can be reexpressed so that the full conditionals are conjugate by adding auxiliary variables.

- For example, consider the zero-inflated Poisson model with $Y_1, ..., Y_n \sim f(y|\theta, \lambda)$ with

$$f(y|\theta, \lambda) = \theta I(y = 0) + (1 - \theta)\text{Poisson}(y|\lambda)$$

  and priors $\theta \sim \text{Beta}(a, b)$ and $\lambda \sim \text{Gamma}(c, d)$.

- The full conditionals for θ and λ are not conjugate.

- Consider the following model

$$Y_i|Z_i \sim \begin{cases} \delta(0) & Z_i = 1 \\ \text{Poisson}(\lambda) & Z_i = 0 \end{cases}$$

  where $\delta(0)$ is the point mass distribution at zero and $Z_i \sim \text{Bernoulli}(\theta)$.

- The Z_i are auxiliary variables.

- After their inclusion, the model is equivalent to the zero-inflated Poisson and conditionally conjugate.

- It is equivalent because:

Auxiliary Variables

The full conditionals are:

- $p(Z_i|\theta, \lambda, \mathbf{y}) =$

- $p(\theta|\lambda, Z_1, ..., Z_n, \mathbf{y}) =$

- $p(\lambda|\theta, Z_1, ..., Z_n, \mathbf{y}) =$

Code

```
#bookkeeping
 n       <- length(y)
 sumy    <- sum(y)
 keepers <- matrix(0,iters,2)

#initial values
 z      <- rep(1,n)
 theta  <- 0.5
 lambda <- 1

for(i in 1:iters){

  #update z

   P1 <- ifelse(y==0,1,0)*theta
   P0 <- dpois(y,lambda)*(1-theta)
   z  <- rbinom(n,1,P1/(P1+P0))

  #update theta and lambda

   n1     <- sum(z==1)
   n0     <- n-n1
   theta  <- rbeta(1,a+n1,b+n0)
   lambda <- rgamma(1,c+sumy,d+n0)

  keepers[i,]<-c(theta,lambda)
}
```

Code is online at

`http://www4.stat.ncsu.edu/~reich/ST740/code/ZIP.R`.

Slice sampling

- It is not obvious how to find auxiliary variables to give conjugacy. Slice sampling is a formulaic way to define auxiliary variables.
- Say the univariate posterior is $p(\theta|\mathbf{y}) \propto h(\theta|\mathbf{y})$.
- We add auxiliary variable $U|\theta, \mathbf{y} \sim \mathrm{U}(0, h(\theta|\mathbf{y}))$.
- The joint distribution is
- The marginal distribution of $\theta|\mathbf{y}$ is
- The full conditional distributions are
- How to sample from $p(\theta|U, \mathbf{y})$?
- It is called slice sampling because:

Metropolis-Hastings sampling

- Metropolis-Hastings sampling is like Gibbs sampling in that you begin with initial values for all parameters, and then update them one at a time conditioned on the current value of all other parameters.

- In this algorithm, we do not need to sample from the full conditionals. All we need is to be able to evaluate the posterior up to a normalizing constant.

- Basic plan for each update:

  1. Draw a candidate for θ_j
  2. Compare the posterior of the candidate and the current value
  3. If the candidate looks good, keep it; if not, keep the current value.

- Formally, the algorithm to update θ_j given the current value $\theta_j^* = \theta_j^{(t-1)}$ is:

  1. Draw a candidate for $\theta_j' \sim g(\theta_j|\boldsymbol{\theta}^*, \mathbf{y})$.
  2. Compute the acceptance ratio
  $$R = \frac{p(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}\frac{g(\theta_j^*|\boldsymbol{\theta}')}{g(\theta_j'|\boldsymbol{\theta}^*)}.$$
  3. Generate $U \sim$ Uniform(0,1).
  4. Set
  $$\theta_j^{(t)} = \begin{cases} \theta_j' & U < R \\ \theta_j^* & U > R \end{cases}$$

  Notation: $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}'$ are current parameter vectors with j^{th} element θ_j^* and θ_j', respectively.

Metropolis-Hastings sampling

- Selecting the candidate distribution is crucial.
- The most common is the Gaussian random walk $\theta_j \sim N(\theta_j^*, c_j^2)$.
- In this case g is symmetric, i.e., $g(\theta_j^*|\boldsymbol{\theta}') = g(\theta_j'|\boldsymbol{\theta}^*)$ and
$$R = \frac{p(\mathbf{y}|\boldsymbol{\theta}')p(\boldsymbol{\theta}')}{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}.$$
- When the candidate distribution is symmetric, the algorithm is the Metropolis algorithm.
- The algorithm has tuning parameter c_j.
- For a random walk candidate, c_j is tuned so that the acceptance probability is 30-50%.
- If c is too large, the acceptance ratio becomes:
- If c is too small, the acceptance ratio becomes:
- Both are bad. Why?

Metropolis-Hastings Sampling

- Consider the logistic regression for binary data $Y_i \in \{0,1\}$ and covariate vector $\mathbf{X}_i = (X_{i1},...,X_{jp})^T$

$$\begin{aligned} \text{Prob}(Y_i = 1|\boldsymbol{\beta}) &= \frac{\exp(\mathbf{X}_i^T\boldsymbol{\beta})}{1+\exp(\mathbf{X}_i^T\boldsymbol{\beta})} & (1) \\ \beta_j &\sim \text{N}(\mu_j, \sigma_j^2). & (2) \end{aligned}$$

- The full posterior is:

Code

```
#full joint log posterior of beta|y for the logistic regression model:
log_post<-function(y,X,beta,prior.mn,prior.sd){
    xbeta  <- X%*%beta
    xbeta  <- ifelse(xbeta>10,10,xbeta)
    like   <- sum(dbinom(y,1,expit(xbeta),log=TRUE))
    prior  <- sum(dnorm(beta,prior.mn,prior.sd,log=TRUE))
like+prior}

logit<-function(x){log(x/(1-x))}
expit<-function(x){exp(x)/(1+exp(x))}


# Start sampling
p<-ncol(X)

#Initial values:
beta      <- rnorm(p,0,1)
keep.beta <- matrix(0,n.samples,p)

acc          <- rep(0,p)
cur_log_post <- log_post(y,X,beta,prior.mn,prior.sd)

for(i in 1:n.samples){

   #Update beta using MH sampling:
   for(j in 1:p){

      # Draw candidate:
       canbeta      <- beta
       canbeta[j]   <- rnorm(1,beta[j],can.sd)
       can_log_post <- log_post(y,X,canbeta,prior.mn,prior.sd)

      # Compute acceptance ratio:
       R <- exp(can_log_post-cur_log_post)
       U <- runif(1)
       if(U<R){
         beta         <- canbeta
         cur_log_post <- can_log_post
         acc[j]       <- acc[j]+1
       }
   }

   keep.beta[i,]<-beta
}
```

Code is online at

http://www4.stat.ncsu.edu/∼reich/ST740/code/Logistic.R.

Metropolis-Hastings sampling

- What is the ideal candidate distribution?
- If the candidate distribution is the full conditional, $g(\theta|\boldsymbol{\theta}_j) = p(\theta_j|\mathbf{y}, \theta_l \text{ for all } l \neq j)$, what is the acceptance ratio?

- Is this a good candidate?
- What does this say about the relationship between Gibbs and Metropolis-Hastings sampling?

Metropolis-Hastings sampling

- Even when the full conditional is not available, with some effort you can find a better candidate distribution than a random walk.

- For example, the Langevin-Hasting algorithm tries to match the candidate to the full conditional using a first-order Taylor approximation,

$$\theta_j' \sim \mathrm{N}\left(\theta_j^* + \frac{c}{2}\frac{\partial f(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}^*)}{\partial \theta_j^*}, c^2\right).$$

- This candidate distribution is not symmetric.

- Hamiltonian MCMC involves an even more elaborate candidate distribution. Gelman et al favor this method and discuss it extensively in their book.

Metropolis-Hastings sampling

Comments:

- How to handle bounded parameters, say $\sigma^2 \in (0, \infty)$?

- Blocked Metropolis is possible. This requires a high-dimensional candidate distribution which is hard to tune.

- Adaptive Metropolis-Hastings tries to match the candidate distribution to the full conditional using past samples. This is performed in the `ARMS` package in `R`.

- You can also keep sampling each parameter until you get a success. This requires a complicated acceptance ratio.

- Combining Metropolis-Hastings and Gibbs sampling:

  - Since Gibbs is a special case of Metropolis-Hastings we can easily combine the two.
  - If some parameters have conjugate full conditionals then they can be updated from their full conditional as in Gibbs; for other parameters you can use an Metropolis step.
  - This is really one large Metropolis sampler, but with the full conditional as the candidate when possible.
  - In the logistic regression example we had prior $\beta_j \sim \mathrm{N}(0, \sigma^2)$ where σ^2 was fixed. What if instead $\sigma^2 \sim$ InvGamma(a, b)?
  - The full conditions for β_j remain unknown, but σ^2 has a conjugate prior.
  - Example code given on the next page.

Code

```
for(i in 1:n.samples){

   #Update beta using MH sampling:

   for(j in 1:p){

      # Draw candidate:
       canbeta      <- beta
       canbeta[j]   <- rnorm(1,beta[j],can.sd)
       can_log_post <- log_post(y,X,canbeta,0,sqrt(sigma2))

      # Compute acceptance ratio:
       R <- exp(can_log_post-cur_log_post)
       U <- runif(1)
       if(U<R){
         beta         <- canbeta
         cur_log_post <- can_log_post
       }
   }

   # Update sigma^2 using Gibbs

    sigma2<-1/rgamma(1,p/2+a,sum(beta^2)/2+b)

}
```