Bayesian Theory and Mechanics for MCMC Methods

A brief review of Bayesian theory

MCMC estimation (Markov Chain Monte Carlo) is based on applying the laws of probability to statistical inference. If we let

$$(\mathcal{Y}, \mathcal{A}_Y, \mathcal{P} = \{P_\theta = f_Y(y; \theta), \theta \in \Theta\})$$

be a parametric statistical model, then, the Bayesian approach would consider $\theta$ to be a random variable with “prior” distribution $f(\theta)$. If we apply Bayes rule to this model we get the “posterior” distribution of $\theta$

$$f(\theta|Y) = \frac{f(Y|\theta)f(\theta)}{f(Y)} = \frac{f(Y|\theta)f(\theta)}{\int f(Y|\theta)f(\theta)d\theta}.$$ (1)

Notice that this all makes sense because we consider $\theta$ to be a random variable so there is no “true” value of the parameter $\theta_0$ out there waiting to be discovered as in the classical approach! Since $\theta$ is random, then the prior expresses my beliefs about $\theta$ before I observe the data $Y$. Once I observe the data, I update my beliefs by using (1). So we have the following definitions.

**Definition 1** $f(\theta)$ is the **prior distribution** of $\theta$, $p(\theta|Y)$ is the **posterior distribution** of $\theta$ and

$$f(Y) = \int f(Y|\theta)f(\theta)d\theta$$

is the **marginal distribution** of the data.

Notice that what we get out of Bayesian estimation is not a point estimate (i.e., is not a number) but rather a distribution. Most of the time what we actually want is a point estimate. Bayesian statistics approach this problem by defining some measure of loss and then minimizing the loss. We do this all the time in economics, except we maximize utility. So, if we view our problem of picking an estimator as a decision problem, all we have to do is pick the estimator (make the decision) that minimizes our loss in some sense. Formally, if we let $D$ be the space of possible decisions we can make (i.e., the space of possible estimators we can pick for example)

**Definition 2** A loss function $L : P \times D \rightarrow \mathbb{R}$ is a non negative mapping that measure the loss incurred from taking decision $d$ when the true distribution is equal to $P$. 

1
Since we are dealing with parametric statistical models, we know that picking $P_\theta$ and picking $\theta$ is equivalent so we redefine the loss function as going from $\Theta \times D \to \mathbb{R}$ and define the remaining elements in terms of this parametric loss function.

**Definition 3** The **posterior expected loss** of a decision $d \in D$ is

$$\rho (d | Y) = \int L (\theta, d) f (\theta | Y) d\theta.$$ 

Finally we define a Bayes estimator.

**Definition 4** Given a sample distribution, a prior and a loss function, a **Bayes estimator** $\hat{\theta} (Y)$ is any function of $Y$ so that

$$\hat{\theta} (Y) = \arg \min_{d \in D} \rho (d | Y).$$

Two common loss functions and their corresponding estimators are

<table>
<thead>
<tr>
<th>Name</th>
<th>$L (\theta, d)$</th>
<th>$\hat{\theta} (Y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quadratic</td>
<td>$(\theta - d)^2$</td>
<td>$E (\theta</td>
</tr>
<tr>
<td>Absolute error</td>
<td>$</td>
<td>\theta - d</td>
</tr>
</tbody>
</table>

**Lemma 1** For $z, a, b$ vectors of size $k$ and $A, B$ symmetric matrices of size $k \times k$ such that $(A + B)^{-1}$ exists. Then

$$(z - a)^\top A (z - a) + (z - b)^\top B (z - b)$$

$$= (z - c)^\top (A + B) (z - c) + (a - b)^\top A (A + B)^{-1} B (a - b)$$

where

$$c = (A + B)^{-1} (Aa + Bb).$$

**Corollary 1** Applying this to the scalar case ($k = 1$) this translates to

$$A (z - a)^2 + B (z - b)^2 = (A + B) \left( z - \frac{Aa + Bb}{A + B} \right)^2 + \frac{AB}{A + B} (a - b)^2.$$ 

**Example 1** Suppose we have an *i.i.d.* sample $Y$ of size $n$ such that $Y_i \sim N (\mu, 1)$. Then

$$f (Y | \mu) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2} \sum_{i=1}^{n} (Y_i - \mu)^2}.$$
Suppose we use the prior
\[ f(\mu) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2} \]
where \( \mu_0 \) and \( \sigma_0^2 \) are the known prior mean and variance. The posterior distribution would be proportional to
\[ f(\mu | Y) = \frac{f(Y | \mu) f(\mu)}{f(Y)}. \]

A couple of important things to notice. Since \( f(\mu | Y) \) is a density it integrates to 1. Notice too that since it is the density of \( \mu \) conditional on the data \( Y \) everything that is not \( \mu \) (i.e., in this case everything that is \( Y \)) is a constant with respect to the random variable \( \mu \). So, we can say that
\[ f(\mu | Y) \propto f(Y | \mu) f(\mu) \]
and all we need to do is find the constant \( K \) that makes this integrate to 1
\[ \int f(\mu | Y) = K \int f(Y | \mu) f(\mu) = 1 \]
\[ \Leftrightarrow K = \frac{1}{\int f(Y | \mu) f(\mu)} = \frac{1}{f(Y)}. \]

You will see why this is useful in a second. All we care about then is the numerator
\[ f(Y | \mu) f(\mu) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2} \left[ \sum_{i=1}^{n} (Y_i - \mu)^2 + \left( \frac{\mu - \mu_0}{\sigma_0} \right)^2 \right]} \]
and it is straightforward to prove that
\[ \sum_{i=1}^{n} (Y_i - \mu)^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 + n (\mu - \bar{Y}) \]
fors \( \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \). Now, remember, only things containing \( \mu \) are random, everything else that can be multiplied and does not contain \( \mu \) is part of the constant that will make this integrate to 1. So
\[ f(\mu | Y) \propto e^{-\frac{1}{2} \left[ \sum_{i=1}^{n} (\mu - \bar{Y}) + \frac{1}{\sigma_0^2}(\mu - \mu_0)^2 \right]} \]
and applying Corollary 1
\[ f(\mu | Y) \propto e^{-\frac{1}{2\sigma_0^2}(\mu - \bar{Y})^2} \]
where

\[ \mu = \frac{n\bar{Y} + \mu_0}{n + \frac{1}{\sigma_0^2}} \]

\[ \sigma^2 = \frac{1}{n + \frac{1}{\sigma_0^2}}. \]

which we recognize a normal distribution so

\[ f(\mu|Y) = N(\mu, \sigma^2). \]

Under quadratic loss then, the Bayes estimator of \( \mu \) would be

\[ E(\mu|Y) = \mu = \frac{n\bar{Y} + \mu_0}{n + \frac{1}{\sigma_0^2}}. \]

Notice that when either \( n \to \infty \) or \( \sigma_0^2 \to \infty \) you get \( E(\mu|Y) = \bar{Y} \) which is the classical (MLE) estimator of the \( \mu \). This is illustrative of a much more general point that we make below.

**Asymptotic Results**

I am not going to give a formal proof. Actually I am not even going to give a hint of a proof. For a hint at a proof (and the regularity conditions required) see Gourieroux and Monfort (1989) for example. Let me then just state that, under regularity conditions, it follows that, if we define \( \theta_0 \) as the true parameter value (in a classical sense) that generates the data from \( f(Y|\theta_0) \); then, for a prior \( \Pi \) with positive probability around \( \theta_0 \) it follows that

\[ \sqrt{n}(\theta^*_n - \hat{\theta}_0) \xrightarrow{a.s.} N(0, I(\theta)^{-1}) \]

where \( \theta^*_n \) is the MLE estimator obtained from the posterior distribution associated with \( \Pi \).

From this result other results follow:

1. The posterior distribution of \( \theta \) as well as that of \( \sqrt{n}(\hat{\theta}_n - \theta_0) \) tend to become symmetric.

2. The posterior mean (i.e., the Bayes estimator under quadratic loss) \( \hat{\theta}_n = E(\theta|Y) \) becomes approximately equal to the MLE estimator \( \theta^*_n \).

3. Since \( \text{plim}(\theta^*_n) = \theta_0 \) then \( \text{plim}(\hat{\theta}_n) = \theta_0 \).
4. The posterior variance \( V(\theta|Y) \) is such that
\[
nV(\theta|Y) = V\left(\sqrt{n}(\theta - \theta^*_n) | Y\right) \approx I(\theta_0)^{-1}.
\]

5. Consistency and asymptotic efficiency is independent of the prior II it has positive density around \( \theta_0 \) so the prior information becomes unimportant.

**Bayesian Mechanics: Normal Linear Regression Revisited**

Suppose we have \( n \) observations such that
\[
f(Y|X, \beta, \tau) \sim N(X\beta, \tau^{-1}I_n)
\]
\[
\propto \tau^{\frac{n}{2}} \exp\left(\frac{-\tau}{2} (Y - X\beta)'(Y - X\beta)\right)
\]
where the precision \( \tau \) is simply the inverse of the variance and \( X \) is \( nxk \) with full rank. Notice that, since we are taking a Bayesian perspective we condition on \( \theta \) because it is a random variable, something we would never do in classical econometrics. We need a prior for \( \theta = (\beta, \tau) \).

1. An non informative prior or improper prior for this case would be
\[
f(\beta, \tau) \propto \tau^{-1}.
\]
This is equivalent to saying that I basically have no prior information regarding my model. If we apply this prior, the posterior would be
\[
f(\beta|X, Y) \propto \tau^{\frac{n}{2}-1} \exp\left(\frac{-\tau}{2} (Y - X\hat{\beta})'(Y - X\hat{\beta})\right)
\]
\[
\propto \tau^{\frac{n}{2}-1} \exp\left(\frac{-\tau}{2} \left\{ (Y - X\hat{\beta})' (Y - X\hat{\beta}) + (\beta - \hat{\beta})' X'X (\beta - \hat{\beta}) \right\}\right)
\]
for \( \hat{\beta} = (X'X)^{-1}X'Y \). Let \( s(Y) = (Y - X\hat{\beta})' (Y - X\hat{\beta}) \). Then if we are interested in \( \beta \) we could integrate \( \tau \) out to obtain
\[
f(\beta|X, Y) = \int f(\beta, \tau|X, Y) \, d\tau
\]
\[
\propto \int \tau^{\frac{n}{2}-1} \exp\left(\frac{-\tau}{2} \left\{ s(Y) + (\beta - \hat{\beta})' X'X (\beta - \hat{\beta}) \right\}\right)
\]
\[
\propto \exp\left(1 + \frac{1}{s(Y)} (\beta - \hat{\beta})' X'X (\beta - \hat{\beta})\right)
\]
which is the kernel of a $t$ distribution

$$f (\beta|Y) = t_{n-k} \left( \hat{\beta}, \hat{\Sigma} \right)$$

with $n - k$ degrees of freedom, mean $\hat{\beta}$ and scale $\hat{\Sigma} = \frac{1}{n-k} s(Y) (X'X)^{-1}$. Doing the same for $\tau$

$$f (\tau|X,Y) \propto \tau^{\frac{(n-k)}{2} - 1} \exp \left( -\frac{\tau}{2} s(Y) \right)$$

which is the kernel of a Gamma distribution. The mean of this distribution is

$$E (\tau|X,Y) = \frac{n - k}{s(Y)} = \frac{1}{\sigma^2}.$$ 

Notice that these are basically the repeated sample distribution estimators of classical statistics.

2. A conjugate prior is one for which the posterior distribution is a member of the same family as the prior. That is, if $P(\theta)$ is say, normal, then the posterior of $\theta$ will also be normal. A common set of proper priors for the linear regression case are independent conjugate priors for $\beta$ and $\tau$.

$$f (\beta) = N (\beta_0, \Lambda_0^{-1})$$
$$f (\tau) = G (\alpha_1, \alpha_2)$$

and the posterior

$$f (\beta, \tau|X,Y) \propto \tau^{\frac{\alpha_1}{2} + \frac{n}{2} - 1} \exp \left\{ -\frac{\tau}{2} \left( s(Y) + \left( \beta - \hat{\beta} \right)' X' X \left( \beta - \hat{\beta} \right)' \right) \right\} \times \exp \left\{ -\frac{1}{2} (\beta - \beta_0)' \Lambda_0 (\beta - \beta_0) \right\} \exp (-\alpha_2 \tau)$$

This posterior does not lead to convenient expressions for the marginals. It will prove useful to instead derive the conditional posterior distributions.

If we start with $\beta$ conditional on $\tau, X, Y$ we can drop every multiplicative term not including $\beta$ so

$$f (\beta|X,Y,\tau) \propto \exp \left\{ -\frac{1}{2} \left[ \left( \beta - \hat{\beta} \right)' \tau X' X \left( \beta - \hat{\beta} \right)' + (\beta - \beta_0)' \Lambda_0 (\beta - \beta_0) \right] \right\}$$

which, by lemma 1

$$f (\beta|X,Y,\tau) \propto \exp \left\{ -\frac{1}{2} \left( \beta - \bar{\beta} \right)' \Sigma^{-1} (\beta - \bar{\beta}) \right\}$$
where
\[
\Sigma = (\tau'X + \Lambda_0)^{-1} \\
\beta = \Sigma(\tau'Y + \Lambda_0\beta_0).
\]
So
\[
f(\beta|Y, X, \tau) \sim N(\beta, \Sigma)
\]
a multivariate normal distribution. In the same manner, for \(\tau\) conditional on \(\beta\)
\[
f(\tau|X, Y, \beta) \propto \tau^{\frac{n}{2} + \alpha_1 - 1} \exp \left\{ -\tau \left[ \frac{1}{2} (Y - X\beta)'(Y - X\beta) + \alpha_2 \right]\right\}
\]
which is the kernel of a gamma distribution so
\[
f(\tau|X, Y, \beta) \sim G\left(\frac{n}{2} + \alpha_1, \frac{\sum_{i=1}^{n} (Y_i - X_i\beta)^2}{2} + \alpha_2\right).
\]

**Practical guide to Gibbs sampling and completion**

Since I need you guys to finish problem set 1 and move on to problem set 2 I am just going to cover the ideas with no justification. Next class I will give you the justification and definitions for all the ideas presented here.

The idea behind the Gibbs sampler is that we want to sample from a general multivariate posterior distribution \(f(\theta|Y)\). Instead of giving drawing an iid sample from the whole distribution, the Gibbs sampler gives a dependent sample out of which we sample repeatedly to form a Markov chain that converges to the distribution \(f(\theta|Y)\). To do this, we break the distribution into \(h < k\) blocks (where \(k\) is the size of \(\theta\)) and sample sequentially from each block conditional on the remaining blocks. Next class I will give you a formal treatment of all of this but for now let me leave it there and go back to my normal linear regression example.

We just derived the posterior for the linear regression example with proper priors and so that it was not a very tractable distribution. We then broke it into 2 blocks \(\beta\) and \(\tau\). We then derived the conditional posteriors for each block (i.e., we derived \(f(\beta|\tau, X, Y)\) and \(f(\tau|\beta, X, Y)\)). Notice of course that this posteriors by themselves are not of very much use. I could get define the Bayes estimator of \(\beta\) but only in terms of \(\tau\) but I do not know \(\tau\)! Enter Gibbs sampling. It turns out, that if I sample sequentially from the conditional posteriors I will eventually get a sample from the
actual joint distributions of $\beta$ and $\tau$. For my linear regression example the Gibbs sampler would be:

1. Choose an initial value $\tau^0 > 0$

2. For $m = 1, 2, ...$

   (a) Sample
   $$\beta^{(m)}|\tau^{(m-1)} \sim N \left( \beta^{(m-1)}, \Sigma(\tau^{(m-1)}) \right),$$

   (b) Sample
   $$\tau^{(m)}|\beta^{(m)} \sim G \left( \frac{n}{2} + \alpha_1, \frac{\sum_{i=1}^n \left( Y_i - X_i' \beta^{(m)} \right)^2}{2} + \alpha_2 \right).$$

That is it. By following this recursive sampling procedure, I will get (after letting the chain converge) a sample from the joint distribution of $(\beta, \tau)$. From this sample I can now compute my estimators by taking the mean of the sample. I can get its standard errors by taking the standard error of the sample and so on!

Let me know look at another idea that I will justify next class called completion by looking at an example that will help you solve the probit case (which I will solve next class). Remember that 2 classes ago we derived the tobit model. The pdf., of $Y_i$ conditional on $X_i$

$$f(Y_i|X_i) = \Phi \left( \tau^{\frac{1}{2}} (-X_i' \beta) \right)^{1(Y_i=0)} \left[ \phi \left( \tau^{\frac{1}{2}} (Y_i - X_i' \beta) \right) \tau^{\frac{1}{2}} \right]^{1(Y_i>0)},$$

and by i.i.d.,

$$f(Y|X) = \prod_{i=1}^n \Phi \left( \tau^{\frac{1}{2}} (-X_i' \beta) \right)^{1(Y_i=0)} \left[ \phi \left( \tau^{\frac{1}{2}} (Y_i - X_i' \beta) \right) \tau^{\frac{1}{2}} \right]^{1(Y_i>0)}.$$

Now, suppose we put a non informative prior on $\beta$ and a proper gamma prior on $\tau$. Then, the posterior distribution of $\tau, \beta$ will be

$$f(\beta, \tau|X, Y) \propto \tau^{\alpha_1-1} \exp (-\alpha_2 \tau) \prod_{i=1}^n \Phi \left( \tau^{\frac{1}{2}} (-X_i' \beta) \right)^{1(Y_i=0)} \left[ \phi \left( \tau^{\frac{1}{2}} (Y_i - X_i' \beta) \right) \tau^{\frac{1}{2}} \right]^{1(Y_i>0)}$$

Needless to say, this does not look like any distribution we can easily sample from. Now, the way we stated the model that time is actually part of the solution to the problem. Remember, we claimed
the existence of a latent random variable $Y^*$ such that

$$
Y^*_i = X'_i \beta + \varepsilon_i \sim N \left( X'_i \beta, \tau^{-1} \right)
$$

$$
Y = Y^* \text{ when } Y^* > 0
$$

$$
Y = 0 \text{ when } Y^* \leq 0.
$$

If we could only observe $Y^*$ we would have a simple linear regression model. What about not observing it but rather conditioning on it? Remember, I am a Bayesian so the parameters are random variables for me, but so is $Y^*$. Why not treat it as simply another parameter of the model? If I do that, the complete (or data augmented) posterior distribution of the model (i.e., the distribution of $\beta, \tau, Y^*$) would be

$$
f (\beta, \tau, Y^* | X, Y) \propto \tau^{\alpha_1-1} \exp \left( -\alpha_2 \tau \right) \tau^\frac{n}{2} \exp \left( -\frac{\tau}{2} \sum_{i=1}^{n} (Y^*_i - X'_i \beta)^2 \right)
$$

where, remember $Y^* = Y$ if it is larger that zero. It is still not a very tractable distribution but now a block structure suggests itself naturally. Lets first look at $\beta$

$$
f (\beta | \tau, Y^*, X, Y) \propto \exp \left( -\frac{\tau}{2} \sum_{i=1}^{n} (Y^*_i - X'_i \beta)^2 \right)
$$

$$
\propto \exp \left( -\frac{\tau}{2} \left( \beta - \hat{\beta} \right)' X' X \left( \beta - \hat{\beta} \right) \right)
$$

which is the kernel of a normal so

$$
f (\beta | \tau, Y^*, X, Y) \sim N \left( \hat{\beta}, \hat{\Sigma} \right)
$$

where

$$
\hat{\beta} = (X'X)^{-1} X'Y^*
$$

$$
\hat{\Sigma} = \tau^{-1} (X'X)^{-1}.
$$

Next block, lets look at $\tau$

$$
f (\tau | \beta, Y^*, X, Y) \propto \tau^{\alpha_1-1} \exp \left( -\alpha_2 \tau \right) \tau^\frac{n}{2} \exp \left( -\frac{\tau}{2} \sum_{i=1}^{n} (Y^*_i - X'_i \beta)^2 \right)
$$

$$
\propto \tau^{\frac{n}{2}+\alpha_1-1} \exp \left( -\tau \left( \sum_{i=1}^{n} \frac{(Y^*_i - X'_i \beta)^2}{2} + \alpha_2 \right) \right)
$$
which is the kernel of a gamma distribution so

\[ f(\tau|\beta,Y^*,X,Y) \sim G \left( \frac{n}{2} + \alpha_1, \sum_{i=1}^{n} \frac{(Y_i^* - X_i'\beta)^2}{2} + \alpha_2 \right). \]

The last block would be \( Y^* \). Notice that we know with certainty what \( Y^* \) is when \( Y > 0 \). We only need to worry about the case when \( Y^* \leq 0 \). So

\[ f(Y_i^*|\beta,\tau,X_i,Y_i) \sim N \left( X_i'\beta, \tau^{-1} \right) \mathbb{1}(Y_i^* \leq 0) \]

that is, \( Y_i^* \) is a truncated normal from \((-\infty, 0)\) for the case in which \( Y_i = 0 \). So, the Gibbs sampling algorithm for the tobit model would be

1. Choose an initial value \( \tau^0 > 0, \beta^0 \)

2. For \( m = 1, 2, ... \)

   (a) For \( i = 1, ..., N \)

      i. If \( Y_i > 0 \) set \( Y_i^* = Y_i \)

      ii. Else if \( Y_i = 0 \) sample

      \[ Y_i^* \sim T N_{(-\infty, 0)} \left( X_i'\beta^{(m)}, (\tau^{(m)})^{-1} \right) \]

   (b) Sample

      \[ \beta^{(m)} \sim N \left( \hat{\beta} \left( \tau^{(m)} \right), \hat{\Sigma} \left( \tau^{(m)} \right) \right) \]

   (c) Sample

      \[ \tau^{(m)} \sim G \left( \frac{n}{2} + \alpha_1, \sum_{i=1}^{n} \frac{(Y_i^* - X_i'\beta^{(m)})^2}{2} + \alpha_2 \right) \]

I have not yet talked about sampling from a truncated normal variable. There is a very efficient method but I will present it next class. For the moment you can do one of three things.

1. You can use the fortran routine that I provide in the class website in the PROBABILITY module. This routine actually implements the algorithm I am going to present next class.
2. Use a simple but slow, accept reject procedure. Suppose you sample from
\[ TN_{(-\infty, 0]} (\mu, \sigma^2) \]
then simply sample from
\[ x \sim N (\mu, \sigma^2) . \]
If \( x < 0 \) then accept the number. If it is not, draw another \( x \) until it is.

3. Use an inverse cdf method. Notice that if
\[ X \sim TN_{(a,b)} (\mu, \sigma^2) \]
then
\[ F_X (x) = \frac{\Phi_{\mu,\sigma^2} (x) - \Phi_{\mu,\sigma^2} (a)}{\Phi_{\mu,\sigma^2} (b) - \Phi_{\mu,\sigma^2} (a)} \]
so, for \( u \) uniform \((0,1)\)
\[ u = F (x) \Rightarrow x = F^{-1} (u) = \Phi^{-1} (\Phi_{\mu,\sigma^2} (a) + u [\Phi_{\mu,\sigma^2} (b) - \Phi_{\mu,\sigma^2} (a)]) \]
would generate a draw \( x \) from the truncated normal distribution.