STAT 515
Example of MCMC with full conditional calculations

This example comes from *Markov Chain Monte Carlo in Practice* by Gilks, Richardson, and Spiegelhalter (1996, first edition), pages 75–76.

Suppose

$$\begin{array}{rcl} Y_1, \ldots, Y_n & \overset{\text{iid}}{\sim} & N(\mu, \tau^{-1}) \\ \mu & \sim & N(0,1) \\ \tau & \sim & \text{Gamma}(2,1), \end{array}$$

and μ and τ are independent (that is, the prior density for (μ, τ) is the product of the individual densities). Let us find the full conditional distributions for μ and τ. First, a bit of preliminary setup:

The likelihood function is the joint density of the data (given the parameters), viewed as a function of the parameters:

$$L(\mu,\tau) = \prod_{i=1}^n f(Y_i \mid \mu, \tau) = \prod_{i=1}^n \frac{\sqrt{\tau}}{\sqrt{2\pi}} \exp\{-\tau(Y_i - \mu)^2/2\} = (2\pi)^{-n/2} \tau^{n/2} \exp\left\{ -\tau \sum_{i=1}^n (Y_i - \mu)^2/2 \right\}.$$

To find the joint density of $(\mathbf{Y}, \mu, \tau)$, we multiply $f(\mathbf{Y} \mid \mu, \tau)$ (the likelihood) by the prior:

$$p(\mathbf{Y}, \mu, \tau) = (\text{constant}) \times \tau^{n/2} \exp\left\{ -\tau \sum_{i=1}^n (Y_i - \mu)^2/2 \right\} \exp\{-\mu^2/2\} \tau \exp^{-\tau}.$$

If the data $\mathbf{Y}$ are observed, then in a Bayesian framework our attention focuses on the conditional density of (μ, τ) given $\mathbf{Y}$. This may be written

$$p(\mu, \tau \mid \mathbf{Y}) = \frac{p(\mathbf{Y}, \mu, \tau)}{p(\mathbf{Y})} = \frac{p(\mathbf{Y}, \mu, \tau)}{\int\int p(\mathbf{Y}, \mu^*, \tau^*)\, d\mu^* d\tau^*},$$

but the denominator above isn't really important for our purposes (because it does not involve μ or τ), so we may simply write

$$p(\mu, \tau \mid \mathbf{Y}) \propto p(\mathbf{Y}, \mu, \tau).$$

Now we can finally talk about the full conditionals. For μ, we obtain

$$p(\mu \mid \tau, \mathbf{Y}) = \frac{p(\mu, \tau \mid \mathbf{Y})}{p(\tau \mid \mathbf{Y})},$$

but once again the denominator does not involve μ so we can combine the expressions above to write, as a function of μ,

$$p(\mu \mid \tau, \mathbf{Y}) \propto p(\mathbf{Y}, \mu, \tau) \propto \exp\left\{ -\tau \sum_{i=1}^n (Y_i - \mu)^2/2 \right\} \exp\{-\mu^2/2\}.$$

(We have ignored all of the factors in $p(\mathbf{Y}, \mu, \tau)$ that do not involve μ.) Similarly, as a function of τ, the full conditional for τ satisfies

$$p(\tau \mid \mu, \mathbf{Y}) \propto p(\mathbf{Y}, \mu, \tau) = \tau^{n/2} \exp\left\{ -\tau \sum_{i=1}^n (Y_i - \mu)^2/2 \right\} \tau \exp^{-\tau}.$$

If we do a little bit of algebra, we get

$$p(\mu \mid \tau, \mathbf{Y}) = \exp\left\{ -\frac{1+n\tau}{2} \left(\mu - \frac{\tau \sum_{i=1}^n Y_i}{1+n\tau} \right)^2 \right\} \times (\text{stuff not involving } \mu)$$

as the full conditional density for μ. We recognize this! It is a normal density function for μ, where the mean is $\tau \sum_{i=1}^n Y_i/(1+n\tau)$ and the variance is $1/(1+n\tau)$. Or, in symbols,

$$\mu \mid \tau, \mathbf{Y} \sim N\left(\frac{\tau \sum_{i=1}^n Y_i}{1+n\tau}, \frac{1}{1+n\tau} \right).$$

Similarly, for τ, we can do some algebra to obtain

$$p(\tau \mid \mu, \mathbf{Y}) = \tau^{1+(n/2)} \exp\left\{-\tau\left(1+\frac{1}{2}\sum_{i=1}^{n}(Y_i-\mu)^2\right)\right\}\exp^{-1/\tau} \times (\text{stuff not involving } \tau)$$

as the full conditional, and we recognize this as a gamma density for τ. In symbols,

$$\tau \mid \mu, \mathbf{Y} \sim \text{Gamma}\left(2+\frac{n}{2}, 1+\frac{1}{2}\sum_{i=1}^{n}(Y_i-\mu)^2\right).$$

Let's use these facts to run a Markov chain. The goal will be to set up the chain using variable-at-a-time Metropolis-Hastings in such a way that its stationary distribution is the posterior density $p(\mu, \tau \mid \mathbf{Y})$. The data $\mathbf{Y}$ are at `http://sites.stat.psu.edu/~dhunter/515/hw/MCMCexampleData.txt`

```
> y <- scan("http://sites.stat.psu.edu/~dhunter/515/hw/MCMCexampleData.txt")
```

First, we must initialize the chain using some $\mu^{(1)}$ and $\tau^{(1)}$. Ideally, we could sample from the true stationary (posterior) distribution to do this, but if we could do that, we wouldn't need MCMC! So let's simply use $(\mu^{(1)}, \tau^{(1)}) = (1, 2)$, which are the prior means. (We could also look at the data and take the starting values to be the sample mean and the reciprocal of the sample variance.)

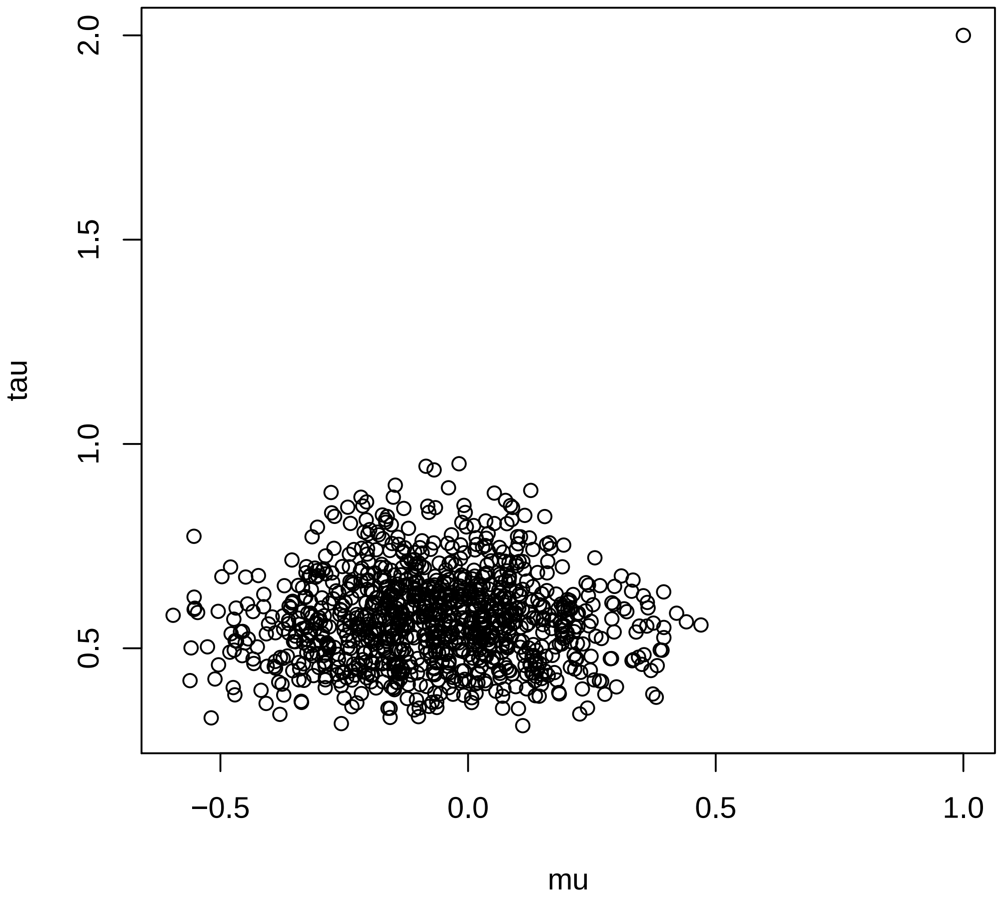
```
> iteration <- 1
> mu <- 1
> tau <- 2
> sumOfY <- sum(y) # We'll need this value repeatedly.
> n <- length(y)
```

We will now run 1000 steps of a Markov chain in which we sample $\mu^{(i+1)}$ from the full conditional based on $\tau^{(i)}$ and $\mathbf{Y}$, then sample $\tau^{(i+1)}$ from the full conditional based on $\mu^{(i+1)}$ and $\mathbf{Y}$:

```
> while (iteration <= 1e3) {
+   tmp <- tau[iteration]
+   mu <- c(mu, rnorm(1, mean=sumOfY * tmp / (1+n*tmp), sd = sqrt(1 / (1 + n*tmp))))
+   tau <- c(tau, rgamma(1, shape=2+n/2, rate=1+sum((y-mu[iteration+1])^2)/2))
+   iteration <- iteration+1
+ }
```

Let us now check on some of the properties of our MCMC sample. To get a sense of the shape of the sample, a simple scatterplot will help:

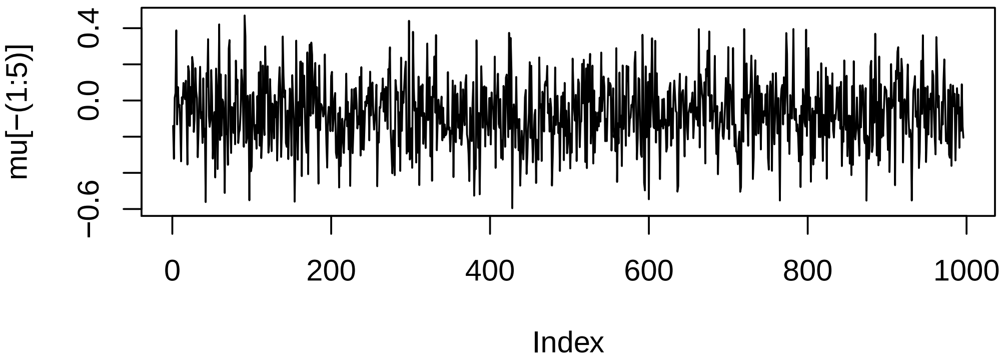
```
> plot(cbind(mu, tau))
```

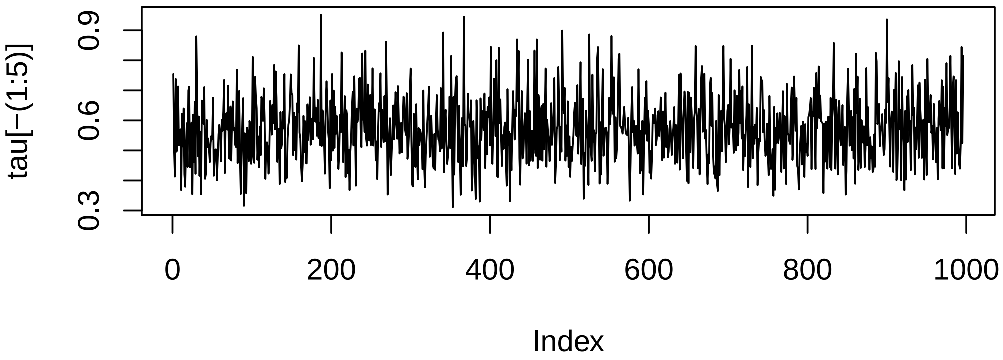


The first thing we see is that the initial point is a long way from the rest of the points. It is partly for this reason that sometimes a burnin period is used. In our case, omitting the first several points appears to be sufficient.

Let us now look at some diagnostics to assess whether the Markov chain seems to be mixing well. Simple plots of iteration against μ or iteration against τ (omitting the first 5 points) give a pretty good sense of this:

```
> par(mfrow=c(2,1))
> plot(mu[-(1:5)], type="l")
> plot(tau[-(1:5)], type="l")
```

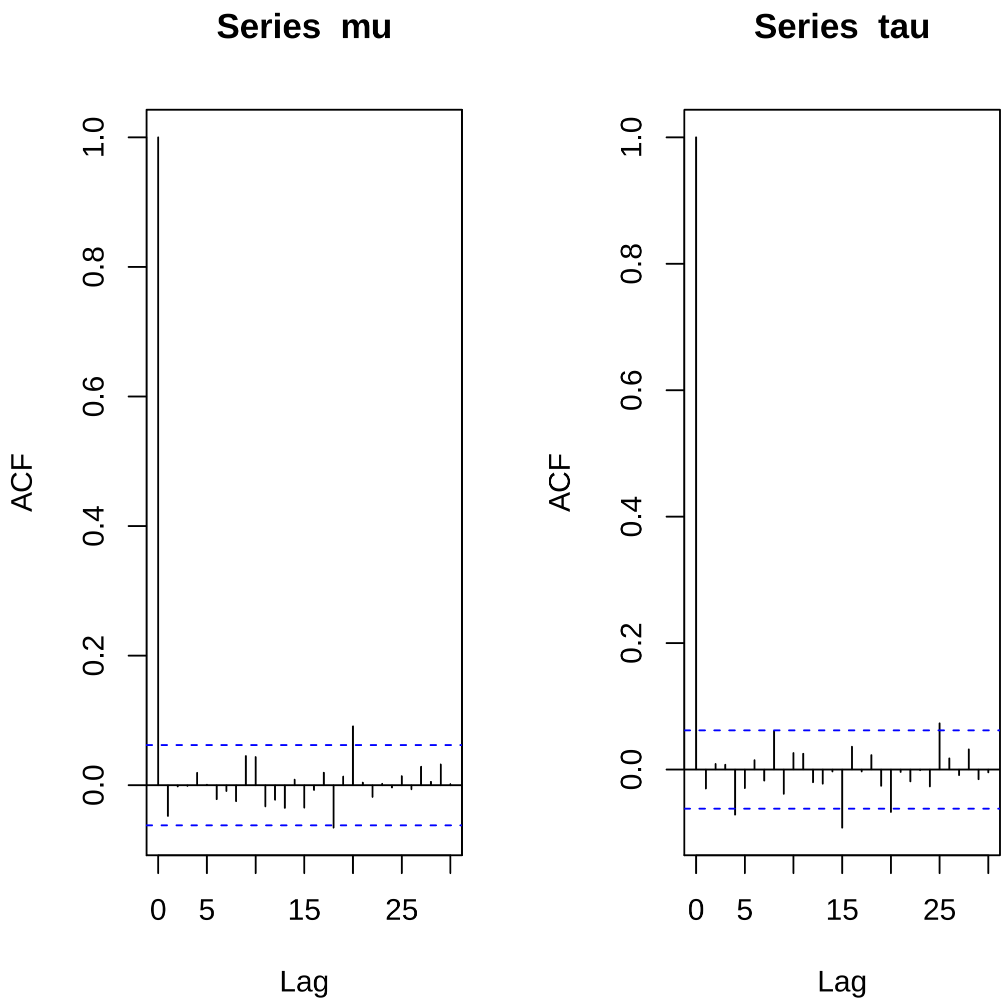




These so-called traceplots do not give the sense of any overall trend, which might indicate that the chain has not yet had a chance to adequately explore the posterior parameter space. (NB: The plot gives no guarantees! Some types of bizarre behavior might not show up on a plot.)

Another helpful plot is a plot of the values of auto-correlation for various values of the "lag" variable. For lag k, the ACF (auto-covariance function) is the sample correlation between the two samples $\theta_1, \ldots, \theta_{n-k}$ and $\theta_{1+k}, \ldots, \theta_n$. In this case, the auto-correlation is close to zero even for a lag of one, which means that successive draws from this chain are nearly uncorrelated. This is a very good thing!

```
> par(mfrow=c(1,2))
> acf(mu)
> acf(tau)
```



Let us now focus on the μ parameter specifically. The Bayes estimator of μ is equal to the mean of the posterior distribution; let us denote this value by $\tilde{\mu}$. We do not actually know what $\tilde{\mu}$ is, but we can estimate $\tilde{\mu}$ using our MCMC sample from the posterior. Think of the following analogy: If you have a random sample of size 1000 from a population with mean $\tilde{\mu}$, then you would use the sample mean as the estimator of $\tilde{\mu}$, and by the central limit theorem you could also construct a 95% confidence interval for $\tilde{\mu}$ using the sample mean plus or minus $1.96\hat{s}$, where $\hat{s}$ is an estimator of the standard deviation of the sample mean (which is a random variable!). In a similar way, we can use the sample mean of our μ sample—let's denote it by $\hat{\mu}$—to estimate $\tilde{\mu}$. However, it is not always straightforward to estimate the standard deviation of $\hat{\mu}$ because the MCMC sample is not an i.i.d. sample. In fact, if we assume (simplistically) that the chain is stationary with distribution π, then

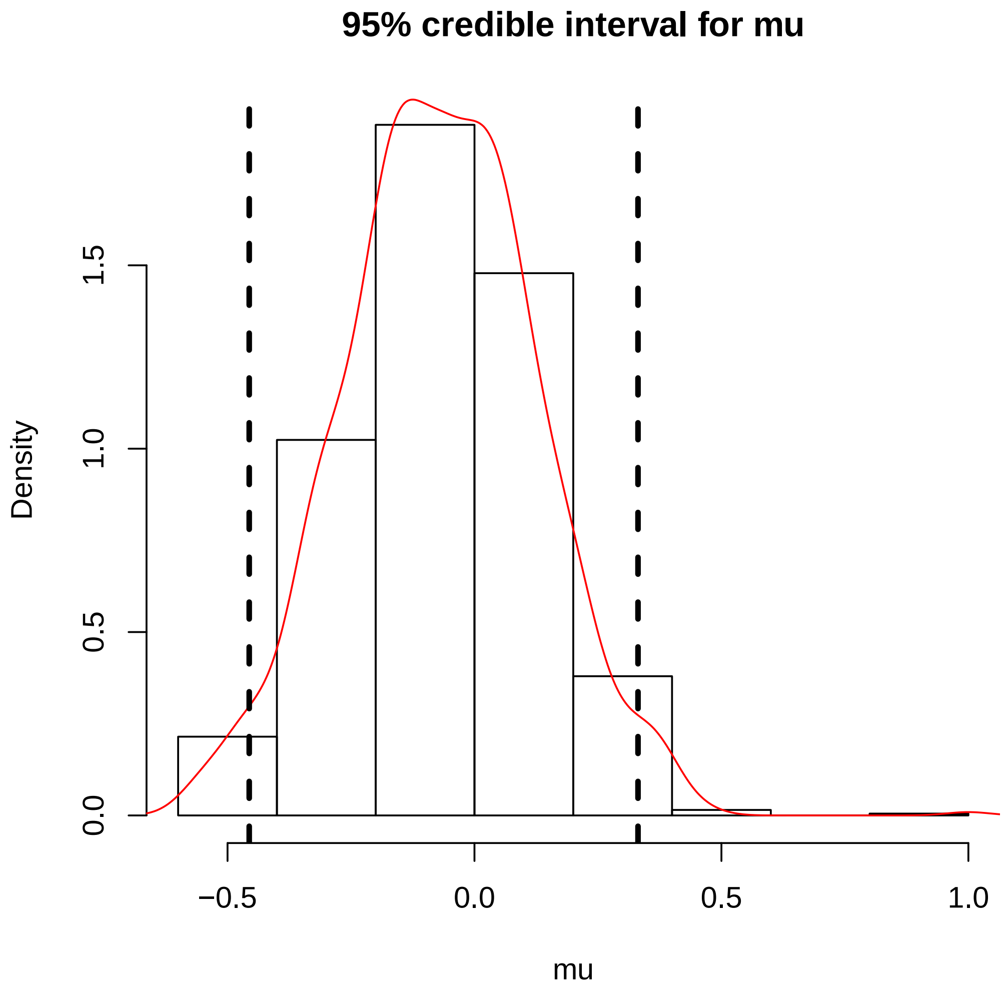
$$\operatorname{Var}\hat{\mu} = \frac{1}{n^2}\operatorname{Var}\sum_{i=1}^{n}\theta_i = \frac{1}{n}\operatorname{Var}_{\pi}\theta_i + \frac{1}{n^2}\sum_{k=1}^{n-1}(n-k)\operatorname{Cov}_{\pi}(\theta_1,\theta_{1+k}).$$

This means that we would need more than an estimate of $\operatorname{Var}_{\pi}\theta_i$ in order to estimate $\operatorname{Var}\hat{\mu}$. Furthermore, if the covariances tend to be larger than zero, which is common in a Markov chain, ignoring them could lead to a dramatic underestimation of the true variance—which in turn would lead to confidence intervals for $\tilde{\mu}$ that are too narrow!

In any event, as $m \to \infty$, any reasonable confidence interval for $\tilde{\mu}$ should get smaller. Such an interval captures only the error inherent in the MCMC approximation of the posterior (which diminishes as m grows); it does not capture the sampling distribution of the posterior distribution itself. We have been assuming throughout this argument that the data $\mathbf{Y}$ are fixed, but of course if we think more broadly, then $\mathbf{Y}$ is random, and a different sample leads to a different posterior. Can we capture this source of variation, which is distinct from MCMC error? One method is to create a 95% credible interval, which is merely a summary of the range of the middle 95% of the posterior distribution. For this purpose, we could proceed as follows:

```
> hist(mu, freq=FALSE, main="95% credible interval for mu")
> lines(density(mu), col=2)
> abline(v=print(quantile(mu, c(.025, .975))), lty=2, lwd=3)

      2.5%      97.5%
-0.4560731  0.3311048
```


95% credible interval for mu
Density
mu