LECTURE 23

Let $Y_1 \sim N(\beta_0 + \beta_1 x_1, \sigma^2)$
$Y_2 \sim N(\beta_0 + \beta_1 x_2, \sigma^2)$
...
$Y_n \sim N(\beta_0 + \beta_1 x_n, \sigma^2)$

where $\beta_0, \beta_1 \overset{iid}{\sim} N(0, \frac{\sigma^2}{m})$ and

$\sigma^2 \sim \text{InvGamma}\left(\frac{n_0}{2}, \frac{n_0 \sigma_0^2}{2}\right)$

It turns out for $Q := -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$, the posterior

$$P(\beta_0, \beta_1, \sigma^2 \mid y, X) \propto P(Y \mid \beta_0, \beta_1, \sigma^2, X)\, P(\beta_0, \beta_1, \sigma^2 \mid X)$$

$$\propto (\sigma^2)^{-\frac{n}{2} - 1 - \frac{n_0}{2} - 1} e^{-\frac{1}{2\sigma^2}\left(Q + m\beta_0^2 + m\beta_1^2 + n_0\sigma_0^2\right)}$$

which is a mess.
We do not know what the posterior is so we cannot sample from.

However, note that

$$P(\sigma^2 \mid \beta_0, \beta_1, X, y) \propto \text{InvGamma}\left(\frac{n+n_0}{2} + 1,\ Q + m\beta_0^2 + m\beta_1^2 + n_0\sigma_0^2\right)$$

$$P(\beta_0 \mid \beta_1, \sigma^2, X, y) \propto N\left(\frac{\bar{y}n - \beta_1 \bar{x}n}{n+m},\ \frac{\sigma^2}{n+m}\right) \quad \text{and}$$

$$P(\beta_1 \mid \beta_0, \sigma^2, X, y) \propto N\left(\frac{\sum x_i y_i - \beta_0 \bar{x} n}{\sum x_i^2 + m},\ \frac{\sigma^2}{\sum x_i^2 + m}\right)$$

Now we can use the Gibbs Sampling Algorithm, followed by Burn & Thinning to get independent samples of our posterior $P(\beta_0, \beta_1, \sigma^2 \mid y, X)$, from which we CAN SAMPLE FROM.

Suppose the true parameters are $\beta_0 = 2$, $\beta_1 = -2$, $\sigma^2 = 1$. Now we do not see this. Instead we see a sample of 3 points, $\langle x_1, y_1 \rangle$, $\langle x_2, y_2 \rangle$, $\langle x_3, y_3 \rangle$. Let us see if Gibbs sampling accompanied by B & T gives us distributions which are close to our true population parameters.
We first plot the three points & impose the line of best fit. (See page 3 in Prof. Kapelner's notes).

Using objective priors $m=1$, $n_0=1$, $\sigma_0^2=1$ in the box on the last page & then running Gibbs yields:

$$\left\langle \begin{bmatrix} \beta_{0,0} \\ \beta_{0,1} \\ \sigma_0^2 \end{bmatrix}, \begin{bmatrix} \beta_{1,0} \\ \beta_{1,1} \\ \sigma_1^2 \end{bmatrix}, \ldots, \begin{bmatrix} \beta_{10000,0} \\ \beta_{10000,1} \\ \sigma_{10000}^2 \end{bmatrix} \right\rangle \quad \star$$

On page 4 of Prof. K's notes, the first graph plots the chain $\langle \beta_{0,0}, \beta_{1,0}, \ldots, \beta_{10000,0} \rangle$
The 2nd graph plots the chain $\langle \beta_{0,1}, \beta_{1,1}, \ldots, \beta_{10000,1} \rangle$
3rd graph plots the chain $\langle \sigma_0^2, \sigma_1^2, \ldots, \sigma_{10000}^2 \rangle$.

From these graphs, we can deduce where our chains burned at. (The graphs look like hairy catapillars b/c that's when we converged) Note that in graph 3, some of the drawn σ^2's are very high up b/c they are drawn from an InverseGamma.

Now that we have burned, let us thin. To do so, we look at the autocorrelation plots of β_0, β_1 & σ^2 after convergence. (This is page 5 of Prof. K's notes). From looking at these graphs, we see $T=5$. After burning & thinning we get a subset of $\star$.

Plotting each chain yields the graphs on Page 6 of Prof. K's notes.

Recall Graph 1 consists of samples from $\beta_1 \mid X, y$
Graph 2 consists of samples from $\beta_2 \mid X, y$
Graph 3 consists of samples from $\sigma^2 \mid X, y$.

For each graph, we look at the mean. A 95% CR around it captures the true parameter! So this really works!

Let us look at another model.

Suppose we have n data points

$$\underbrace{1 \; 2 \; 3 \; \cdots \; m}_{\sim} \quad \underbrace{m+1 \; m+2 \; \cdots \; n}_{\sim}$$

s.t. $X_1, \ldots, X_m \overset{iid}{\sim} \text{Poisson}(\lambda_1)$ then $X_{m+1}, X_{m+2}, \ldots, X_n \overset{iid}{\sim} \text{Poisson}(\lambda_2)$
m is our stopping point.
Suppose we don't know where m is, & we don't know λ_1 & λ_2. We want to estimate m.

Since λ_1, λ_2 have a Poisson likelihood, a nice prior is a Gamma prior since this is conjugate. So let

$$\lambda_1, \lambda_2 \overset{iid}{\sim} \text{Gamma}(\alpha, \beta)$$

A good prior for m is $m \sim \text{uni}(1, 2, 3, \ldots, n)$

Now remember, λ_1, λ_2, & m are our parameters. It turns out

$$P(\lambda_1, \lambda_2, m \mid \vec{X}) \propto P(X_1, \ldots, X_n \mid \lambda_1, \lambda_2, m) \cdot P(\lambda_1) P(\lambda_2) P(m)$$

$$\propto e^{-m\lambda_1 - (n-m)\lambda_2 - \beta\lambda_1 - \beta\lambda_2} \lambda_1^{\alpha - 1 + \sum_{i=1}^{m} x_i} \lambda_2^{\alpha - 1 + \sum_{i=m+1}^{n} x_i}$$

(Note that our likelihood is $P(x_1, \ldots, x_n \mid \lambda_1, \lambda_2, m)$
$= P(x_1, \ldots, x_m \mid \lambda_1) \; P(x_{m+1}, \ldots, x_n \mid \lambda_2)$.)

Our posterior is too messy to sample from. But from the expression above, we see

$$P(\lambda_1 \mid \lambda_2, m, X) \propto \text{Gamma}\left(\alpha + \sum_{i=1}^{m} x_i, \; m + \beta\right)$$

$$P(\lambda_2 \mid \lambda_1, m, X) \propto \text{Gamma}\left(\alpha + \sum_{i=m+1}^{n} x_i, \; n - m + \beta\right) \quad \xi$$

$$P(m \mid \lambda_1, \lambda_2, X) \propto e^{m(\lambda_2 - \lambda_1)} \lambda_1^{\sum_{i=1}^{m} x_i} \lambda_2^{\sum_{i=m+1}^{n} x_i}$$

$\not\propto$ Known Distribution!

We haven't encountered a case like this before! Usually the conditionals were proportional to a known distribution. However $\text{Supp}[m] = \{1, 2, 3, \ldots, n\}$ is FINITE which makes it easy to grid sample. How? Like this:

STEP ONE: Evaluate $P(m=1 \mid \lambda_1, \lambda_2, X) = e^{1(\lambda_2 - \lambda_1)} \lambda_1^{\sum_{i=1}^{m} x_i} \lambda_2^{\sum_{i=m+1}^{n}}$

Evaluate $P(m=2 \mid \lambda_1, \lambda_2, X)$

...

Evaluate $P(m=n \mid \lambda_1, \lambda_2, X)$

STEP TWO: $N := \sum_{j=1}^{n} P(m=j \mid \lambda_1, \lambda_2, X)$.

STEP THREE: $P(m=j) = \frac{1}{N} P(m=j \mid \lambda_1, \lambda_2, X)$ for $j = 1, 2, \ldots, n$

NOTE we get the exact distribution b/c support is finite. ↳ not an approximation!

Hence we can run the Gibbs Sampler.

Now if $\lambda_1 = 2$, $\lambda_2 = 4$, $m = 10$, $n = 30$. Suppose we only see $n = 30$ and a sample drawn from, $x_1, x_2, \ldots, x_{30}$.
(This is page 9 in Prof. K's notes).
Our goal is to guess $\lambda_1 = 2$, $\lambda_2 = 4$, $m = 10$.
Instead of sampling from the 3-Dimensional conditional (which is nontrivial), we run the Gibbs sampler with $\alpha = 1$, $\beta = 1$.
This is page 10 on Prof. K's notes. The G.S. yields

$$\left\langle \begin{bmatrix} \lambda_{0,1} \\ \lambda_{0,2} \\ m_0 \end{bmatrix}, \begin{bmatrix} \lambda_{1,1} \\ \lambda_{1,2} \\ m_1 \end{bmatrix}, \ldots, \begin{bmatrix} \lambda_{10000,1} \\ \lambda_{10000,2} \\ m_{10000} \end{bmatrix} \right\rangle$$

Graph 1 plots $\lambda_{0,1}, \ldots, \lambda_{10000,1}$ (our chain of λ_1)
Graph 2 plots $\lambda_{0,2}, \ldots, \lambda_{10000,2}$ (our chain of λ_2)
Graph 3 plots $m_0, \ldots, m_{10000}$ (our chain of m)
We burn almost immediately.

Page 11 suggests we "thin out" with $T = 160$

Thinning & plotting the individual parameters yields the graphs on page 12.

Note if our C.R. does NOT include the true value of β_0, β_1 or σ^2, then this might be b/c the data has yet to overpower our prior.
Note the CR for m is Bad since $CR_{95\%,m} = [1, 29]$.
But this is a good property in the sense that the G.-S. is telling us that the sampler does NOT know m.

Imagine we have a model like this:

$$Y_t \sim \text{Poisson}\left(\lambda = \beta_0 + \beta_1 t\right)$$

Where our true population parameters are $\beta_0 = 1$, $\beta_1 = 3$
This is the linear line on page 16. Our goal is to guess $P(\beta_0, \beta_1 \mid X, t)$

Assume $p(\beta_0) = p(\beta_1) \propto 1$ ← noninformative prior

Then

$$P(\beta_0, \beta_1 \mid Y, t) \propto P(Y \mid \beta_0, \beta_1, t)\, P(\beta_0, \beta_1 \mid t)$$

$$\propto \prod_{i=1}^{n} (\beta_0 + \beta_1 t_i)^{y_i} \; e^{-(\beta_0 + \beta_1 t_i) y_i}$$

$\propto$ known distribution.

So we cannot sample from the posterior. Let us try to use the G.S.

So what about the conditionals?

SAME

$$P(\beta_0 \mid \beta_1, Y, t) \propto \prod_{i=1}^{n} (\beta_0 + \beta_1 t_i)^{y_i} \; e^{-(\beta_0 + \beta_1 t_i) y_i}$$

$$P(\beta_1 \mid \beta_0, Y, t) \propto \prod_{i=1}^{n} (\beta_0 + \beta_1 t_i)^{y_i} \; e^{-(\beta_0 + \beta_1 t_i) y_i}$$

Which did not simplify anything! Remember, Gibbs essentially works by conditioning on parameters so they become constants in the joint posterior distribution which simplifies to be proportional to some known distribution. Here, this did not work.

So the Gibbs Sampler has failed.
What do we do?

METROPOLIS HASTINGS ALGORITHM

$\beta_{i,j}$ ← i: guess #, j: parameter #

STEP ONE: Guess $\beta_{0,0}$ & $\beta_{0,1}$.

STEP TWO: Let q be your candidate density relying on some known parameters $\vec{\theta}$.
Draw $\beta_{1,0}$ from $q(\beta_{1,0}; \beta_{0,0}, \vec{\theta})$

($\beta_{1,0}$: free variable; $\beta_{0,0}, \vec{\theta}$: parameters)

Ex: $\beta_{1,0} \sim N(\beta_{0,0}, \sigma^2 = 1)$

STEP THREE: Compute

$$r := \frac{\dfrac{P(\beta_0 = \beta_{1,0}, \beta_1 = \beta_{0,1} \mid Y, t)}{q(\beta_{1,0}; \beta_{0,0}, \sigma^2)}}{\dfrac{P(\beta_0 = \beta_{0,0}, \beta_1 = \beta_{0,1} \mid Y, t)}{q(\beta_{0,0}; \beta_{1,0}, \sigma^2)}}$$

← posterior
← prob of going there
← prob of staying
← prob of going back to that state.

STEP FOUR: Accept $\beta_0 = \beta_{1,0}$ w. probability r

STEPS 5-7: Repeat for β_{11}

STEP EIGHT: Repeat iterations ② → ⑦

Note that r is

$$\frac{\text{prob of posterior if you use the draws}}{\text{prob of going from } \beta_{00} \to \beta_{10}} \Big/ \frac{\text{prob of sticking with previous value of parameter}}{\text{prob of going from } \beta_{10} \to \beta_{00}}$$

If r is small, then it is NOT a good idea to move to the value you drew.
If r is large, then it is a good idea to move.

How is this different from a Gibbs Sampler?
In the G.S. the β_0 is our θ_1
& the β_1 is our θ_2.

So r becomes $= 1$ meaning you always accept.

Now remember, the linear line on page 16 is from $Y_t \sim \text{Poisson}(\beta_0 + \beta_1 t)$ where $\beta_0 = 1$, $\beta_1 = 3$. Again, we don't see this. We see samples drawn from it which are our circles on that graph. We guess $\beta_{0,0} = 2$, $\beta_{0,1} = 2$.
Now we perform the algorithm above. We get a set

$$\left\langle \begin{bmatrix} \beta_{00} \\ \beta_{01} \end{bmatrix}, \begin{bmatrix} \beta_{00} \\ \beta_{01} \end{bmatrix}, \dots \right\rangle$$

where it is possible to have a string of equal terms b/c we are able to reject under this algorithm.

Plotting the chains yields page 17.
Plotting the autocorrelation yields page 18. Note that there is strong autocorrelation. Why? $T = 50$ since we are rejecting quite frequently. That is we are rejecting the drawn sample with probability $1-r$. So you are staying at the same spot. This happens because our candidate density is not a good indicator of where our parameters want to go. Since we are staying at the same value of β_0 or β_1 for long periods of time, $\exists$ high autocorrelation. Page 19 plots the values of β_0 & β_1 after burn & thinning.

Metropolized - within - Gibbs

Say you are using Gibbs and so you are conditioning. If you reach a step where you can't get the conditional, you use Metropolis.