LECTURE 22

Last time, we dealt with $P(\theta, \sigma^2 | X)$, i.e. 2 parameters. So a general posterior with k parameters is $P(\theta_1, \ldots, \theta_k | X)$. We have

$$P(\theta_1, \ldots, \theta_k | X) \propto P(X | \theta_1, \ldots, \theta_k)\, P(\theta_1, \ldots, \theta_k) \quad \text{given } X$$
$$\propto P(\theta_1, \ldots, \theta_k | X)$$
$$\propto K(\theta_1, \ldots, \theta_k | X)$$

Suppose we know

$$P(\theta_1 | \theta_2, \theta_3, \ldots, \theta_k, X)$$
$$P(\theta_2 | \theta_1, \theta_3, \ldots \theta_k, X)$$
$$\ldots$$
$$P(\theta_k | \theta_1, \theta_2, \ldots, \theta_{k-1}, X)$$

We then used the following algorithm:

GIBBS SAMPLING SYSTEMATIC SWEEP GIBBS SAMPLING

STEP ONE: Initialize (Guess) $\vec{\theta}_0 = \begin{bmatrix} \theta_{0,1} \\ \theta_{0,2} \\ \theta_{0,3} \\ \vdots \\ \theta_{0,k} \end{bmatrix}$ where $\theta_{i,j}$ (i: guess number, j: parameter number)

STEP TWO: Draw $\theta_{1,1}$ from $P(\theta_1 | \theta_2 = \theta_{0,2}, \ldots, \theta_k = \theta_{0,k}, X)$
$\theta_{1,2}$ from $P(\theta_2 | \theta_1 = \theta_{1,1}, \ldots, \theta_k = \theta_{0,k}, X)$
$\ldots$
$\theta_{1,k}$ from $P(\theta_k | \theta_1 = \theta_{1,1}, \ldots, \theta_{k-1} = \theta_{0,k-1}, X)$

STEP THREE: Repeat step 2 until convergence (recall the discussion about convergence from last time.)

We get an ordered set

$$\left\langle \begin{bmatrix} \theta_{0,1} \\ \theta_{0,2} \\ \vdots \\ \theta_{0,k} \end{bmatrix}, \begin{bmatrix} \theta_{1,1} \\ \theta_{1,2} \\ \theta_{1,k} \end{bmatrix}, \ldots\ldots, \begin{bmatrix} \theta_{\alpha,1} \\ \theta_{\alpha,2} \\ \theta_{\alpha,k} \end{bmatrix} \right\rangle$$

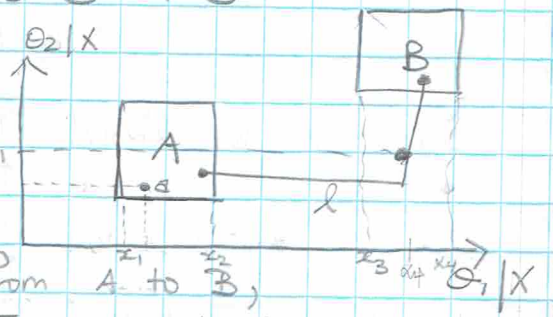
(α^{th} iteration)

where after the $t = B$ iteration transpires, the process converges. Then $\{\theta_{B,1}, \theta_{B+1,1}, \ldots, \theta_{B+N,1}\}$, $\{\theta_{B,2}, \theta_{B+1,2}, \ldots, \theta_{B+N,2}\}$, $\ldots$, $\{\theta_{B,k}, \theta_{B+1,k}, \ldots, \theta_{B+N,k}\}$ are samples of $P(\theta_1 | X), P(\theta_2 | X), \ldots, P(\theta_k | X)$ respectively.

Hence

$$\left\langle \begin{bmatrix} \theta_{B,1} \\ \theta_{B,2} \\ \vdots \\ \theta_{B,K} \end{bmatrix}, \begin{bmatrix} \theta_{B+1,1} \\ \theta_{B+1,2} \\ \vdots \\ \theta_{B+1,K} \end{bmatrix}, \ldots, \begin{bmatrix} \theta_{B+K,1} \\ \theta_{B+K,2} \\ \vdots \\ \theta_{B+K,K} \end{bmatrix} \right\rangle$$ are samples from $P(\theta_1, \theta_2, \ldots, \theta_K \mid X)$

Let us turn our attention to Problems ① & ② (i.e. Bad Mixing & Local Modes)



Suppose we have the following: If we start at the point $a \in A$, then we hop around in A & never reach B. WHY? Because to go from A to B, you must go through some line ℓ. This is outside the support of $\theta_1|X$ since $\theta_1|X$ does not assume values between x_2 & x_3. Hence we violate the POSITIVITY CONDITION! $f_X(x_{4+}) > 0$ but $f(x_{4+}, y_1) = 0$

Even if $\theta_1|X$ & $\theta_2|X$ assumed very low probabilities, we would run into this problem.

PROBLEM 3

Recall our samples once the algorithm converged.

$$\left\{ \begin{bmatrix} \theta_{B,1} \\ \vdots \\ \theta_{B,K} \end{bmatrix}, \begin{bmatrix} \theta_{B+1,1} \\ \vdots \\ \theta_{B+1,K} \end{bmatrix}, \ldots, \begin{bmatrix} \theta_{B+N,1} \\ \vdots \\ \theta_{B+N,K} \end{bmatrix} \right\}$$

We always indexed our samples after convergence by B. Why? Because B stands for BURNED IN.

The above is a BURNED IN CHAIN which refers to a chain of samples once the process converged.

However, these samples are NOT independent. Why?

$\theta_{2,1}$ is drawn from $P(\theta_1 \mid \theta_2 = \theta_{1,2}, \ldots, \theta_K = \theta_{1,K}, X)$
but $\theta_{1,2}$ was drawn from $P(\theta_2 \mid \theta_1 = \theta_{1,1}, \ldots, \theta_L = \theta_{0,K}, X)$
$\theta_{1,3}, \theta_{1,4}, \ldots, \theta_{1,K}$ all relied on $\theta_{1,1}$.
So $\theta_{2,1}$ depends on $\theta_{1,1}$. That is, the 2^{nd} draw of θ_1 depends on the first draw of θ_1 and in general, θ_t depends on θ_{t-1} so these samples

are not independent. How do we solve this?

Recall for r.v.'s X & Y,

$$\text{Corr}[X,Y] := \frac{\text{Cov}(X,Y)}{SE[X]\,SE[Y]} = \frac{E(X-\mu_X)\,E(Y-\mu_Y)}{SE(X)\;SE(Y)}$$

is estimated by

$$r := \frac{S_{xy}}{S_x S_y} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum (x_i-\bar{x})^2 \sum (y_i-\bar{y})^2}}$$

We will now define an analog to our situation:

Def: Autocorrelation → correlation with itself at different points in time. Informally, it is the similarity between observations as a function of the time lag between them.

Def: The sample lag 1 autocorrelation is

$$r_{a1} := \frac{\sum_{t=B}^{B+N-1} (\theta_t - \bar{\theta})(\theta_{t+1} - \bar{\theta})}{\sum_{t=B}^{B+N} (\theta_t - \bar{\theta})^2}$$

which measures how much θ_{t+1} depends on θ_t.

(Note $\bar{\theta} := \frac{1}{N} \sum_{t=1}^{B+N} \theta_t$)

Lag 2 autocorrelation is

$$r_{a2} := \frac{\sum_{t=B}^{B+N-2} (\theta_t - \bar{\theta})(\theta_{t+2} - \bar{\theta})}{\sum_{t=B}^{B+N} (\theta_t - \bar{\theta})^2}$$

which measures how much θ_{t+2} depends on θ_t.

and in general, lag l autocorrelation is

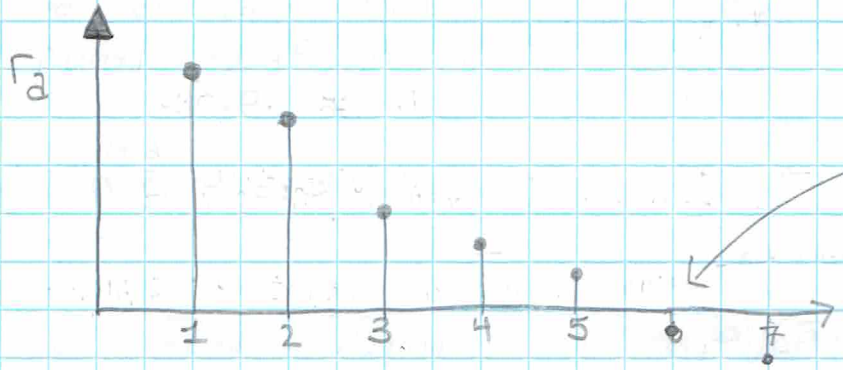
$$r_{al} = \frac{\sum_{t=B}^{B+N-l} (\theta_t - \bar{\theta})(\theta_{k+l} - \bar{\theta})}{\sum_{t=B}^{N} (\theta_t - \bar{\theta})}$$

Now remember we are looking at

$$\left\{ \begin{bmatrix} \theta_{B,1} \\ \theta_{B,2} \\ \vdots \\ \theta_{B,K} \end{bmatrix}, \begin{bmatrix} \theta_{B+1,1} \\ \theta_{B+1,2} \\ \vdots \\ \theta_{B+1,K} \end{bmatrix}, \ldots\ldots, \begin{bmatrix} \theta_{B+N,1} \\ \theta_{B+N,2} \\ \vdots \\ \theta_{B+N,K} \end{bmatrix} \right\}$$

Recall to assess convergence we looked at the "rows", i.e., we looked at the chains, meaning you look at the values of your first parameter across the board. Similarly, we look at each chain to assess autocorrelation.

For each chain we get the following picture:



We say lag 6 is INCONSEQUENTIAL "TOO SMALL TO MATTER" We denote this point by T_j.

The autocorrelation goes down b/c the 2nd & 4th position are more related than the 2nd & 50th position.

Define $T = \max\{T_1, T_2, \ldots, T_k\}$

Now from the Burned In (i.e. converged) set,

$$\left\{ \begin{bmatrix} \theta_{B,1} \\ \vdots \\ \theta_{B,K} \end{bmatrix}, \begin{bmatrix} \theta_{B+1,1} \\ \vdots \\ \theta_{B+1,K} \end{bmatrix}, \ldots \begin{bmatrix} \theta_{B+N,1} \\ \vdots \\ \theta_{B+N,K} \end{bmatrix} \right\}$$

We extract a SEQUENCE (which makes the above set smaller & hence THINS the Burned in set) by starting with B & taking $B+T$, $B+2T$, ... as follows:

$$\left\{ \begin{bmatrix} \theta_{B,1} \\ \vdots \\ \theta_{B,K} \end{bmatrix}, \begin{bmatrix} \theta_{B+T,1} \\ \vdots \\ \theta_{B+T,K} \end{bmatrix}, \begin{bmatrix} \theta_{B+2T,1} \\ \vdots \\ \theta_{B+2T,K} \end{bmatrix}, \dots \right\}$$

DRAWS FROM MARGINAL DENSITY IF WE DROP OTHER PARAMETERS (OR) DRAWS FROM POSTERIOR.

We call the above our BURNED IN AND THINNED MODEL. These are independent samples from our posterior.

$$\{\theta_{B,1}, \theta_{B+T,1}, \dots, \theta_{B+2T,1}, \dots\}$$ ← 1 Burned & Thinned chain.

$$\{\theta_{B,2}, \theta_{B+T,2}, \dots, \theta_{B+2T,2}, \dots\}$$

$$\dots$$

$$\{\theta_{B,K}, \theta_{B+T,K}, \dots, \theta_{B+2T,K}, \dots\}$$

} Burned & Thinned CHAINS

For each parameter, $\theta_1, \theta_2, \dots \theta_K$, we can compute

$$E[\theta_j | X] \approx \frac{1}{N} \sum_{BTC} \theta_{t,j}$$

$\hat{\theta}_{mmse} = med(\theta_j | X) \approx$ middle $\theta_{t,j}$ after sorting

$$P(\theta_j > c | X) = \frac{1}{N} \sum_{BTC} \mathbb{1}_{\theta_j > c} = \frac{\text{frequency}}{\text{sum}}$$

$\hat{\theta}_{j,MAP} \approx$  pick width & take midpoint.

How would we compute $P(X^* | X)$? where $\dim[X^*] = n$

Well

$$P(X^*|X) = \int_{\theta_1} \dots \int_{\theta_K} P(X^*|\theta_1, \theta_2, \dots, \theta_K) P(\theta_1, \dots, \theta_K | X) \, d\theta_1 \dots d\theta_K$$

We cannot do this integral, we only have samples. We have the following procedure:

① Run Gibb Sampler to get Burned & Thinned Chains.

② We get N samples.

③ Compute

$$\frac{1}{N} \left[P(X^* | \theta_1 = \theta_{B,1}, \dots \theta_K = \theta_{B,K}) + \dots + P(X^* | \theta_1 = \theta_{B+nT,1}, \dots, \theta_K = \theta_{B+nT,K}) \right.$$