Lecture 22 5/11/16 Math 390.03-02

Posterior Inference

values of posterior samples after sharing

① $E[\Theta | X] \approx \frac{1}{N_T} \sum_{i=1}^{N_T} \theta_i$

② $\text{Quantile}[\Theta | X, p] = \{ p \cdot N_T^{th}$ value after sorting $\}$

Covered in Lecture 21

③ 95% CR = ditto

④ Hypothesis tests, etc .. same as before

Gibbs sampling offers inference for previously intractable models!

Gibbs Sampling algorithm (systematic sweep) Lecture START

$$p(\theta_1, \ldots, \theta_K | X) \propto p(X | \theta_1, \ldots, \theta_K)\, p(\theta_1, \ldots, \theta_K) \propto k(X | \theta_1, \ldots, \theta_K)$$

Assume

$p(\theta_i | \theta_{-i}, X)$ can be sampled from $\forall i$

Then Gibbs Sampler is:

Step 1. Initialize $\vec{\theta}_0 = [\theta_{1,0}, \ldots, \theta_{K,0}]$

Step 2: Draw $\theta_{1,1}$ from $p(\theta_1 | \theta_{2,0}, \ldots, \theta_{K,0}, X)$

Draw $\theta_{1,2}$ from $p(\theta_2 | \theta_{1,1}, \theta_{3,0}, \ldots, \theta_{K,0}, X)$

$\theta_{1,K}$ from $p(\theta_K | \theta_{1,1}, \ldots, \theta_{1,K-1}, X)$

Step 3: Record $\vec{\theta}_1$ as a sample Step 4: Repeat 2-3 until "convergence"!!

Proof of convergence

Def: Markov Chain on space $\mathcal{X}$ is a seq. of r.v.'s $X_0, X_1, \ldots$ s.t.

$$P(X_{t+1} \in A \mid X_0, X_1, \ldots, X_t) = P(X_{t+1} \in A \mid X_t) \quad \forall A \subseteq \mathcal{X}$$

Once you're at a certain state, you lose memory of all prior states.

Further ... $= \int_A P(X_{t+1} \mid X_t) f(X_t)\,dx$ where $P(X_{t+1} \mid X_t)$ is called the "transition kernel"

If $\mathcal{X}$ is ctbl, it's called the "transition matrix".

Def: f is the chain's stationary distribution if:

$$f(X_{t+1}) = \int_{\mathcal{X}} f(X_t)\, P(X_{t+1} \mid X_t)\,dx$$

Thm: $\forall\, g(X_0)$ (i.e. starting position), $f(x) = \lim_{T \to \infty} \int_{\mathcal{X}} \prod_{t=0}^{T} P(X_{t+1} \mid X_t)\, g(X_0)\,dx$ (stationary distr) w/ reg. cond's.

$\mathcal{X}$ ← avg. over everywhere X_t could be!

$\Rightarrow$ No matter where the Markov chain begins, it converges to the same distr.

Def: A jdf $f(X_1, \ldots, X_k)$ satisfies the positivity condition if $f(X_i) > 0 \;\forall\, i = 1 \ldots k$

$$\Rightarrow f(x_1, \ldots, x_k) > 0 \iff \mathrm{Supp}[X_1, \ldots, X_k] = \mathrm{Supp}[X_1] \times \ldots \times \mathrm{Supp}[X_k]$$

Thm: If $f(x_1, \ldots, x_k)$ satisfies the pos. cond., then $\forall\, (a_1, \ldots, a_k) \in \mathrm{Supp}[X_1, \ldots, X_k]$

$$\Rightarrow f(x_1, \ldots, x_k) \propto \prod_{i=1}^{K} \frac{f(x_j \mid x_1, \ldots, x_{j-1}, x_{j+1} = a_{j+1}, \ldots, x_K = a_K)}{f(x_j = a_j \mid \;''\;\;''\;\;''\;\;'')} > 0$$

Con: If jdf has pos, all conditional densities are non-zero. If not, then $\perp$.

$\Rightarrow$ Before running Gibbs sampler, ask... is there mass everywhere in the support???

Thm: the kernel of a Gibbs sampler can be expressed as

Recall the Gibbs sampler in K dimensions ... the transition kernel would be:

$$P(\underbrace{\theta_{t+1,1}, \dots \theta_{t+1,K}}_{\text{future iteration}} \mid \underbrace{\theta_{t,1}, \dots \theta_{t,K}}_{\text{previous iteration}}, \underbrace{X}_{\text{data}})$$

$\theta_{1,0}\ \theta_{2,0}\ \cdots\ \theta_{K,0}$	$t=0$
$\theta_{1}\ \theta_{2,0}\ \cdots\ \theta_{K,0}$	$t=1.1$

$$= P(\theta_{t+1,K} \mid \theta_{t+1,1}, \dots, \theta_{t+1,K-1}, X) \cdot$$
$$P(\theta_{t+1,K-1} \mid \theta_{t+1,1}, \dots \theta_{t+1,K-2}, \theta_{t,K}, X) \cdot$$
$$P(\theta_{t+1,K-2} \mid \theta_{t+1,1}, \dots, \theta_{t+1,K-3}, \theta_{t,K-1}, \theta_{t,K}, X) \cdot \dots \cdot$$
$$P(\theta_{t+1,2} \mid \theta_{t+1,1}, \theta_{t,2}, \dots, \theta_{t,K}, X) \cdot$$
$$P(\theta_{t+1,1} \mid \theta_{t,2}, \dots \theta_{t,K}, X)$$

K steps

$\Rightarrow$ a Gibbs sampler is a Markov chain

Thm: The Gibbs sampler converges to the posterior if $f(\theta_1, \dots, \theta_K \mid X)$ is the "invariant distr." Proof:

$$f(\theta_1 = \theta_{t+1,1}, \dots \theta_K = \theta_{t+1,K} \mid X) \overset{?}{=} \int_{\theta_1} \cdots \int_{\theta_K} f(\theta_1 = \theta_{t,1}, \dots \mid X)\,(\text{kernel})\, d\theta_{t,1} \cdots d\theta_{t,K}$$

Recall since $f(\theta_1, \dots \theta_K \mid X) > 0 \Rightarrow$ all cond's are > 0

$$\int_{\theta_2} f(\theta_1 = \theta_{t+1,1}, \theta_2 = \theta_{t,2}, \dots \theta_K = \theta_{t,K} \mid X)\, d\theta_{t,2}$$
$$\|$$
$$P(A)\, P(B \mid A) = P(B, A)$$

$$= \int_{\theta_2} \cdots \int_{\theta_K} \int_{\theta_1} f(\theta_1 = \theta_{t,1}, \dots \theta_K = \theta_{t,K} \mid X)\, d\theta_{t,1}\ \text{kernel}\ d\theta_{t,2} \cdots d\theta_{t,K} = \int_{\theta_3} \cdots \int_{\theta_K} \int_{\theta_2} f(\theta_2 = \theta_{t,2}, \dots \theta_K = \theta_{t,K} \mid X)\, P(\theta_{t+1,1} \mid \theta_{t,2}, \dots \theta_{t,K}, X)\, d\theta_{t,2} \cdots d\theta_{t,K}$$

rest of kernel

$$= \int_{\theta_4} \cdots \int_{\theta_K} \underbrace{\int_{\theta_3} f(\theta_1 = \theta_{t+1}, \theta_3 = \theta_{t,3}, \dots, \theta_K = \theta_{t,K} \mid X)\, P(\theta_{t+1,2} \mid \theta_{t+1,1}, \theta_{t,2}, \dots \theta_{t,K}, X)\, d\theta_{t,3}}_{\int_{\theta_3} f(\theta_1 = \theta_{t+1}, \theta_2 = \theta_{t+1,2}, \theta_3 = \theta_{3,t}, \theta_4 = \theta_{4,t}, \dots \theta_K = \theta_{K,t} \mid X)\, d\theta_3} \cdots d\theta_{t,4} \cdots d\theta_{t,K}$$

rest of kernel

$$= \int_{\theta_5} \cdots \int_{\theta_K} \int_{\theta_4} f(\theta_1 = \theta_{t+1}, \theta_2 = \theta_{t+1,2}, \theta_4 = \theta_{4,t}, \dots \theta_K = \theta_{t,K} \mid X) \cdots d\theta_{t,4} \cdots d\theta_{t,K}$$

$\vdots$

$$= f(\theta_1 = \theta_{t+1,1}, \dots, \theta_K = \theta_{t+1,K} \mid X) \quad \blacksquare$$

So it converges to stationary distr which is the pdf we seek. Let B be the iteration of convergence, thus...

$$\left\{ \begin{bmatrix} \theta_B \\ \sigma^2_B \end{bmatrix}, \ldots, \begin{bmatrix} \theta_{B+N} \\ \sigma^2_{B+N} \end{bmatrix} \right\}$$

Constitute samples from $P(\theta, \sigma^2 | X)$, the density which previously defined

$B :=$ "Burn in"

$N := \#$ of posterior samples

Simulation

$\boxed{\text{Gibbs problem \#3}}$ θ_{t+1} is dependent on θ_t which is dep on θ_{t-1}, etc.

At what point are they independent?

Recall the $\text{Corr}[X, Y] := \dfrac{\text{Cov}[X,Y]}{SE[X]\,SE[Y]} := \dfrac{E(X-\mu_x)(Y-\mu_Y)}{\sqrt{\text{Var}[X]\,\text{Var}[Y]}}$

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

Here we care about "autocorrelation"

auto - self (Greek prefix)

automatic

self move

For θ, Autocorrelation for lag 1 is:

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

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

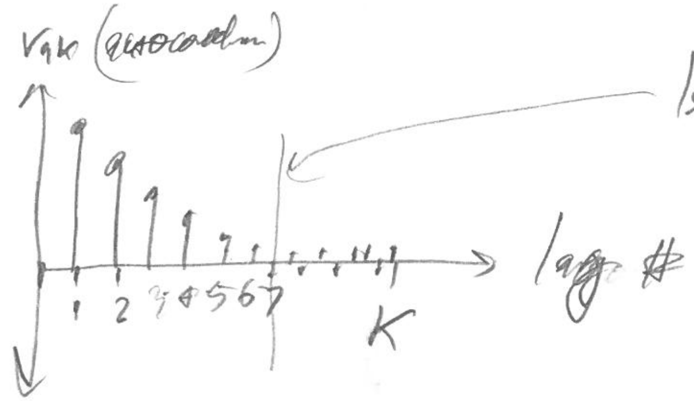
and autocorrelation for lag 2 is:

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

And for the k^{th} lag:

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

Pick a max k, and look at



looks like no meaningful autocorrelation is found at $k \geq 7 = T$

$\Rightarrow$ "Thin" the chains by throwing out all non-multiples of 7:

$$\left\{ \begin{bmatrix} \theta_B \\ \sigma^2_B \end{bmatrix}, \ldots\ldots, \begin{bmatrix} \theta_{B+7} \\ \sigma^2_{B+7} \end{bmatrix}, \ldots, \begin{bmatrix} \theta_{B+14} \\ \sigma^2_{B+14} \end{bmatrix}, \ldots \right\}$$

$$\left\{ \begin{bmatrix} \theta_B \\ \sigma^2_B \end{bmatrix}, \begin{bmatrix} \theta_{B+7} \\ \sigma^2_{B+7} \end{bmatrix}, \begin{bmatrix} \theta_{B+14} \\ \sigma^2_{B+14} \end{bmatrix}, \ldots \right\}$$

Our "burned and thinned" chains

which can be used as draws from joint or marginal densities.

$\hat{\theta}_{1,MAP}$? Bin:

take midpoint of largest bin

$P(X^*|X)$?

$$= \int P(X^*|\theta_1,\dots,\theta_k)\, P(\theta_1,\dots,\theta_k|X)\, d\theta_1\dots d\theta_k$$

Procedure

① Run gibbs sampler until convergence and get "burned-in" chain

② Sample $\vec{\theta}$ from chain $\vec{\theta}_1$

③ Sample X^* from Lik. model, X^*_1 (most often this is possible)

④ Repeat the above steps 2-3 N times, $X^*_1, \dots, X^*_N$

What if you could want...

$$P(X^* \geq c|X) \approx \frac{1}{N}\sum \mathbb{1}_{X^* \geq c}$$

First example was

$X|\theta,\sigma^2 \sim N(\theta,\sigma^2)$

$\theta \sim N(\mu_0,\tau^2)$, $\sigma^2 \sim \text{InvGamma}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$