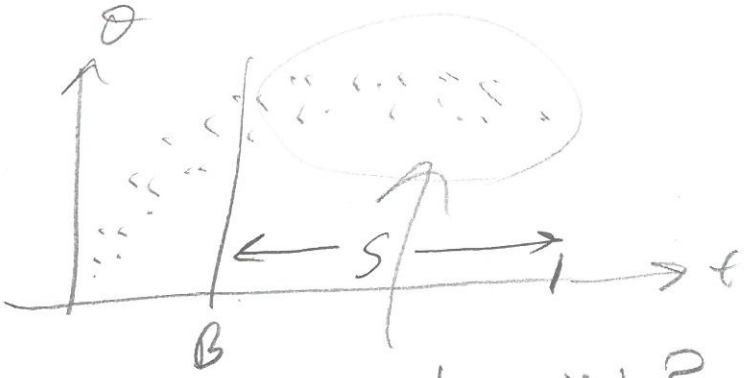
Lec 20 Math 341 5/2/18 Gibbs Sampling



Are these iid? No!

They are dependent on the previous realizations.

How do we measure dependence?

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \in [-1, 1]$$ if $r \approx 0 \Rightarrow X, Y$ mostly independent

---

Use this to see how are iterates depends on the previous...

$$r_{g1} := \frac{\sum_{t=B+1}^{B+S-1} (\theta_t - \bar{\theta})(\theta_{t+1} - \bar{\theta})}{\sum_{t=B+1}^{B+S} (\theta_t - \bar{\theta})^2}$$

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

this is called "autocorrelation" with lag = 1

self!

It measures how θ_t varies with θ_{t-1} - What do you think r_{g1} will be? Large & Positive!

Autocorrelation for lag 2 is:

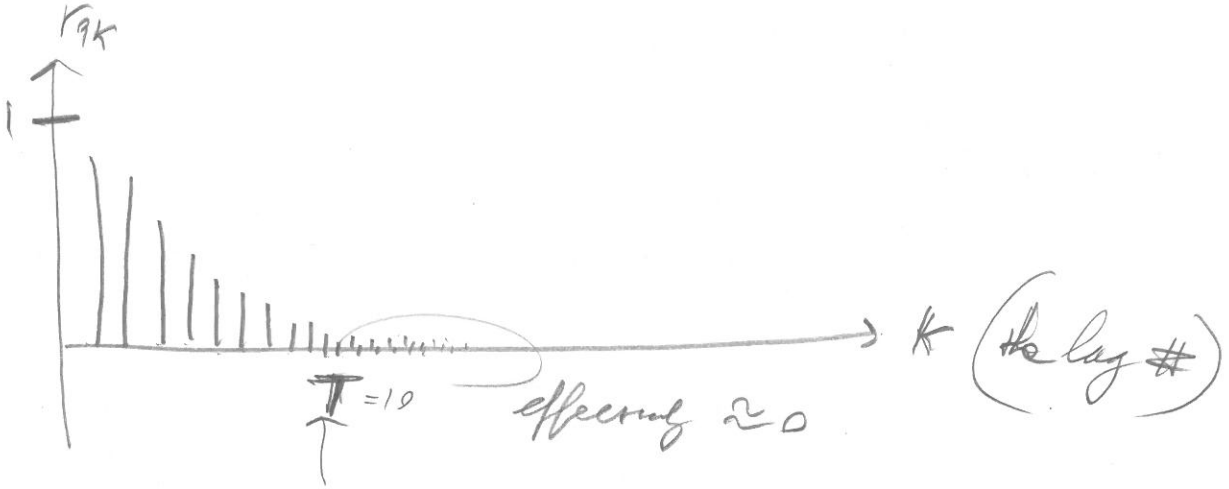
$$r_{g2} := \frac{\sum_{t=B}^{B+S-2} (\theta_t - \bar{\theta})(\theta_{t+2} - \bar{\theta})}{\sum_{t=B}^{B+S} (\theta_t - \bar{\theta})^2}$$

$$\vdots$$

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

at some k, $r_{gk} \simeq 0$. Why? Eventually the dependence is lost...

How to see? Autocorrelation plot



at around T iterations ago, the draws effectively are independent of

In order to make the chain report all independent samples from posterior, we need to throw out all samples except those that are multiple of T after B. This is known as "thinning".

$$\left\{ \begin{bmatrix} \theta_B \\ \sigma^2_B \end{bmatrix}, \begin{bmatrix} \theta_{B+T} \\ \sigma^2_{B+T} \end{bmatrix}, \begin{bmatrix} \theta_{B+2T} \\ \sigma^2_{B+2T} \end{bmatrix}, \dots \right\} \rightarrow \text{the burned and thinned chain.}$$

let $\ell = 1 \ldots L$ be the index on the burned and thinned chain. This is almost as good as having $P(\theta|x)$ directly:

How to get $\hat{\theta}_{MMSE} = E(\theta|x) \approx \bar{\theta} = \frac{1}{L}\sum_{\ell=1}^{L} \theta_\ell$

$\hat{\theta}_{MAE} = \text{Med}(\theta|x)$ = order from smallest to largest $\theta_{(1)}, \ldots, \theta_{(L)}$ and return $\theta_{\left(\frac{L}{2}\right)}$ (round)

$CR_{\theta, 1-\alpha}$ = order from smallest to largest

$$\left[\theta_{\left(\frac{\alpha}{2}L\right)},\ \theta_{\left(\left(1-\frac{\alpha}{2}\right)L\right)}\right]$$

round... round..

$$pval = P(H_0|x) = P(\theta \in \Theta_0|x) \approx \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{1}_{\theta_\ell \in \Theta_0}$$

e.g. $P(\theta > 3|x) \approx \frac{1}{L}\sum_{\ell=1}^{L} \mathbb{1}_{\theta_\ell > 3}$ i.e. prop. of L s.t. $\theta > 3$

$$P(X^*|x)\ ? = \int_{\Theta} P(X^*|\theta)\, P(\theta|x)\, d\theta$$

to sample from this...

① Pick $\ell \in \{1, \ldots, L\}$ to get θ_ℓ

② Draw X^* from $P(X^*|\theta = \theta_\ell)$

③ Repeat steps 1,2 over and over...

Def: Systematic Sweep Gibbs Sampler

Assume posterior $P(\theta_1, \ldots, \theta_p|X)$ is unknown but

$P(\theta_j \mid \theta_{-j}, X)$ s.t. $\theta_{-j} = \{\theta_1, \ldots \theta_{j-1}, \theta_{j+1}, \ldots, \theta_p\}$ i.e. all θ's except θ_j

is known $\forall j$.

Step 1: Initialize $\vec{\theta}_0 = \langle \theta_{1,0}, \theta_{2,0}, \ldots, \theta_{p,0} \rangle$

Step 2: Sample $\theta_{1,1}$ from $P(\theta_1 \mid \theta_2 = \theta_{2,0}, \ldots \theta_p = \theta_{p,0}, X)$

sample $\theta_{2,1}$ from $P(\theta_2 \mid \theta_1 = \theta_{1,1}, \theta_3 = \theta_{3,0}, \ldots, \theta_p = \theta_{p,0}, X)$

$\vdots$

$\theta_{p,1}$ from $P(\theta_p \mid \theta_1 = \theta_{1,1}, \ldots \theta_{p-1} = \theta_{p-1,1}, X)$

and record $\langle \theta_{1,1}, \ldots \theta_{p,1} \rangle$ as a "sample"

Step 3: repeat step 2 for many times...

The claim is that the samples, given $t \to \infty$ come from $P(\theta_1, \ldots, \theta_p \mid X)$. Proof...

Def. Consider $X_0, X_1, X_2, \ldots$ a sequence of r.v.'s (scalar or vector) with support $\mathcal{X}$.

If $P(X_t \in A \mid X_{t-1}, X_{t-2}, X_{t-3}, \ldots, X_{t-s}) = P(X_t \in A \mid X_{t-1})$ $\forall t, s$ $\forall A \subset \mathcal{X}$

then $X_1, X_2, \ldots$ is called a "Markov Chain".

(trans. matrix of discr. / trans kernel of cont.)

Is the Gibbs sampler a Markov Chain?? YES!!! "MCMC"

Def. A Markov chain's "invariant distribution" is defined as:

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

margin out the effect of the previous value

same $= P_{inv}(X)$ known

$$= \int_{\mathcal{X}} P(X_{t+1} \mid X_t)\, P(X_t \mid X_{t-1})\, P(X_{t-1})\, d\ldots$$

$$= \int \left(\prod_{i=0}^{t+1} P(X_i \mid X_{i-1}) \right) P(X_0)\, dx$$

Cool equation huh?

Thm. for any starting distr. $P(X_0)$,

$$P(X) = \lim_{t\to\infty} \int_{\mathrm{Supp}(X)} \prod_{i=0}^{t+1} P(X_i|X_{i-1})\, P(X_0 = x)\, dx$$

$\Rightarrow$ doesn't matter where you start, given enough "hops" or time, you wind up in the same steady state distr. AKA "long-term" / "limiting" / "stationary" distr. / "equilibrium"

Ref: A jdf $P(X_1, \dots, X_p)$ has the positivity cond.

if $\forall j, P(X_j) > 0 \quad \forall x_j \in \mathrm{Supp}(X_j)$

Thm: Consider $P(X_1, \dots, X_p)$ which has the positivity cond. then, $\forall \vec{q} \in \mathrm{Supp}(\vec{X})$,

$$P(X_1, \dots, X_p) \propto \prod_{j=1}^{p} \frac{P(X_j \mid X_1, \dots, X_{j-1}, X_{j+1} = q_{j+1}, \dots, X_p = q_p)}{P(X_j = q_j \mid X_1, \dots, X_{j-1}, X_{j+1} = q_{j+1}, \dots, X_p = q_p)}$$

Corr: If $P(X_1, \dots, X_p)$ has the positivity cond.

$\Rightarrow P(X_j \mid X_{-j}) > 0 \quad \forall x_j \in \mathcal{X}$ i.e. all cond. densities are nonzero.

We need this for the proof...

What is the transition kernel for the Gibbs sampler?

$$P(\vec{\theta}_{t+1} \mid \vec{\theta}_t, X) = P(\theta_{t+1,1}, \dots, \theta_{t+1,p} \mid \theta_{t,1}, \dots, \theta_{t,p}, X)$$

I will not write X anymore... pretend it is there