Math 650. 5/11/17

Algorithm: Systematic Sweep
: Gibbs Sampler for $P(\theta_1, \dots, \theta_p | X)$, the unknown posterior w/ p parameters.

Here all conditionals, $P(\theta_j | \theta_{-j})$ where $\theta_{-j} = \{\theta_1, \dots, \theta_{j-1}, \theta_{j+1}, \dots, \theta_p\}$ are known and can be "easily" sampled from.

step 1: Initialize $\theta = (\theta_{0,1}, \theta_{0,2}, \dots \theta_{0,p})$ (conditional on data X...)

step 2: Sample $\theta_{1,1}$ from $P(\theta_{0,1} | \theta_2 = \theta_{0,2}, \dots, \theta_p = \theta_{0,p})$ — $P(\vec{\theta}_{t+1} | \vec{\theta}_t)$

Sample $\theta_{1,2}$ from $P(\theta_{0,2} | \theta_1 = \theta_{1,1}, \theta_3 = \theta_{0,3}, \dots \theta_p = \theta_{0,p})$

⋮

Sample $\theta_{1,p}$ from $P(\theta_p | \theta_1 = \theta_{1,1}, \dots, \theta_{p-1} = \theta_{1,p-1})$

step 3: Repeat 2 until "convergence".

(Not Required) Proof: Consider $X_0, X_1, X_2, \dots$ a sample of random variables.
Each have sample X. If $P(X_t \in A | X_{t-1}, X_{t-2}, \dots, X_0) = P(X_t \in A | X_{t-1}) \; \forall t, \forall A \subset X$
then the sample sequence is called a discrete-time Markov chain.
The Gibbs sampler is a Markov Chain.
This is why the Gibbs sampler is a form of "Markov Chain Monte Carlo" MCMC.

$$P(X_{t+1}) = \int_X P(X_{t+1}, X_t)\, dx = \int_X \underbrace{P(X_{t+1} | X_t)}_{\text{transition kernel}} P(X_t)\, dx$$

If $P(X_{t+1}) \overset{X}{=} P(X_t)$ then this distribution is deemed the "invariant" "equilibrium". "stationary". "long-term".

$$P(X_{t+1}) = P(X_t | X_{t-1}) P(X_{t-1} | X_{t-2}) \cdots P(X_1 | X_0) P(X_0)$$

integrating whole thing

$$P(X)_{inv} = \lim_{t \to \infty} \int_X dx_0$$ you'll get invariant distribution.

$$= P(\theta_{t+1,1} | \theta_{t,2}, \dots, \theta_{t,p}) \cdot P(\theta_{t+1,2} | \theta_{t+1,1}, \theta_{t,3}, \dots, \theta_{t,p}) \cdot \dots \cdot P(\theta_{t+1,p-1} | \theta_{t+1,1}, \dots \theta_{t+1,p-1}, \theta_{t,p}) \cdot$$
$$P(\theta_{t+1,p} | \theta_{t+1,1}, \dots \theta_{t+1,p-1}, \theta_{t+1,p})$$

(vector notation) $P(\vec{\theta}_{t+1}) = \int P(\vec{\theta}_{t+1} | \vec{\theta}_t) \cdot P(\vec{\theta}_t)\, d\vec{\theta}$

(scalar notation) $P(\theta_{t+1,1}, \dots \theta_{t+1,p}) = \int_{H_1} \int_{H_2} \cdots \int_{H_p} \cdot P(\theta_{t+1,1}, \dots \theta_{t+1,p}) \cdots d\theta_1 \cdots d\theta_p.$

$$= \int_{H_2} \cdots \int_{H_p} \text{kernel} \int_{H_1} \underbrace{P(\theta_{t,1} | \theta_{t,2}, \dots, \theta_{t,p}) \cdot P(\theta_{t,2}, \dots, \theta_{t,p})}_{P(\theta_1, \dots \theta_p)} \theta_1)$$
$$= 1$$
$$P(\theta_1, \dots \theta_p)\, d\theta_2 \cdots d\theta_p.$$

$$= \int_{H_5} \int_{H_p} \text{rest of kernel} \left[\int_{H_2} p(\theta_{t+1,1} \mid \theta_{t,2}, \dots \theta_{t,p})\, p(\theta_{t,2}, \dots, \theta_{t,p})\, d\theta_t \right] d\theta_5 \cdots d\theta_p$$

$$\frac{p(\theta_{t+1}, \theta_{t+2}, \theta_{t,p})}{p(\theta_{t+2}, \theta_{t,3}, \dots \theta_{t,p})}$$

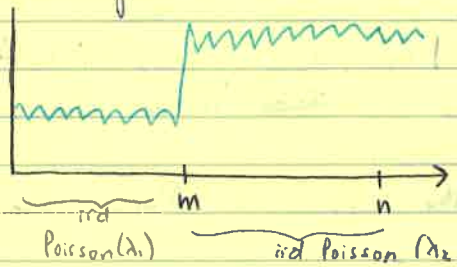
$$= \int_{H_4} \cdots \int_{H_p} \text{rest of kernel} \left[\int_{H_3} p(\theta_{t+1,2} \mid \theta_{t+1,1}, \theta_{t,3}, \dots \theta_{t,p})\, p(\theta_{t+1,1}, \theta_{t,3}, \dots \theta_{t,p})\, d\theta_4 \cdots d\theta_p \right]$$

$$\frac{p(\theta_{t+1,1}, \theta_{t+1,2}, \theta_{t,3}, \dots \theta_{t,p})}{p(\theta_{t+1,1}, \theta_{t+1,2}, \theta_{t,4}, \dots \theta_{t,p})}$$

and the process repeats ... proof of eventual convergence.

$$= p(\theta_{t+1,1}, \dots, \theta_{t+1,p})$$

Change Point Model



Parameters: λ_1 - mean of "first process"
λ_2 - mean of "second process"
m - "change point."

Priors: $p(\lambda_1) = \text{Gamma}(\alpha, \beta)$
$p(\lambda_2) = \text{Gamma}(\alpha, \beta)$
$p(m) = \text{Uniform}\{0, \dots, n\} = \frac{1}{n} \quad \forall\ m$

Posterior: $p(\lambda_1, \lambda_2, m \mid X_1 \cdots X_n) \propto p(X_1 \cdots X_n \mid \lambda_1, \lambda_2, m) \cdot p(\lambda_1, \lambda_2, m)$

$$\propto \left(\prod_{i=1}^{m} \frac{e^{-\lambda_1} \lambda_1^{X_i}}{X_i!} \right) \left(\prod_{i=m+1}^{n} \frac{e^{-\lambda_2} \lambda_2^{X_i}}{X_i!} \right) p(\lambda_1)\, p(\lambda_2)\, p(m)$$

$$\propto \left(\prod_{i=1}^{m} \frac{e^{-\lambda_1} \lambda_1^{X_i}}{X_i!} \right) \left(\prod_{i=m+1}^{n} \frac{e^{-\lambda_2} \lambda_2^{X_i}}{X_i!} \right) \left(\lambda_1^{\alpha-1} e^{-\beta\lambda_1} \right) \left(\lambda_2^{\alpha-1} e^{-\beta\lambda_2} \right)$$

$\sum_{i=1}^{m} X_i = a$, $\sum_{i=m+1}^{n} X_i = b$

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

$$= e^{-(m+\beta)\lambda_1} \lambda_1^{a+\alpha-1} e^{-(n-m+\beta)\lambda_2} \lambda_2^{b+\alpha-1}$$

"Unknown Distribution" that's the best we can do...

Need conditionals: $P(\lambda_1 | X_1, \cdots, X_n, \lambda_2, m) \propto e^{-(m+\beta)\lambda_1} \lambda_1^{a+\alpha-1} \propto \text{Gamma}(a+\alpha, m+\beta)$

$P(\lambda_2 | X_1, \cdots, X_n, \lambda_1, m) \propto e^{-(n-m+\beta)\lambda_2} \lambda_2^{b+\alpha-1} \propto \text{Gamma}(b+\alpha, n-m+\beta)$

$P(m | X_1 \cdots X_n, \lambda_1, \lambda_2) \propto e^{-(m+\beta)\lambda_1} e^{-(n-m+\beta)\lambda_2} \propto e^{-m\lambda_1} e^{\lambda_2 m}$

$\propto \underbrace{e^{-m(\lambda_1-\lambda_2)} \lambda_1^a \lambda_2^b}$ ← dependent on m.

$\propto$ call $h(m)$

$\propto \frac{h(m)}{\sum_{k=0}^{n} h(k)}$

then. Pick λ_1 and a starting point. Plug in to get next round. and repeat. ✓

$< \begin{bmatrix} \lambda_{0,1} \\ \lambda_{0,2} \\ m_0 \end{bmatrix}, \begin{bmatrix} \lambda_{1,1} \\ \lambda_{1,2} \\ m_1 \end{bmatrix}, \cdots >$

Drawing vertical line through the 3 constitutes 1 data point.

All have same burn-in.

converges quickly.

Discard the data points before burn-in.

- Problem: When do you discard data points. ≈ 4.5.

(thining) When does it dip below significance level.

- Then Posterior ↦ get C. Regions.

Color on Github.

↳ Dark line = mean. ($\hat{\theta}_{MMSE}$)

the real mean $\neq \hat{\theta}_{MMSE}$ but still in the Credible Region.

Machine Learning Club ✓