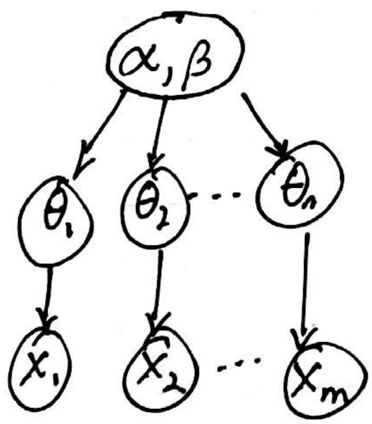
Feynman Liang
STAT 210A
Lecture #11
9/29/2016

1 Review: Hierarchical Bayes

$$\theta_i \sim \text{Beta}(\alpha, \beta) \tag{1.1}$$
$$X_i \mid \theta_i \sim \text{Binom}(n_i, \theta_i) \tag{1.2}$$

$$\hat{\theta}_i = \frac{X_i}{n_i}\left(\frac{n_1}{n_1+\alpha+\beta}\right) + \frac{\alpha}{\alpha+\beta}\left(\frac{\alpha+\beta}{n_1+\alpha+\beta}\right) \tag{1.3}$$



2 MCMC / Gibbs Sampler

MC transition kernel $Q(x \mid y)$.
$\pi(x)$ stationary for Q if

$$\pi(x) = \int_{\mathcal{X}} Q(x|y)\pi(y)dy \tag{2.1}$$

2.1 Gibbs Sampler

Definition 2.1. A transition kernel Q satisfies *detailed balance* if

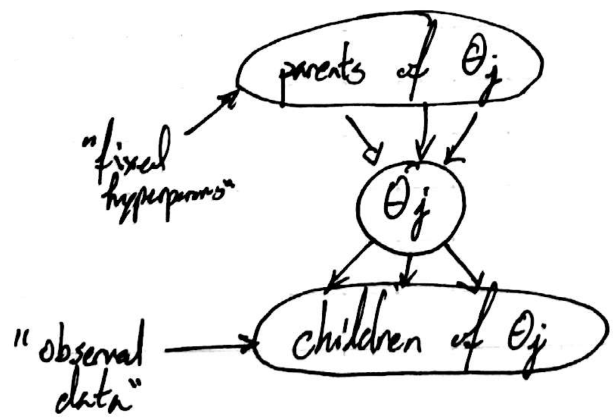
$$\forall x, y : \pi(x)Q(y \mid x) = \pi(y)Q(x \mid y) \tag{2.2}$$

Detailed balance implies the existance of a stationary distribution.
Parameter vector $(\theta_i)_{i=1}^d$, data X.

2.1.1 Gibbs rule

for $j = 1$ to d **do**
  Sample $\theta_1^{(t+1)} \sim \lambda(\theta_1|\theta_{2:d}^{(t)}, X)$
  Sample $\theta_2^{(t+1)} \sim \lambda(\theta_2|\theta_{1:1}^{(t+1)}, \theta_{3:d}^{(t)}, \cdots, \theta_d^{(t)}, X)$
  $\vdots$
  Sample $\theta_d^{(t+1)} \sim \lambda(\theta_d|\theta_{1:d-1}^{(t+1)}, X)$
end for

Gibbs sampling is efficient and can be done in parallel:



Burn-in is an issue:

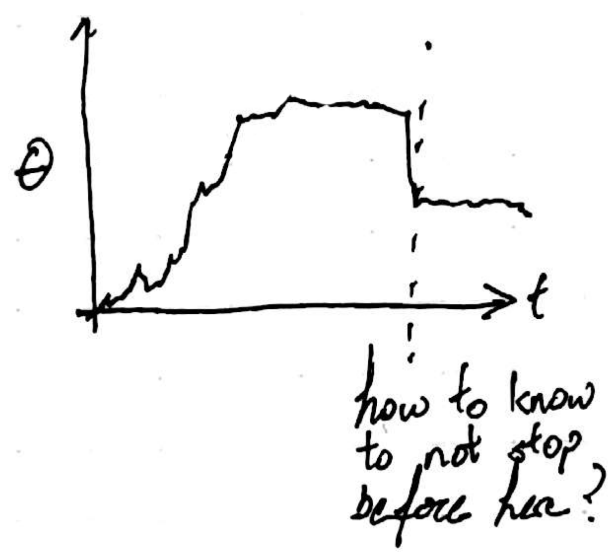


Figure 1: Trace plot of θ vs t showing that MCMC may be terminated before burn in i.e. sufficient t for initial conditions to be negligible and θ to converge

The posterior distribution $\lambda(\cdot \mid X)$ is invariant under this update rule:

$$\theta = (\zeta \in \mathbb{R}, \gamma \in \mathbb{R}^{d-1}) \tag{2.3}$$

$$\lambda(\zeta, \gamma \mid X) Q(\hat{\zeta}, \gamma, \zeta, \gamma) = \lambda(\zeta, \gamma \mid X) \lambda(\hat{\zeta} \mid \gamma, X) \tag{2.4}$$

$$= \lambda(\zeta, \gamma \mid X) \frac{\lambda(\hat{\zeta}, \gamma \mid X)}{\int \lambda(u, \gamma \mid X) du} \tag{2.5}$$

$$\lambda(\hat{\zeta}, \gamma \mid X) Q(\zeta, \gamma \mid \hat{\zeta}, \gamma) = \text{same thing} \tag{2.6}$$

Example 2.2 (Normal Means Model)**.** $X_i \overset{\text{indep}}{\sim} \mathcal{N}(\theta_i, 1)$ for $i = 1, \cdots, n$, $\theta \in \mathbb{R}^n$

Multivariate squared-error loss (Notation $\|\cdot\| = \|\cdot\|_2$)

$$L(\theta, \delta(X)) = \|\theta - \delta(X)\|_2^2 \tag{2.7}$$

$$= \sum_{i=1}^{n} (\theta_i - \delta_i(X))^2 \tag{2.8}$$

$\delta_i(X)$ could depend on X_j for $j \neq i$.

Bayes approach

$$X_i \mid \theta_i \sim \mathcal{N}(\theta_i, 1) \tag{2.9}$$

$$\theta_i \sim \mathcal{N}(0, \tau^2) \tag{2.10}$$

The Bayes estimator is

$$\lambda(\theta \mid X) \propto_\theta \exp\left\{\frac{-1}{2}\|X - \theta\|_2^2 - \frac{1}{2\tau^2}\|\theta\|^2\right\} \tag{2.11}$$

$$= \exp\left\{\frac{-1}{2}\left(1 + \frac{1}{\tau^2}\right)\|\theta\|^2 + X'\theta - \frac{1}{2}\|X\|^2\right\} \tag{2.12}$$

$$\propto_\theta \exp\left\{\frac{-1}{2}\left(1 + \frac{1}{\tau^2}\right)\left\|\theta - \frac{X}{1 + \frac{1}{\tau^2}}\right\|^2\right\} \tag{2.13}$$

$$\theta | X \sim \mathcal{N}\left(\left(1 + \frac{1}{\tau^2}\right)^{-1} X, \left(1 + \frac{1}{\tau^2}\right)^{-1} \vec{I}_n\right) \tag{2.14}$$

$$= \mathcal{N}\left(\left(1 - \frac{1}{1 + \tau^2}\right) X, \left(1 - \frac{1}{1 + \tau^2}\right) \vec{I}_n\right) \tag{2.15}$$

$$\delta_i(X) = \left(1 - \frac{1}{1 + \tau^2})\right) X_i = (1 - \zeta) X_i \tag{2.16}$$

This is an example of a *shrinkage estimator* where we are shrinking towards our prior mean (= 0 in this case) by $(1 - \zeta)$.

Hierarchical-Bayes approach

$$\tau^2 \sim \Lambda(\tau^2) \tag{2.17}$$

$$\theta_i \mid \tau^2 \sim \mathcal{N}(0, \tau^2) \tag{2.18}$$

$$X_i \mid \theta_i \sim \mathcal{N}(\theta_i, 1) \tag{2.19}$$

The estimator is

$$\begin{aligned}
\delta_i(X) &= \mathbb{E}[\theta_i \mid X] && (2.20)\\
&= \mathbb{E}[\mathbb{E}[\theta_i \mid \tau^2, X] \mid X] && (2.21)\\
&= \mathbb{E}[\left(1 - \frac{1}{1+\tau^2}\right) X_i \mid X] && (2.22)\\
&= \left(1 - \mathbb{E}\left[\frac{1}{1+\tau^2} \mid X\right]\right) X && (2.23)
\end{aligned}$$