Feynman Liang
STAT 210A
Lecture #10
9/27/2016

1 Review

Likelihood $p_\theta(x) = p(x|\theta)$

Prior $\lambda(\theta)$

Marginal $q(x) = \int_\Omega p_\theta(x)\lambda(\theta)d\theta$

Posterior $\lambda(\theta|x) = \frac{p_\theta(x)\lambda(\theta)}{q(x)}$

2 Bayes Pros/Cons

Example 2.1. $(X_i)_1^n \overset{\text{iid}}{\sim} p(x)$ non-parametrci model.
Estimate: $\mathbb{E}X, \text{Median}(X), \text{Var}(X)$.
Clear how to estimate with frequentist methods, but choice of prior for p has strong influence on Bayesian inference's results.

Sometimes Jeffrey's priors or uninformative priors can go wrong:

Example 2.2. $p(x) = e^{\beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots - A(\beta_1, \beta_2, \cdots)}$

Example 2.3. $X_i \sim N(\theta_i, 1)$ indep. $i = 1, \cdots, d$.
$(\theta_1, \cdots, \theta_1) \sim$ flat prior
$\Theta \mid X \sim \mathcal{N}_d(X, \vec{I}_d)$, so posterior mean $= X$
Estimate $\|\theta\|_2^2 = \sum_{i=1}^d \theta_i^2$

$$\begin{aligned}
\mathbb{E}[\sum_i \theta_i^2 | X] &= \sum_{i=1}^d \mathbb{E}[\theta_i^2 | X] && (2.1)\\
&= \sum_{i=1}^d (1 + X_i^2) && (2.2)\\
&= \|X\|_2^2 + d && (2.3)\\
MSE &= 2d + (2d)^2 && (2.4)
\end{aligned}$$

$$\mathbb{E}_\theta[\|X\|_2^2] = \sum_i \mathbb{E}_{\theta_i}[X_i^2] \tag{2.5}$$
$$= \sum_{i=1}^d (1+\theta_i^2) \tag{2.6}$$
$$= \|\theta\|_2^2 + d \tag{2.7}$$
$$\implies \|X\|_2^2 - d \text{ is UMVU} \tag{2.8}$$
$$MSE = 2d \tag{2.9}$$

The Bayes estimator has higher MSE: the two estimators have same variance but second is unbiased

3 Hierarchical Bayes

$$X|\theta \sim \text{Binom}(n,\theta) \tag{3.1}$$
$$\Theta \sim \text{Beta}(\alpha,\beta) \tag{3.2}$$

Usually α, β are hyperparameters known in advance, but we can introduce priors on them as well:

$$\Gamma \sim \phi(\gamma) \tag{3.3}$$
$$\Theta|\Gamma = \gamma \sim \lambda_\gamma(\theta) \tag{3.4}$$
$$X|\Theta = \theta \sim p_\theta(x) \tag{3.5}$$

Formally, hierarchical prior no more general (can always marginalize):

$$\theta \sim \tilde{\lambda}(\theta) = \int \lambda_\gamma(\theta)\phi(\gamma)d\gamma \tag{3.6}$$
$$\gamma|\Theta = \theta \sim p_\theta(x) \tag{3.7}$$

Same posterior distribution for $\Theta|X$.

However, hierarchical models are useful for incorporating complex structured prior knowledge into the problem.

Example 3.1. Predict a hitter's "true" batting averge fron n at-bats.

$$X)i = \text{\# hits} \tag{3.8}$$
$$\sim \text{Binom}(n_i,\theta_i) \tag{3.9}$$
$$\Theta_i \sim \text{Beta}(\alpha,\beta) \tag{3.10}$$

$i = 1, \cdots, m$ different batters. Can we pool information across all batters when coming up with a reasonable prior? Let us set

$$\alpha, \beta \overset{\text{indep}}{\sim} \text{Gamma}(k,\theta) \tag{3.11}$$
$$\theta_i|\alpha,\beta \overset{\text{iid}}{\sim} \text{Beta}(\alpha,\beta) \tag{3.12}$$
$$X_i|\theta_i \sim \text{Binom}(n,\theta_i) \tag{3.13}$$

Example 3.2. Can build DAG for PGM.

$$p(\alpha, \beta, \theta_1, \cdots, \theta_n, x_1, \cdots, x_n) = p(\alpha|\beta)p(\theta_1|\alpha, \beta) \cdots p(\theta_n|\alpha, \beta)p(x_1|\alpha, \beta) \cdots p(x_n|\alpha, \beta) \tag{3.14}$$

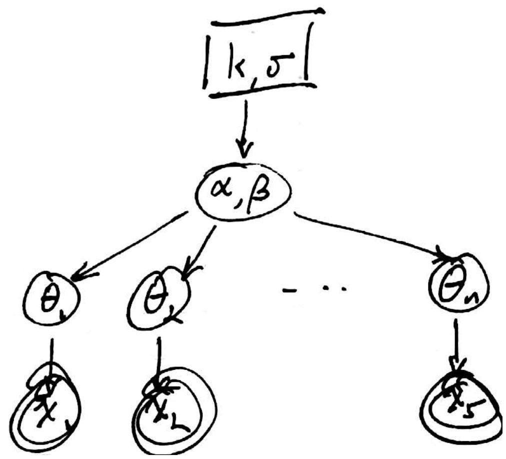

Figure 1: Probabilistc graphical model for the hierarchical Bayes model

4 Markov Chain Monte Carlo

$\lambda(\Theta|X)$ where Θ is some (high dimensional) parameter and X some (high dimensional) data.

The posterior

$$\lambda(\Theta|X) = \frac{p_\theta(x)\lambda(\theta)}{\int_\Omega p_\theta(x)\lambda(\zeta)d\zeta} \tag{4.1}$$

The denominator (aka *partition function*) involves an integration over Ω (which may be high dimensional) and the integrand may be very highly spiked, making symbolic integration intractible and standard grid-based numeric integration computationally difficult.

The typical way to estimate the partition function is MCMC.

4.1 Review: Markov Chains

Definition 4.1. A *(stationary) Markov Chain* with transition kernel $Q(y|x)$ and initial distribution $\pi_0(x)$ is a sequence of RVs

$$X^{(0} \to X^{(1)} \to X^{(2)} \to \cdots \tag{4.2}$$

where $X^{(0)} \sim \pi_0$ and

$$X^{(i+1)}|X^{(0)}, \cdots, X^{(i)} \sim Q(\cdot|X^{(i)}) \tag{4.3}$$

Definition 4.2. If $\pi(y) = \int_{\mathcal{X}} Q(y|x)\pi(x)dx$, then π is *stationary* for Q.

Under mild conditions:

$$P(X^{(t)}) \approx \pi(x) \quad \text{as } t \to \infty \tag{4.4}$$

regardless of $\pi_0(x)$.

$$X^{(1)} \sim \int Q(y|x)\pi_0(x)dx = \pi_1(x) \tag{4.5}$$

Strategy: find some $Q(\cdot \mid \cdot)$ for which the posterior $\lambda(\Theta|X)$ is stationary.

(1) Initialize arbitrarily $\Theta^{(0)}$

(2) Run chain for B steps, B large $\to \Theta^{(B)}$, a sample from posterior

(3) Repeat to get multiple samples.

(4) Can perform MC integration to approximate integral