Outline

9/29/2021

1) Hierarchical Bayes
2) Markov Chain Monte Carlo
3) Gibbs Samples

Hierarchical Bayes

[Full power of Bayes is realized in large, complex problems with repeat structure, allowing us to pool information across many observations.]

Ex Predict a batter's "true" batting average from n at-bats. $X = \#$ of hits $\sim \text{Binom}(n, \theta)$

Prior info: Most batting avg.s are between 0.1 and 0.3, $\theta = 0.8$ very unlikely. Can represent using a Beta dist., but how to pick α, β?

Solution: Pool info across players $i = 1, \ldots, m$ via hierarchical model

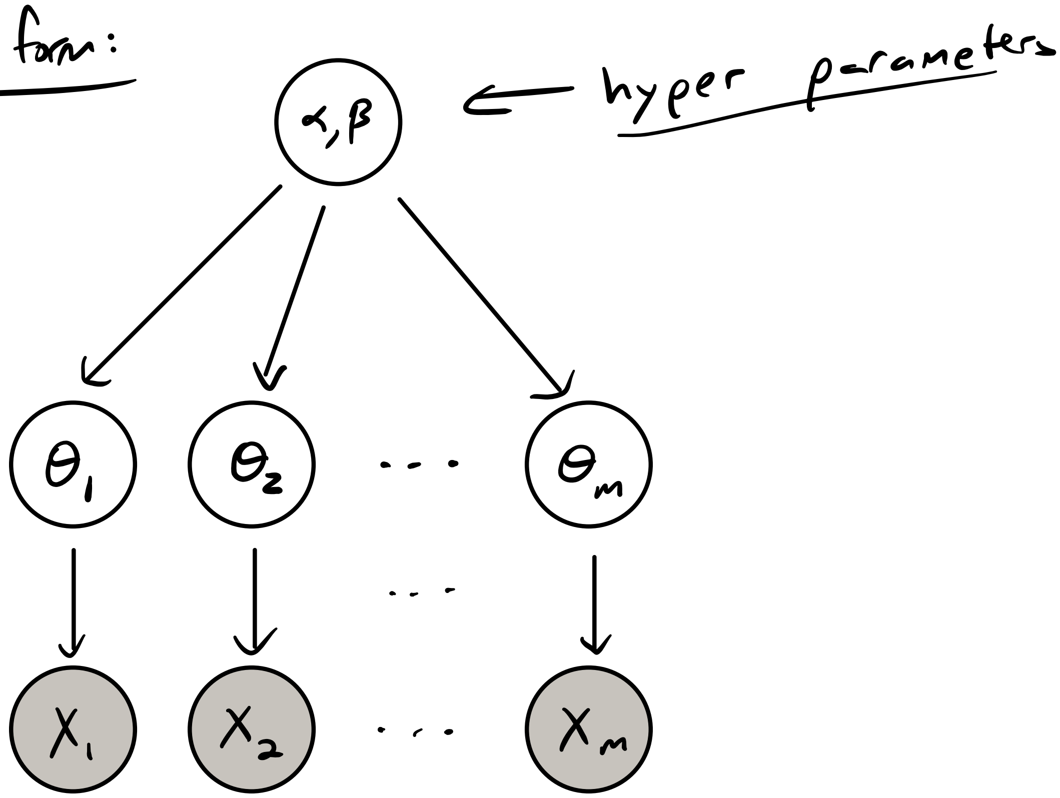
$$\alpha, \beta \sim \lambda_{\alpha,\beta} \qquad (\text{say, indep. } \text{Exp}(1))$$

$$\theta_i \mid \alpha, \beta \overset{iid}{\sim} \text{Beta}(\alpha, \beta) \qquad i \leq m$$

$$X_i \mid \theta_i \overset{indep}{\sim} \text{Binom}(n_i, \theta_i) \qquad i \leq m$$

[Note: there is always an equivalent model where we marginalize over α, β and just write a more complicated prior on Θ. Hierarchical version often gives better intuition or suggests computational strategies]

Graphical form:



This is a directed graphical model. Implies the distribution may be factorized with one factor for each vertex in a DAG (V, E)

$$p(z_1, \dots, z_{|V|}) = \prod_{i=1}^{|V|} p_i(z_i \mid P_a(z_i))$$

parents

For this model,

$$p(\alpha, \beta, \theta_1, \dots, \theta_m, X_1, \dots, X_m)$$
$$= p(\alpha, \beta) \cdot \prod_i p(\theta_i \mid \alpha, \beta) \cdot \prod_i p(x_i \mid \theta_i)$$

Practical implication:

$X_2, \dots, X_m$ indirectly influence the estimate of X_1, by teaching us what values of θ are plausible.

Markov Chain Monte Carlo

Hierarchical models can get very complex very fast, creating big computational headaches

$$\lambda(\theta / x) = \frac{p_\theta(x)\,\lambda(\theta) \quad \longleftarrow \text{usually nice}}{\int_\Omega p_\xi(x)\,\lambda(\xi)\,d\xi \quad \longleftarrow \text{often intractable.}}$$

Computational strategy: set up a Markov chain with stationary dist $\propto p_\theta(x)\lambda(\theta)$, run it to get approximate samples from $\lambda(\theta | x)$

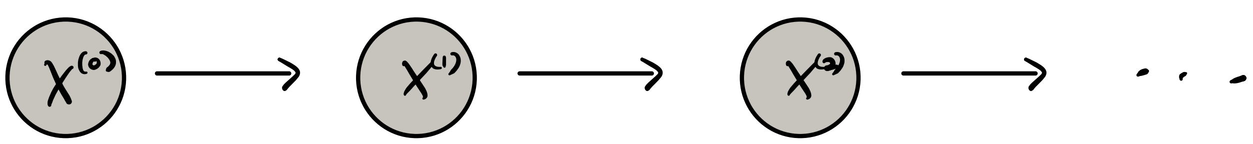
Definition: A (stationary) Markov chain with trans. kernel $Q(y|x)$ and initial dist. $\pi_0(x)$ is a sequence of r.v.s $X^{(0)}, X^{(1)}, \ldots$ where $X^{(0)} \sim \pi_0$ and $X^{(t+1)} \mid X^{(0)}, \ldots, X^{(t)} \sim Q(\cdot \mid X^{(t)})$

$$Q(y|x) = \mathbb{P}(X^{(t+1)} = y \mid X^{(t)} = x)$$

Marginal dist. of $X^{(1)}$:

$$\pi_1(y) = \mathbb{P}(X^{(1)} = y) = \int_{\mathcal{X}} Q(y|x)\,\pi_0(x)\,d\mu(x)$$

This is a directed graphical model:



If $\pi(y) = \int_{\mathcal{X}} Q(y|x)\,\pi(x)\,d\mu(x)$ we say π is a stationary distribution for Q

Sufficient condition is detailed balance:

$$\pi(x)\,Q(y|x) = \pi(y)\,Q(x|y) \qquad \forall\, x,y$$

$$\Rightarrow \int_{\mathcal{X}} Q(y|x)\,\pi(x)\,d\mu(x) = \pi(y)\int_{\mathcal{X}} Q(x|y)\,d\mu(x) = \pi(y)$$

A Markov chain with detailed balance is called reversible: $(X^{(0)},\dots,X^{(t)}) \overset{D}{=} (X^{(t)},\dots,X^{(0)})$ if $\pi_0 = \pi$

$$\mathbb{P}(X^{(t)} = x \mid X^{(t+1)} = y) = \frac{\mathbb{P}(X^{(t)} = x)\,\mathbb{P}(X^{(t+1)} = y \mid X^{(t)} = x)}{\mathbb{P}(X^{(t+1)} = y)} = \frac{\pi(x)\,Q(y|x)}{\pi(y)}$$

Theorem: If an MC with stationary dist. π is:

1) Irreducible: $\forall x,y\ \exists n:\ p(X^{(n)} \overset{\in A \text{ for cts } \mathcal{X}}{=} y \mid X^{(0)} = x) > 0$

2) Aperiodic: $\forall x,\ \gcd\{n > 0 : p(X^{(n)} = x \mid X^{(0)} = x) > 0\} = 1$

can be generalized to cts $\mathcal{X}$

Then $\mathcal{L}(X^{(t)}) \xrightarrow{t\to\infty} \pi$ (in TV distance),

regardless of π_0 (chain "forgets" π_0)

[Proof beyond scope of our class]

Strategy: Find Q with stationary dist $\lambda(\theta|X)$, start at any X, run chain for a long time

$\rightsquigarrow$ $X^{(t)} \approx$ sample from posterior, for large t.

Gibbs sampler

Parameter vector $\theta = (\theta_1, \dots, \theta_d)$

Algorithm:

```
Initialize $\theta = \theta^{(0)}$
For $t = 1, \dots, T$:
    For $j = 1, \dots, d$:                                   $\}$ (*)
        Sample $\theta_j \sim \lambda(\theta_j \mid \theta_{\sim j}, X)$
    Record $\theta^{(t)} = \theta$
```

Variations on (*):

- Update one random coordinate $J^{(t)} \sim \text{Unif}\{0, \dots, d\}$
- Update coordinates in random order

Advantage for hierarchical priors: only need to sample low-dimensional conditional dists:

$$\lambda(\theta_j \mid \theta_{\sim j}, X) \propto p(\theta_j \mid \theta_{Pa(j)}) \cdot \prod_{i : j \in Pa(i)} p(\theta_i \mid \theta_{Pa(i)})$$

Especially easy if using conjugate priors at all levels, often can be parallelized.

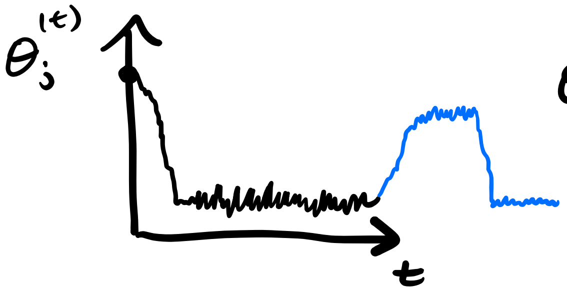
MCMC in Practice

In theory: Pick any initialization $\theta^{(0)}$ and valid kernel Q, sample long enough* ⇝ $\theta^{(t)} \approx \lambda(\theta \mid x)$

Do it again N more times ⇝ N samples from $\lambda(\theta \mid x)$

In practice, how do we know we've sampled long enough?

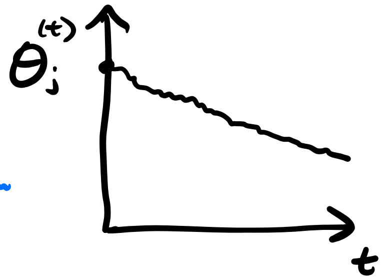
Trace plots: Show how fast the MC mixes



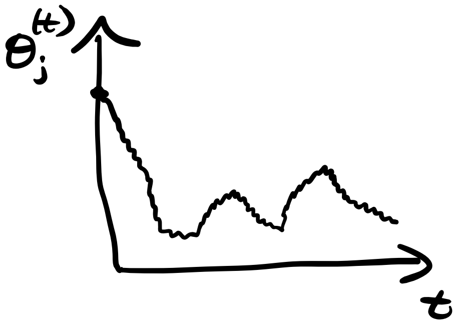
GOOD (?)

Can be deceived!

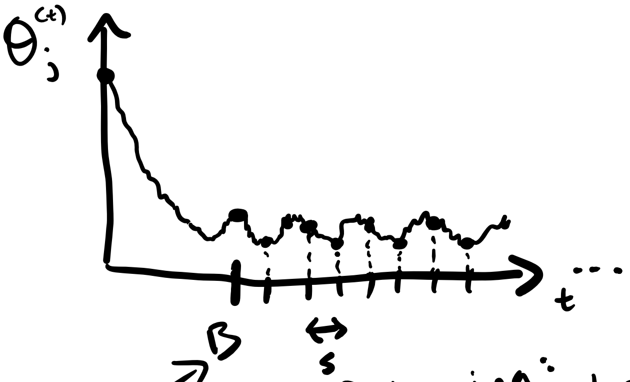
Esp. for bimodal posterior



BAD



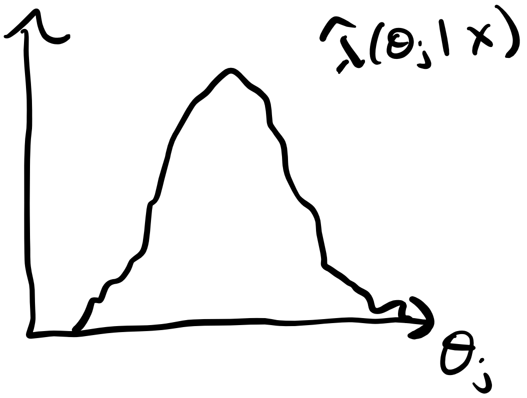
NOT GREAT



Burn-in: "Forget" initialization

thinning: makes samples more independent

⇝



Estimate posterior based on $\{\theta_j^{(B)}, \theta_j^{(B+s)}, \dots, \theta_j^{(B+Ns)}\}$

Posterior mean: $\frac{1}{N+1}\sum_{k=0}^{N}\theta_j^{(B+ks)} \xrightarrow[N\to\infty]{} \mathbb{E}[\theta_j \mid x]$

Implementation details matter!

$$\theta_1, \theta_2 \overset{ind.}{\sim} N(0, 1)$$

$$X_i | \theta \overset{iid}{\sim} N(\theta_1 + \theta_2, 1) \qquad i = 1, \dots, n$$

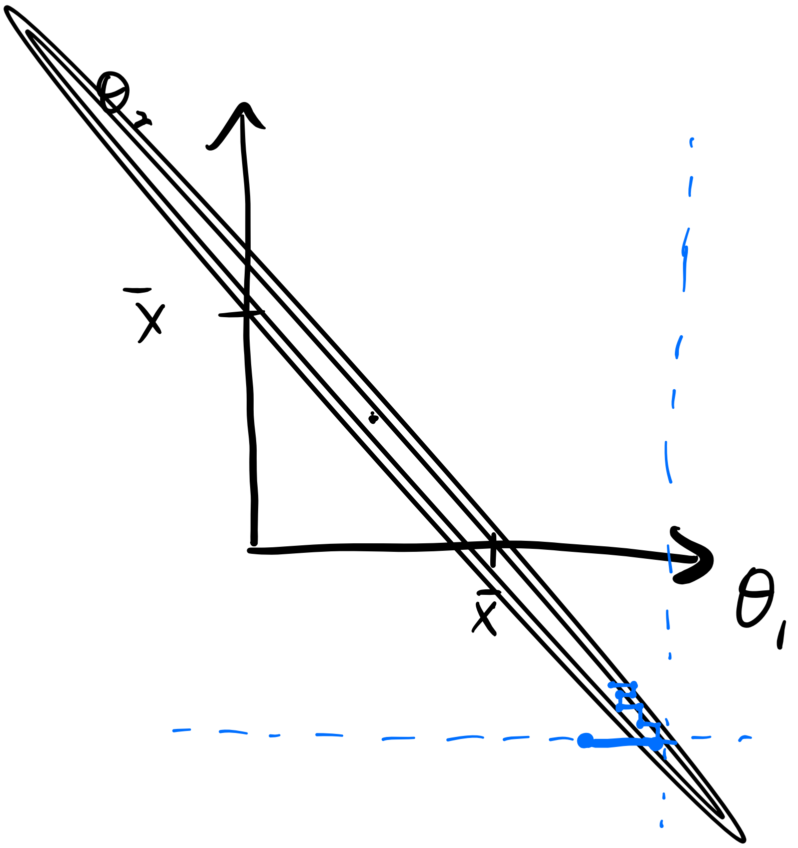
$$\Rightarrow \begin{pmatrix} \theta_1 \\ \theta_2 \\ \bar{X} \end{pmatrix} \sim N_3\left(0, \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 + \frac{1}{n} \end{pmatrix} \right)$$

$$\theta | \bar{X} \sim N_2(\mu(\bar{x}), \Sigma(\bar{x}))$$

$$\mu(\bar{x}) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \left(2 + \tfrac{1}{n}\right)^{-1} \bar{x} = \frac{n\bar{x}}{2n+1} \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\Sigma(\bar{x}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 \\ 1 \end{pmatrix} \left(2 + \tfrac{1}{n}\right)^{-1} (1 \ 1)$$

$$= \frac{n+1}{2n+1} \begin{pmatrix} 1 & \frac{-n}{n+1} \\ \frac{-n}{n+1} & 1 \end{pmatrix}$$



Gibbs takes a long time to mix

Better parameterization:

$$\beta_1 = \theta_1 + \theta_2$$
$$\beta_2 = \theta_1 - \theta_2$$

$$\beta_1 \perp\!\!\!\perp \beta_2 \mid X$$

Gibbs ⇔ Directly sampling from posterior.

Gaussian Hierarchical Model:

$\tau^2 \sim \lambda(\tau)$ e.g. $\frac{1}{\tau^2} \sim \text{Gamma}(k, s)$

$\theta_i \mid \tau^2 \overset{\text{iid}}{\sim} N(0, \tau^2)$ $i \le d$

$X_i \mid \tau^2, \theta \overset{\text{ind.}}{\sim} N(\theta_i, 1)$

Posterior mean:

$$\delta(X_i) = \mathbb{E}[\theta_i \mid X]$$

$$= \mathbb{E}\{\mathbb{E}[\theta_i \mid X, \tau^2\} \mid X]$$

$$= \mathbb{E}\left[\frac{\tau^2}{1+\tau^2} X_i \mid X\right]$$

$$= \underbrace{\left(\mathbb{E}\left[\frac{\tau^2}{1+\tau^2} \mid X\right]\right)}_{\text{Bayes estimate of optimal ``shrinkage'' } \zeta} \cdot X_i$$

Define $\zeta = \frac{1}{1+\tau^2}$ ($\zeta = 0 \Leftrightarrow$ no shrinkage)

$X_i \mid \tau^2 \overset{\text{iid}}{\sim} N(0, 1+\tau^2)$

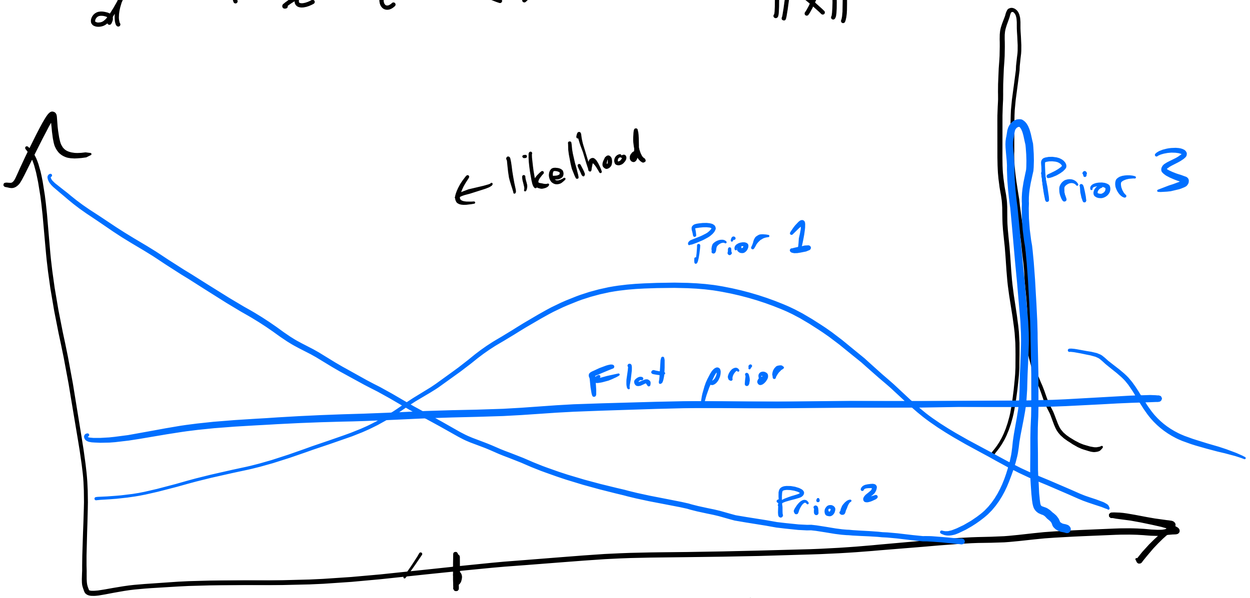
$\frac{\|X\|^2}{d} \mid \tau^2 \sim \frac{1+\tau^2}{d} \chi_d^2$

mean $1+\tau^2$ var $2(1+\tau^2)^2/d$

This likelihood has a sharp peak at

$$\frac{\|X\|^2}{d} - 1 \approx \tau^2 \iff \zeta \approx \frac{d}{\|X\|^2}$$

(large d)



For any reasonably "open-minded" prior (not Prior 3),

$$\hat{\theta}_i \approx \left(1 - \frac{d}{\|X\|^2}\right) X_i \approx (1 - \zeta) X_i$$

If prior doesn't matter much, why use one?

Could just estimate ζ from data however we want, "plug it in"

Called "Empirical Bayes" a hybrid approach in which hyperparameters treated as fixed, others treated as random.