Outline

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_i at-bats. X_i = # of hits $\sim \text{Binom}(n_i, \theta_i)$

Pool info across players $i = 1, \dots, m$ via hierarchical model

$$\alpha, \beta \sim \lambda_0(\alpha, \beta)$$

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

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

$$\mathbb{E}[\theta_i \mid X] = \mathbb{E}\big[\mathbb{E}[\theta_i \mid X, \alpha, \beta] \mid X\big]$$

$$= \mathbb{E}\left[\frac{\overset{\text{fixed}}{X_i} + \overset{\text{sampled} \sim \lambda(\alpha,\beta \mid X)}{\alpha}}{n_i + \alpha + \beta} \,\middle|\, X\right]$$

Intuition: Use all $X_1, \dots, X_m$ to learn good prior on Θ_i

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

Gaussian Hierarchical Model:

$$\tau^2 \sim \lambda_0$$

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

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

Posterior mean:

$$\begin{aligned} \delta(X_i) &= \mathbb{E}[\theta_i \mid X] \\ &= \mathbb{E}\{\mathbb{E}[\theta_i \mid X, \tau^2] \mid X] \\ &= \mathbb{E}\left[\frac{\tau^2}{1+\tau^2} X_i \mid X\right] \\ &= \mathbb{E}\left[\frac{\tau^2}{1+\tau^2} \mid X\right] \cdot X_i \end{aligned}$$

Linear shrinkage estimator,

Bayes-optimal shrinkage estimated from data

Likelihood for τ^2 : marginalize over θ_i

$$X_i \mid \tau^2 \sim N(0, 1+\tau^2)$$

$$\Rightarrow \frac{1}{d}\|X\|^2 \sim \frac{1+\tau^2}{d}\chi^2_d$$

$$\sim \left(1+\tau^2, \frac{2+2\tau^2}{d}\right)$$

notation
(mean, variance)

Define $\zeta(\tau^2) = \frac{1}{1+\tau^2}$ "amount of shrinkage"

$$\Rightarrow \delta(X) = \Big(1 - \underbrace{E[\zeta \mid X]}_{\text{learned from entire data set}}\Big) X_i$$

$$X \mid \zeta \sim N_d\left(0, \tfrac{1}{\zeta} I_d\right) = \frac{1}{(2\pi/\zeta)^{d/2}} e^{-\|X\|^2/(2/\zeta)}$$

$$\propto_\zeta \zeta^{d/2} e^{-\zeta \|X\|^2/2}$$

Conjugate prior:

$$\zeta \sim \frac{1}{s^2}\chi^2_k = \Gamma\left(\tfrac{k}{2}, \overset{\text{Scale}}{2/s^2}\right) = \frac{(s^2)^{k/2}}{\Gamma(k/2)} \zeta^{k/2-1} e^{-s^2\zeta/2}$$

$$\Rightarrow \zeta \mid \|X\|^2 \propto_\zeta \zeta^{\frac{k+d}{2}-1} e^{-(s^2+\|X\|^2)\zeta/2}$$

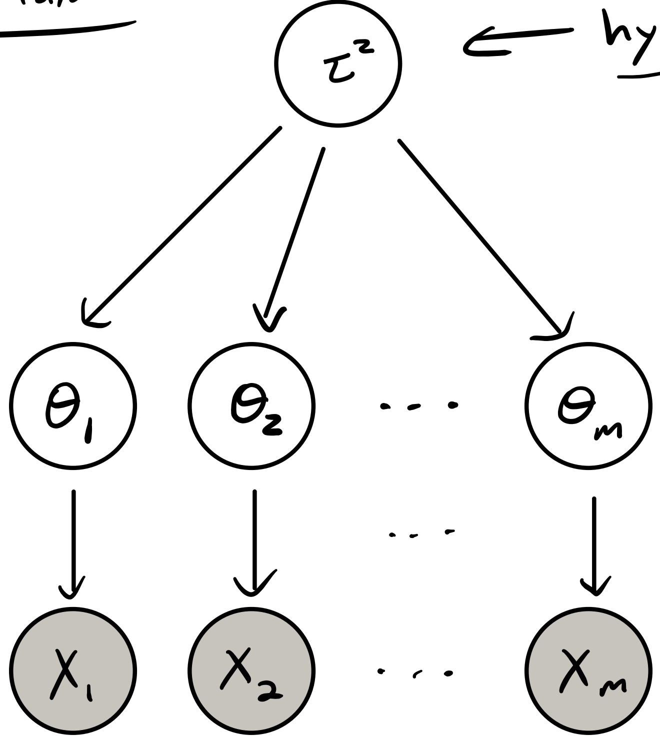
$$\sim \frac{1}{s^2+\|X\|^2} \chi^2_{k+d}$$

$$E[\zeta \mid \|X\|^2] = \frac{k+d}{s^2 + \underbrace{\|X\|^2}} \approx d(1+\tau^2) + O(d^{1/2})$$

"pseudo-data" = $Y_1, \dots, Y_k$ with $\|Y\|^2 = s^2$

[might want to truncate prior to $[0, 1]$ if d small]

Graphical form:



← hyper parameters

These are directed graphical models. Implies the distribution may be factorized with one factor for each vertex in a DAG (V, E)

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

$$Pa(i) = \{ j : j \to i \}$$

For this model,

$$p(\tau^2, \theta_1, \dots, \theta_m, x_1, \dots, x_m)$$

$$= p(\tau^2) \cdot \prod_i p(\theta_i \mid \tau^2) \cdot \prod_i p(x_i \mid \theta_i)$$

Markov Chain Monte Carlo

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

$$\lambda(\theta | x) = \frac{p_\theta(x)\,\lambda(\theta)}{\int_\Omega p_\xi(x)\,\lambda(\xi)\,d\xi} \quad \begin{array}{l} \leftarrow \text{usually nice} \\ \leftarrow \text{often intractable.} \end{array}$$

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

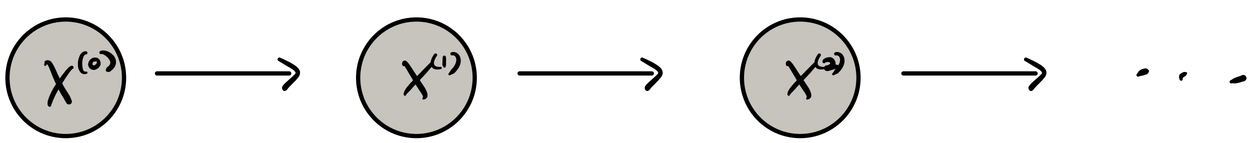
$$X^{(t+1)} \mid X^{(0)}, \ldots, X^{(t)} \sim Q(\cdot \mid X^{(t)})$$

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

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:

$\in A$ for cts $\mathcal{X}$

1) Irreducible: $\forall x,y\ \exists n:\ p(X^{(n)}=y \mid X^{(0)}=x) > 0$
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 $\theta_j \sim \lambda(\theta_j \mid \theta_{\setminus j}, X)$   $\}$ (*)
    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:

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

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

Gibbs: Stationarity of $\lambda(\theta | X)$

Claim: If $\theta^{(t)} \sim \lambda(\theta | X)$ then $\theta^{(t+1)} \sim \lambda(\theta | X)$

Proof:

Consider updating only one (fixed) coordinate j:

$$\eta_{-j} = \theta_{-j}$$

$$\eta_j \sim \lambda(\theta_j | \theta_{-j}, X)$$

If $\theta \sim \lambda(\theta | X) = \lambda(\theta_{-j} | X)\, \lambda(\theta_j | \theta_{-j}, X)$

then $\eta_{-j} = \theta_{-j} \sim \lambda(\theta_{-j} | X)$

$$\eta_j | \eta_{-j} \sim \lambda(\theta_j | \theta_{-j}, X)$$

$$= \lambda(\theta_j | \eta_{-j}, X)$$

$\Rightarrow \eta \sim \lambda(\theta | X)$

$\Rightarrow$ updating <u>any</u> coord preserves posterior dist.

$\Rightarrow$ updating coords in any order also does.

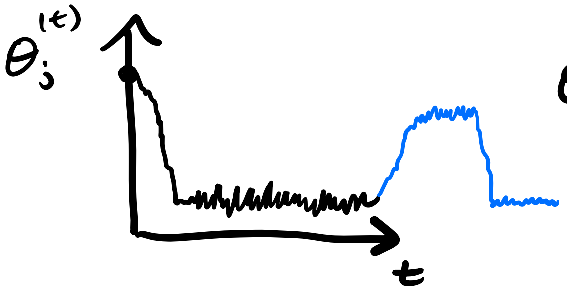
MCMC in Practice

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

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

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

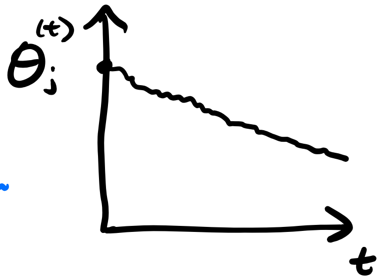
Trace plots: Show how fast the MC mixes



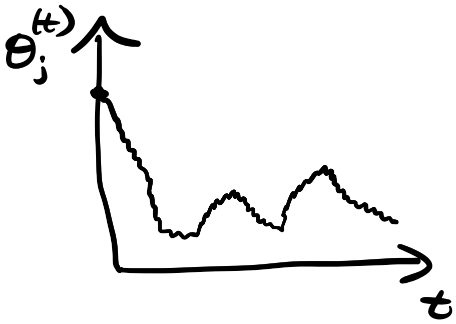
GOOD (?)

Can be deceived!

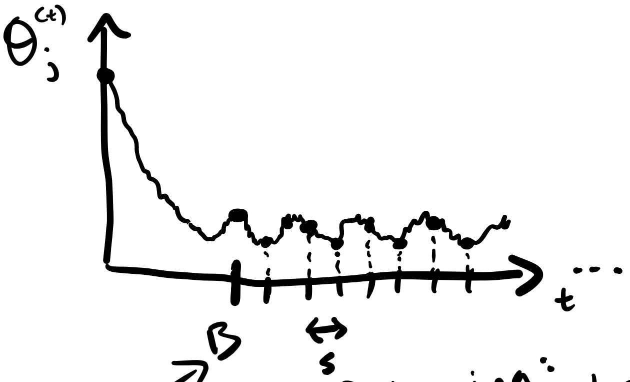
Esp. for bimodal posterior



BAD



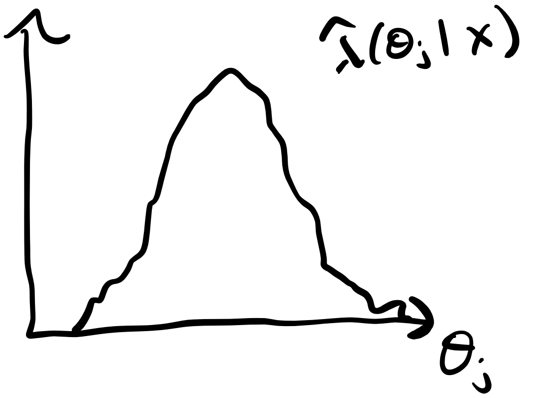
NOT GREAT



Burn-in: "Forget" initialization

thinning: makes samples more independent

$\rightsquigarrow$



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!

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

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

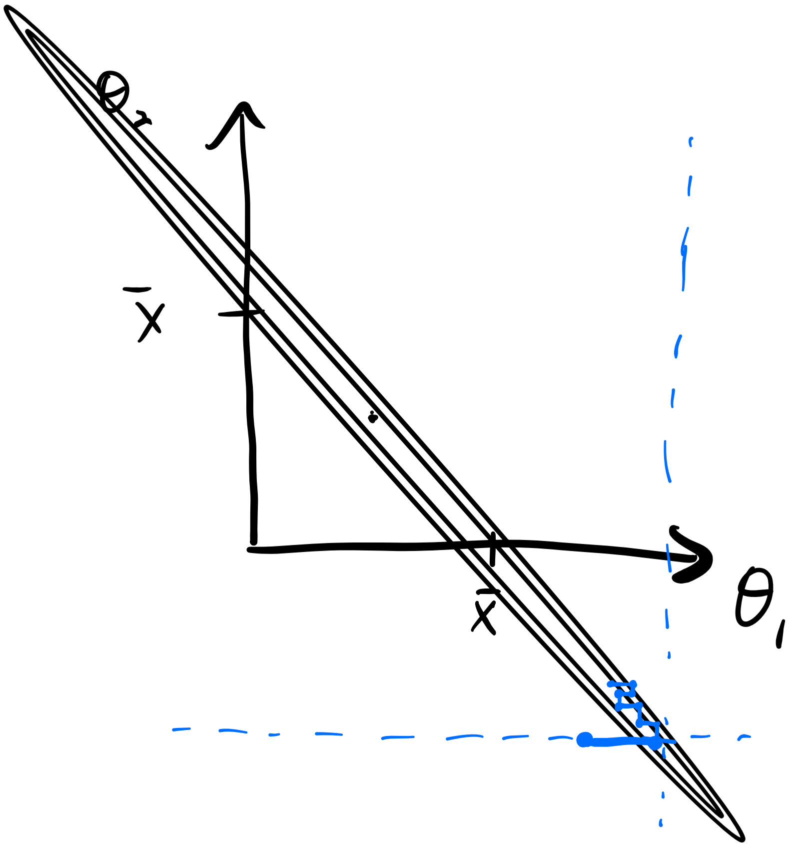
$$\Rightarrow \begin{pmatrix} \theta_1 \\ \theta_2 \\ \bar{X} \end{pmatrix} \sim N_3\left(0, \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2+\frac{1}{n} \end{pmatrix}\right)$$

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

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

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

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



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 $\Leftrightarrow$ Directly sampling from posterior.

Empirical Bayes

Back to Gaussian hierarchical model

$$\frac{1}{d}\|X\|^2 \sim \frac{1+\tau^2}{d}\chi^2_d \qquad \text{MLE for } 1+\tau^2 \text{ is } \frac{1}{d}\|x\|^2$$

$$\sim \left(1+\tau^2, \frac{2+2\tau^2}{d}\right)$$

For any "reasonable" prior, $\mathbb{E}[\zeta \mid X] \approx \overset{(\text{MLE})}{\frac{d}{\|X\|^2}}$

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

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

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

UMVU estimator is $\hat{\zeta} = \frac{d-2}{\|X\|^2}$

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