

Outline

2/28/2023

- 1) Hierarchical Bayes
- 2) Markov Chain Monte Carlo
- 3) Gibbs Sampler
- 4) Empirical Bayes

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

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

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

$$\theta_i | \alpha, \beta \stackrel{\text{iid}}{\sim} \text{Beta}(\alpha, \beta) \quad i \leq m$$

$$X_i | \theta_i \stackrel{\text{indep}}{\sim} \text{Binom}(n_i, \theta_i) \quad i \leq m$$

$$\begin{aligned} \mathbb{E}[\theta_i | X] &= \mathbb{E}[\mathbb{E}[\theta_i | X, \alpha, \beta] | X] \\ &= \mathbb{E}\left[\frac{\overset{\text{fixed}}{\underbrace{X_i}_{\text{fixed}}} + \underbrace{\alpha}_{\text{sampled} \sim \lambda(\alpha, \beta | X)}}{\underbrace{n_i + \alpha + \beta}_{\text{sampled} \sim \lambda(\alpha, \beta | X)}} \mid X\right] \end{aligned}$$

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:

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

$$\theta_i | \tau^2 \stackrel{iid}{\sim} N(0, \tau^2) \quad i \leq d$$

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

Posterior mean:

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

Linear shrinkage estimator,

Bayes-optimal shrinkage estimated from data

Likelihood for τ^2 : marginalize over θ_i

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

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

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

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

$$\Rightarrow \mathcal{J}(X) = \left(1 - \underbrace{\mathbb{E}[\zeta | X]}_{\text{learned from entire data set}}\right) X;$$

learned from entire data set

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

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

Conjugate prior:

$$\zeta \sim \frac{1}{s^2} \chi_k^2 = \Gamma\left(\frac{k}{2}, \frac{2}{s^2}\right) = \frac{(s^2)^{k/2}}{\Gamma(k/2)} \zeta^{k/2-1} e^{-s^2 \zeta/2}$$

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

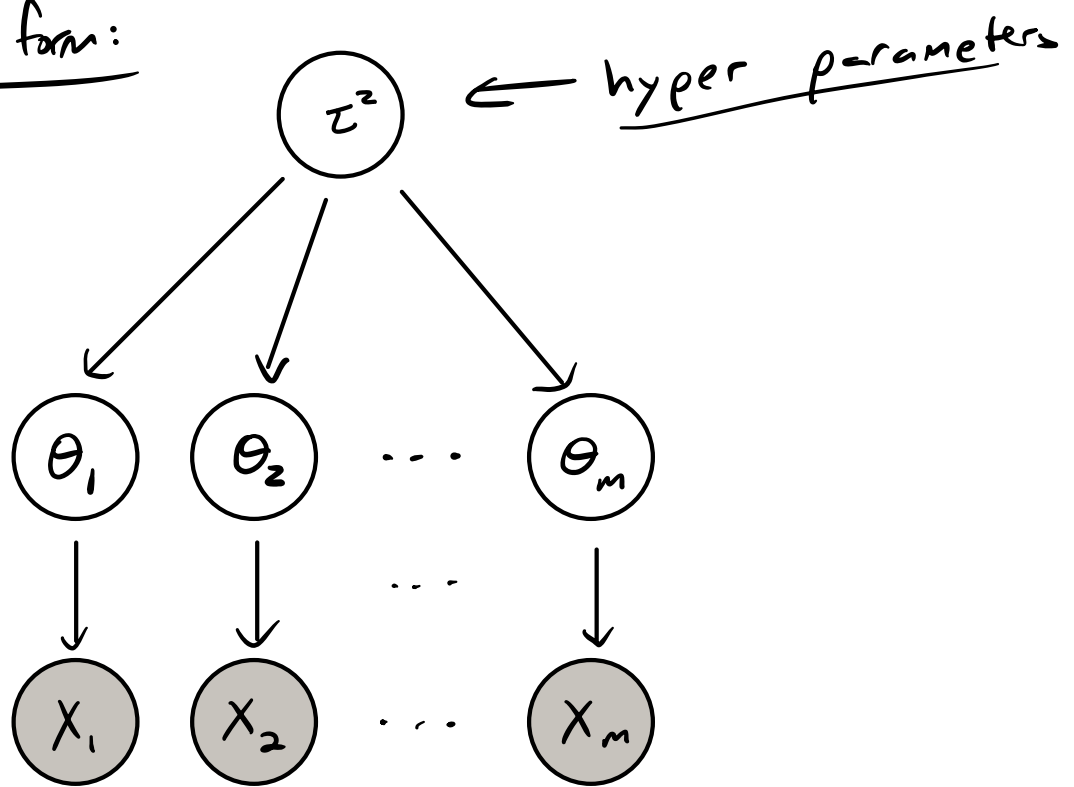
$$\sim \frac{1}{s^2 + \|X\|^2} \chi_{k+d}^2$$

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

"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:



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

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

For this model, $Pa(i) = \{j : j \rightarrow i\}$

$$\begin{aligned} p(z^2, \theta_1, \dots, \theta_m, x_1, \dots, x_m) \\ = p(z^2) \cdot \prod_i p(\theta_i | z^2) \cdot \prod_i p(x_i | \theta_i) \end{aligned}$$

Markov Chain Monte Carlo

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

$$\lambda(\theta|x) = \frac{p_\theta(x) \lambda(\theta)}{\int_{\Omega} p_\theta(x) \lambda(s) ds}$$

← usually nice
← 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 $\lambda(\theta|x)$

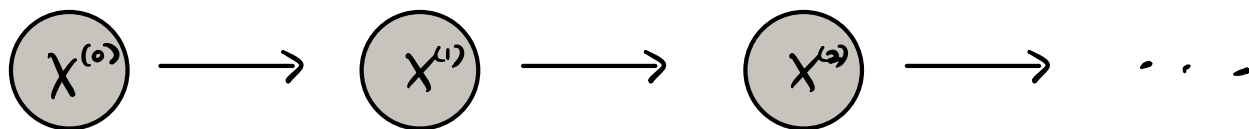
Definition: A (stationary) Markov chain with trans.
kernel $Q(y|x)$ and initial dist. $\pi_0(x)$ is
a sequence of r.v.s $X^{(0)}, X^{(1)}, \dots$ where $X^{(0)} \sim \pi_0$
and $X^{(t+1)} | X^{(0)}, \dots, X^{(t)} \sim Q(\cdot | X^{(t)})$

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

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

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

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:

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

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

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

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

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

- 1) Irreducible: $\forall x, y \exists n: \rho(X^{(n)} = y | X^{(0)} = x) > 0$ (for cts \mathcal{X})
- 2) Aperiodic: $\forall x, \gcd \{n > 0: \rho(X^{(n)} = x | X^{(0)} = x) > 0\} = 1$ can be generalized to cts \mathcal{X}

Then $\mathcal{L}(X^{(t)}) \xrightarrow{t \rightarrow \infty} \pi$ (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
 $\Rightarrow X^{(t)} \approx$ 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 $\theta_j \sim \lambda(\theta_j | \theta_{-j}, x)$ } (*)

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:

$$\lambda(\theta_j | \theta_{-j}, x) \propto p(\theta_j | \theta_{P_2(j)}) \cdot \prod_{i: j \in P_2(i)} p(\theta_i | \theta_{P_2(i)})$$

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 :

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

$$\gamma_j \sim \lambda(\theta_j | \theta_{-j}, x)$$

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

$$\text{then } \gamma_{-j} = \theta_{-j} \sim \lambda(\theta_{-j}|x)$$

$$\gamma_j | \gamma_{-j} \sim \lambda(\theta_j | \theta_{-j}, x)$$

$$= \lambda(\theta_j | \gamma_{-j}, x)$$

$$\Rightarrow \gamma \sim \lambda(\theta|x)$$

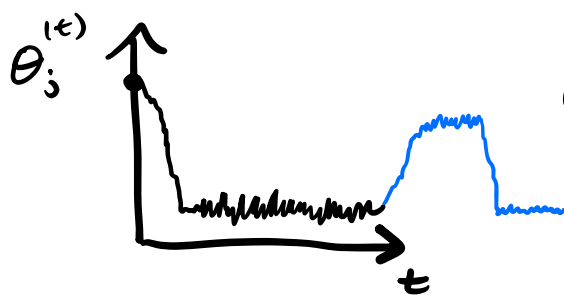
\Rightarrow updating any coord preserves posterior dist.

\Rightarrow updating words in any order also does.

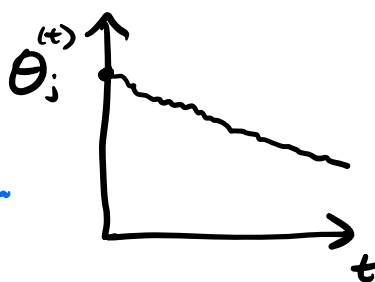
MCMC in Practice

In theory: Pick any initialization $\theta^{(0)}$ and valid kernel Q , sample long enough $\rightarrow \theta^{(t)} \approx \lambda(\theta | x)$
 Do it again N more times $\rightarrow N$ samples from $\lambda(\theta | x)$
 In practice, how do we know we've sampled long enough?

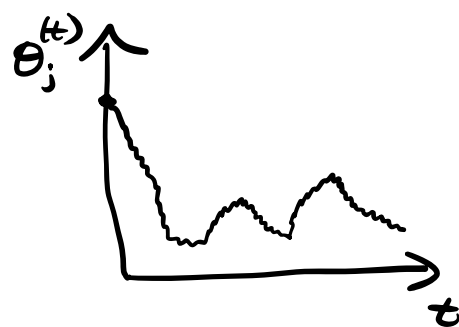
Trace plots: Show how fast the MC mixes



GOOD (?)



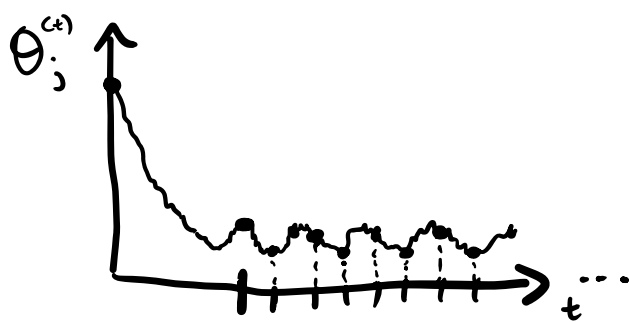
BAD



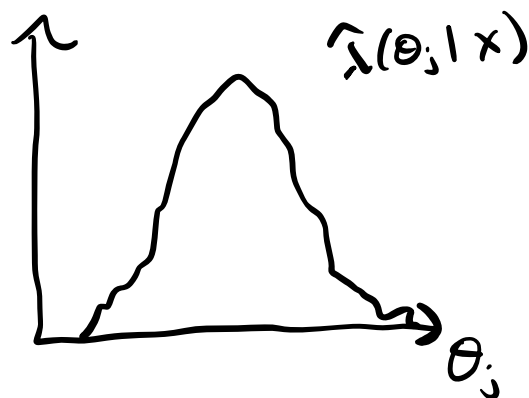
NOT GREAT

Can be deceived!

Esp. for bimodal posterior



\rightsquigarrow



Burn-in: "Forget" initialization
 $\rightarrow B$
 $\leftrightarrow s$
 Thinning: makes samples more independent

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

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

Implementation details matter!

$$\theta_1, \theta_2 \stackrel{\text{iid.}}{\sim} N(0, 1)$$

$$X_i | \theta \stackrel{\text{iid.}}{\sim} N(\theta_1 + \theta_2, 1) \quad i=1, \dots, n$$

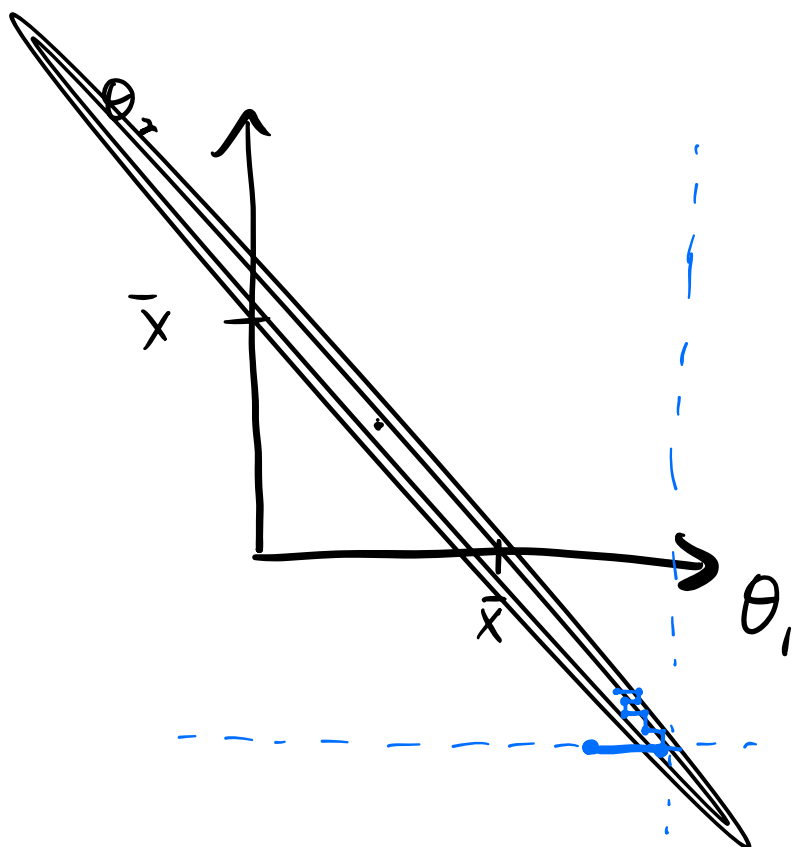
$$\Rightarrow \begin{pmatrix} \theta_1 \\ \theta_2 \\ \bar{x} \end{pmatrix} \sim N_3 \left(0, \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 2 + \frac{1}{n} \end{pmatrix} \right)$$

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

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

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

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



Gibbs takes a long time to mix

Better parameterization:

$$\beta_1 = \theta_1 + \theta_2$$

$$\beta_2 = \theta_1 - \theta_2$$

$$\beta_1 \perp \beta_2 | x$$

Gibbs \Leftrightarrow Directly sampling from posterior.

Empirical Bayes

Back to Gaussian hierarchical model

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

MLE for $1+\tau^2$
is $\frac{1}{d} \|X\|^2$

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

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

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

Could just estimate S from data

however we want, "plug it in"

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

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