

Lecture 6

Outline:

- 1) Bayesian posteriors on Λ CDM parameters
- 2) Metropolis-Hastings Markov chain sampling.
- 3) Affine invariant ensemble MCMC.

Posterior $P(\theta | \text{data}^s)$ on the Λ CDM parameters $\theta = (R_{ch}^2, R_b h^2, \dots, n_s)$

Recall: our idealized data \vec{d} is a vector of data_l , $l=0, 1, \dots, l_{\max}$.
 $-l \leq m \leq l$ where

$$\text{data}_m = T_m + \varepsilon_m \quad \text{s.t.} \quad E(T_m | T_{m'}) = \sum_{\ell} S_{mm'} C_{\ell}^{T, \theta} \quad E(\varepsilon_m | \varepsilon_{m'}) = \sum_{\ell} S_{mm'} \sigma^2 e^{\frac{b^2}{8 \log 2} \ell (\ell+1)}$$

depends on Λ CDM parameters

Given a prior $\pi(\theta)$ on the Λ CDM parameters the posterior $P(\theta | \vec{d})$ has the form

$$P(\theta | \vec{d}) \propto P(\vec{d} | \theta) \pi(\theta)$$

$$\therefore P(\theta | \vec{d}) \propto \pi(\theta) \prod_{l=0}^{l_{\max}} \left(C_l^{T, \theta} + C_l^{\varepsilon \varepsilon} \right)^{-\frac{2m}{2}} \exp \left(-\frac{2\ell+1}{2} \frac{\partial \ell}{(C_l^{T, \theta} + C_l^{\varepsilon \varepsilon})} \right)$$

Sampling $P(\theta | \vec{d})$ with Metropolis-Hastings (MH for short)

MH is one of the easiest & most basic techniques for setting up a Markov chain $\theta_1, \theta_2, \theta_3, \dots$ s.t. in the limit as $N \rightarrow \infty$, $\theta_N \sim P(\theta | \vec{d})$.

* Note: Many of these sampling algorithms are designed to circumvent having to compute the normalization factor in $P(\theta | \vec{d})$.

To run a MH chain you need a proposal density: $P_{\text{prop}}(\theta | \theta_0)$

\curvearrowleft depends on a previously sampled θ_0 .

Algorithm: Basic MH

Input: initial start θ_0 , proposal density $P_{\text{prop}}(\cdot)$

Output: $\theta_1, \theta_2, \theta_3, \dots$

for $i = 1, 2, \dots$

* Set $\theta_{\text{curr}} = \theta_{i-1}$

* Sample $\theta_{\text{prop}} \sim P_{\text{prop}}(\theta | \theta_{\text{curr}})$

* Draw $U \sim \text{Uniform}(0, 1)$ & define

$$\alpha := \min \left(\frac{P(\theta_{\text{prop}} | \vec{d})}{P(\theta_{\text{curr}} | \vec{d})} \frac{P_{\text{prop}}(\theta_{\text{curr}} | \theta_{\text{prop}})}{P_{\text{prop}}(\theta_{\text{prop}} | \theta_{\text{curr}})}, 1 \right)$$

* if $U \leq \alpha$

 set $\theta_i = \theta_{\text{prop}}$

else

 set $\theta_i = \theta_{\text{curr}}$

end

Fact 1:

$$\text{if } p_{\text{prop}}(\theta | \theta_{\text{curr}}) = p_{\text{prop}}(\theta_{\text{curr}} | \theta)$$

then α can be simplified to

$$\alpha := \min \left(\frac{p_{\text{prop}}(\theta_{\text{prop}} | \theta)}{p_{\text{prop}}(\theta_{\text{curr}} | \theta)}, 1 \right)$$

Fact 2:

MH does not require evaluation of the normalizing constant in $p(\theta | d)$

Fact 3:

The hard part about MH is choosing a p_{prop} which gives good acceptance rates. This is difficult when degeneracies are present.

Why does MH work?

Claim: Let x_1, x_2, \dots be a markov chain [i.e. that $p(x_{i+1} | x_1, x_2, \dots, x_i) = p(x_{i+1} | x_i)$] with invariant density f [i.e. $x_i \sim f = x_{i+1} \sim f$]. Then $x_n \xrightarrow{n \rightarrow \infty} f$.

see p. 163 in Robert & Casella's book "Monte Carlo Statistical Methods"

: All we need to show is that MH has the posterior $p(\theta | d)$ has an invariant density.

(3)

(4)

Claim:

Let x_1, x_2, \dots be a MH markov chain for sampling from some density f where the support of $p_{\text{prop}}(x | y)$ contains the support of f . Let $K(x|y)$ denote the transition density. Then

i) $K(x|y)f(y) = K(y|x)f(x)$
(detailed balance ... same thing as reversibility)

which implies

ii) f is an invariant density.

Proof:

Let $\alpha(x|y) = \min \left(\frac{f(x)}{f(y)} \frac{p_{\text{prop}}(y|x)}{p_{\text{prop}}(x|y)}, 1 \right)$.

Fix y .

Let A, B be the events

$$A := \{x_{\text{prop}} \in x + dx\}$$

$$B := \{U \leq \alpha(x_{\text{prop}} | y)\}$$

Notice we are working with densities so we can suppose w.l.o.g that $x \neq y$

$$\therefore K(x|y)dx = P(A \cap B)$$

$$= \int P(A \cap B | X_{\text{prop}}) p_{\text{prop}}(x_{\text{prop}} | y) dx_{\text{prop}}$$

↓
This factors to
 $P(A | x_{\text{prop}}) P(B | x_{\text{prop}})$

$$= I_{(x, x+dx)} \alpha(x_{\text{prop}} | y)$$

$$= \alpha(x|y) p_{\text{prop}}(x|y) dx$$

Now we show detailed balance for K

(5)

$$\begin{aligned} K(x|y) f(y) &= \alpha(x|y) p_{\text{prop}}(x|y) f(y) \\ &= \min \left(f(x) p_{\text{prop}}(y|x), p_{\text{prop}}(x|y) f(y) \right) \\ &= \alpha(y|x) p_{\text{prop}}(y|x) f(x) \\ &= K(y|x) f(x). \end{aligned}$$

This shows i). For ii) just notice

$$\begin{aligned} x_i \sim f(x) &\Rightarrow x_{i+1} \sim \int K(x|y) f(y) dy \\ &= \int K(y|x) f(x) dy \\ &= f(x). \end{aligned}$$

□

Choosing the proposal density p_{prop}

- if you have an idea of the posterior cov Σ and the parameters are not too nonlinear a common choice is

$$p_{\text{prop}}(\theta | \theta_{\text{curr}}) \sim N(\theta_{\text{curr}}, g\Sigma)$$

↑
tune this
to get
a good
acceptance
rate...

- If the data model is $\vec{d} = X(\theta - \theta_0) + \vec{\epsilon}$ try

$$p_{\text{prop}}(\cdot | \theta_{\text{curr}}) \sim N\left(\theta_{\text{curr}}, g(X^T N^{-1} X)^{-1}\right)$$

where $N = E(\vec{\epsilon} \vec{\epsilon}^T)$.

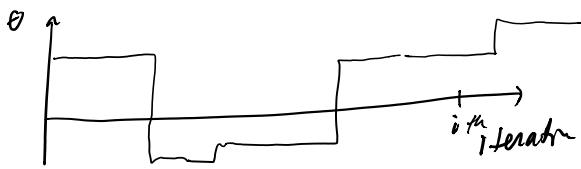
encourages
linear dependence

Note: Non informative priors of the form

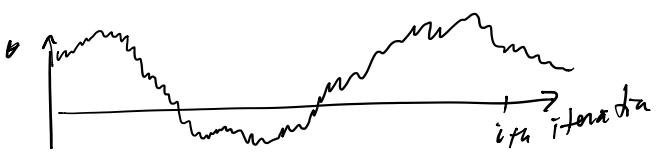
$$\pi(\beta) \sim N(\beta_0, g(X^T N^{-1} X)^{-1})$$

for the regression Model $\vec{d} = X\beta + \epsilon$
are called "g-priors".

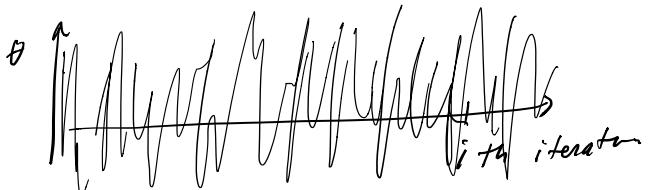
If the acceptance rate is too small the chain will look like



If too large ...



A good rule of thumb is an acceptance of 20%



Burn in, thinning & initial start.

- For LCDM it can help to start at a numerical approximation to $\theta_{\text{max}} = \arg \max_{\theta} P(\theta | \vec{d})$.
- Discard initial steps to avoid dependence on initial starting point
- thin the chain to reduce within chain dependence

Affine-Invariant ensemble MCMC

(7)

Ref: Goodman & Weare (2010).

A MCMC chain whose performance is invariant under linear transformations
... useful for degenerate posteriors.

Algorithm

Input: $a > 0, n \in \mathbb{Z}^+$ and an ensemble of "walkers" $\vec{\theta} := (\theta_1, \dots, \theta_n)$ where each θ_i is in the parameter space (e.g. $a=2$, $n=100$).
Output: $\vec{\theta}_1, \vec{\theta}_2, \vec{\theta}_3, \dots$

for $i = 1, 2, \dots$ ← i indexes iteration

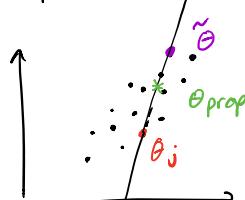
* set $\vec{\theta}_{i+1} = \vec{\theta}_i$

for $j = 1, \dots, n$ ← j indexes walkers

* draw z with density proportional to $\frac{1}{\sqrt{z}} I_{[\frac{1}{a}, a]}(z)$

* Randomly choose a walker $\tilde{\theta}$ from $\{\theta_1, \dots, \theta_n\} \setminus \{\theta_j\}$

* set $\theta_{\text{prop}} = \tilde{\theta} + z(\theta_j - \tilde{\theta})$



* Draw $U \sim \text{Uniform}(0, 1)$ & define

$$\alpha := \min\left(2^{d-1} \frac{P(\theta_{\text{prop}} | \vec{d})}{P(\theta_j | \vec{d})}, 1\right)$$

$d :=$ dim of the parameter space

* if $U \leq \alpha$
| set $(\vec{\theta}_{i+1})_j = \theta_{\text{prop}}$
| end

end

end

Note: The stationary distribution $\vec{\theta}$ is

$$\vec{\theta} \sim \prod_{i=1}^N P(\theta_i | \vec{d}) = \text{iid copies of } p(\vec{\theta} | \vec{d}).$$

(8)