Metropolis Hastings Sampling for Bayesian Inference

Simon Wood, University of Edinburgh, U.K.

Bayesian Inference

- Suppose we have data $\mathbf{y}$ and parameters of a model for the data $\boldsymbol{\theta}$.
- Suppose that we *treat the parameters as random* and describe our beliefs/knowledge about $\boldsymbol{\theta}$, *prior* to observing $\mathbf{y}$, by p.d.f. $\pi(\boldsymbol{\theta})$.
- Denoting densities by $\pi(\cdot)$, recall that from basic conditional probability the joint density of $\mathbf{y}$ and $\boldsymbol{\theta}$ can be written

$$\pi(\mathbf{y}, \boldsymbol{\theta}) = \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|\mathbf{y})\pi(\mathbf{y})$$

- Re-arranging gives Bayes theorem

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})/\pi(\mathbf{y})$$

- The *posterior* density on the left describes our knowledge about $\boldsymbol{\theta}$ after having observed $\mathbf{y}$. Notice that $\pi(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood.

Simulating from the posterior, $\pi(\boldsymbol{\theta}|\mathbf{y})$

- For most interesting models there is no closed form for $\pi(\boldsymbol{\theta}|\mathbf{y})$.
- Even evaluating $\pi(\boldsymbol{\theta}|\mathbf{y})$ is usually impractical as

$$\pi(\mathbf{y}) = \int \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$

  is usually intractable.
- But it turns out that we can simulate samples from $\pi(\boldsymbol{\theta}|\mathbf{y})$, in a way that only requires evaluation of $\pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ (at the observed $\mathbf{y}$), thereby bypassing $\pi(\mathbf{y})$.
- We simulate sequences of random vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \ldots$ so that:
  1. $\pi(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i-2}, \ldots) = P(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1})$ (Markov property).
  2. $\boldsymbol{\theta}_i \sim \pi(\boldsymbol{\theta}|\mathbf{y})$.

  This is known as Markov Chain Monte Carlo (MCMC).

The condition for MCMC to work: Reversibility

- A Markov chain, with transition kernel $P(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1})$, will generate from $\pi(\boldsymbol{\theta}|\mathbf{y})$ if it satisfies the reversibility condition

$$P(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1})\pi(\boldsymbol{\theta}_{i-1}|\mathbf{y}) = P(\boldsymbol{\theta}_{i-1}|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i|\mathbf{y})$$

- Why? If $\boldsymbol{\theta}_{i-1} \sim \pi(\boldsymbol{\theta}|\mathbf{y})$ then LHS is joint density of $\boldsymbol{\theta}_i, \boldsymbol{\theta}_{i-1}$ from the chain. Integrating out $\boldsymbol{\theta}_{i-1}$ we get the marginal for $\boldsymbol{\theta}_i$

$$\int P(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1})\pi(\boldsymbol{\theta}_{i-1}|\mathbf{y})d\boldsymbol{\theta}_{i-1} = \int P(\boldsymbol{\theta}_{i-1}|\boldsymbol{\theta}_i)\pi(\boldsymbol{\theta}_i|\mathbf{y})d\boldsymbol{\theta}_{i-1} = \pi(\boldsymbol{\theta}_i|\mathbf{y})$$

  i.e. $\boldsymbol{\theta}_i \sim \pi(\boldsymbol{\theta}|\mathbf{y})$.
- So given reversibility, if $\boldsymbol{\theta}_1$ is not impossible under $\pi(\boldsymbol{\theta}|\mathbf{y})$ then $\boldsymbol{\theta}_i \sim \pi(\boldsymbol{\theta}|\mathbf{y})$ for $i \geq 1$.

Constructing a reversible $P(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1})$: Metropolis Hastings

- We can construct an appropriate P, based on making a random proposal for $\boldsymbol{\theta}_i$ and then accepting or rejecting the proposal with an appropriately tuned probability.
- Let $q(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1})$ be a *proposal* distribution, chosen for convenience. e.g. $\boldsymbol{\theta}_i \sim N(\boldsymbol{\theta}_{i-1}, \mathbf{I}\sigma_\theta^2)$ for some σ_θ^2.
- Metropolis Hastings iterates the following two steps, starting from some $\boldsymbol{\theta}_0$ and $i = 1\ldots$
  1. Generate a proposal $\boldsymbol{\theta}_i' \sim q(\boldsymbol{\theta}_i|\boldsymbol{\theta}_{i-1})$.
  2. Accept and set $\boldsymbol{\theta}_i = \boldsymbol{\theta}_i'$ with probability

$$\alpha = \min\left\{1, \frac{\pi(\mathbf{y}|\boldsymbol{\theta}_i')\pi(\boldsymbol{\theta}_i')q(\boldsymbol{\theta}_{i-1}|\boldsymbol{\theta}_i')}{\pi(\mathbf{y}|\boldsymbol{\theta}_{i-1})\pi(\boldsymbol{\theta}_{i-1})q(\boldsymbol{\theta}_i'|\boldsymbol{\theta}_{i-1})}\right\}$$

otherwise set $\boldsymbol{\theta}_i = \boldsymbol{\theta}_{i-1}$. Increment i by 1.

Why Metropolis Hastings works in theory

- Let's compress notation writing $\Pi(\boldsymbol{\theta})$ for $\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \pi(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})$ ($\mathbf{y}$ is fixed, after all).
- So the MH acceptance probability for $\boldsymbol{\theta}'$ in place of $\boldsymbol{\theta}$ is

$$\alpha(\boldsymbol{\theta}', \boldsymbol{\theta}) = \min\left\{1, \frac{\Pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\Pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}$$

- $P(\boldsymbol{\theta}'|\boldsymbol{\theta}) = q(\boldsymbol{\theta}'|\boldsymbol{\theta})\alpha(\boldsymbol{\theta}', \boldsymbol{\theta})$ so for $\boldsymbol{\theta} \neq \boldsymbol{\theta}'$

$$\begin{aligned}\Pi(\boldsymbol{\theta})P(\boldsymbol{\theta}'|\boldsymbol{\theta}) &= \Pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})\min\left\{1, \frac{\Pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{\Pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\} \\ &= \min\{\Pi(\boldsymbol{\theta})q(\boldsymbol{\theta}'|\boldsymbol{\theta}), \Pi(\boldsymbol{\theta}')q(\boldsymbol{\theta}|\boldsymbol{\theta}')\} = \Pi(\boldsymbol{\theta}')P(\boldsymbol{\theta}|\boldsymbol{\theta}')\end{aligned}$$

(last equality by symmetry) — reversibility! Trivial if $\boldsymbol{\theta} = \boldsymbol{\theta}'$.

Making Metropolis Hastings work in practice

- The chain output will be correlated. It may take a long time to reach the high probability region of $\pi(\boldsymbol{\theta}|\mathbf{y})$ from a poor $\boldsymbol{\theta}_0$.
- So we usually have to discard some initial portion of the simulation (burn in).
- The proposal distribution will make a big difference to how rapidly the chain explores $\pi(\boldsymbol{\theta}|\mathbf{y})$
  - Large ambitious proposals will result in frequent rejection, and the chain remaining stuck for many steps.
  - Small over cautious proposals will lead to high acceptance, but slow movement as each step is small.
- It is necessary to examine chain output to see how quickly the sampler is exploring $\pi(\boldsymbol{\theta}|\mathbf{y})$ (how well it is *mixing*), and to tune the proposal if necessary.
- Output must also be checked for convergence to the high probability region of $\pi(\boldsymbol{\theta}|\mathbf{y})$.

An example

- Consider the `nhtemp` supplied with base R, giving annual mean temperatures, T_i, in New Haven over several years.
- Suppose we want to model the data using a heavy tailed distribution, and adopt the model

$$(T_i - \mu)/\sigma \underset{\text{i.i.d.}}{\sim} t_\nu$$

  where μ, σ and ν are parameters. If f_ν is the p.d.f. of a t_ν distribution then the p.d.f. for T_i is $f(t) = f_\nu((t-\mu)/\sigma)/\sigma$
- The log likelihood for this model can be coded:

```
ll <- function(theta,temp) {
  mu <- theta[1];sig <- exp(theta[2])
  df = 1 + exp(theta[3])
  sum(dt((temp-mu)/sig,df=df,log=TRUE) - log(sig))
}
```

Priors and Proposals

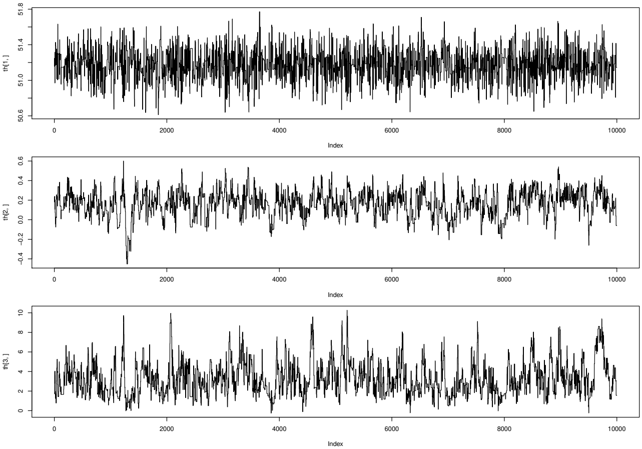
- To complete the model, we need priors for the parameters.
- Let's use improper uniform priors for $\theta_1 = \mu$ and $\theta_2 = \log(\sigma)$.
  - Note: these parts of the prior cancel in the MH acceptance ratio.
- ν becomes somewhat unidentifiable if it is too high, so for convenience, let's assume a prior $\log \nu = \theta_3 \sim N(3, 2^2)$.
- Now the Bayesian model is complete, we need to pick a proposal distribution to form the basis for an MH sampler.
- Let's use the *random walk* proposal $\boldsymbol{\theta}'_i \sim N(\boldsymbol{\theta}_{i-1}, \mathbf{D})$, where $\mathbf{D}$ is diagonal, and we will need to tune its elements.
  - Note that for this proposal $q(\boldsymbol{\theta}'_i|\boldsymbol{\theta}_{i-1}) = q(\boldsymbol{\theta}_{i-1}|\boldsymbol{\theta}'_i)$, so q cancels in MH acceptance ratio.
- Remember that the proposal does not change the posterior, but will affect how quickly the chain explores the posterior.

MH sampler code

```
ns <- 10000; th <- matrix(0,3,ns)
th[,1] <- c(mean(nhtemp),log(sd(nhtemp)),log(6))
llth <- ll(th[,1],nhtemp) ## initial log likelihood
lprior.th <- dnorm(th[3,1],mean=3,sd=2,log=TRUE)
p.sd <- c(.5,.1,1.2) ## proposal SD (tuned)
accept <- 0 ## acceptance counter
for (i in 2:ns) { ## MH sampler loop
  thp <- th[,i-1] + rnorm(3)*p.sd ## proposal
  lprior.p <- dnorm(thp[3],mean=3,sd=2,log=TRUE)
  llp <- ll(thp,nhtemp) ## log lik of proposal
  if (runif(1) < exp(llp + lprior.p - llth - lprior.th)) {
    th[,i] <- thp;llth <- llp;lprior.th <- lprior.p
    accept <- accept  + 1
  } else { ## reject
    th[,i] <- th[,i-1]
  }
}
accept/ns ## about 1/4 is ideal
```

Checking the chains

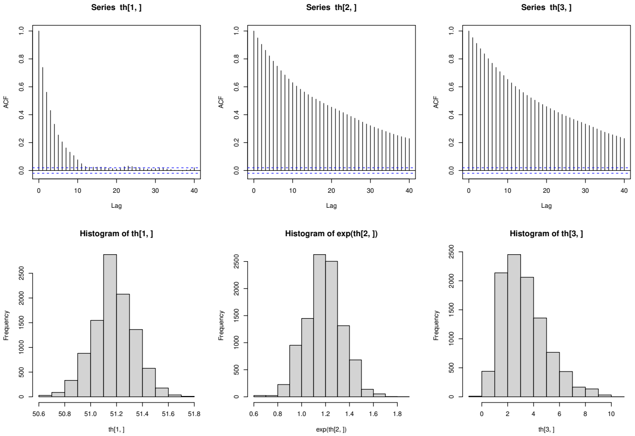
```
par(mfrow=c(3,1),mar=c(4,4,1,1))
plot(th[1,],type="l")
plot(th[2,],type="l")
plot(th[3,],type="l")
```



…quick convergence, but mixing fairly slow.

Chain correlation and marginal posteriors

```
par(mfrow=c(2,3))
acf(th[1,]);acf(th[2,]);acf(th[3,]);
hist(th[1,]);hist(exp(th[2,]));hist(th[3,]);
```



…standard deviation and degrees of freedom quite highly correlated.

CIs, posterior means etc.

```
> pm <- rowMeans(th) ## posterior mean
> ## transform to original scale...
> pm[2:3] <- exp(pm[2:3])
> pm[3] <- pm[3] + 1
> names(pm) <- c("mu","sig","df")
> pm
       mu       sig        df
51.175612  1.176176 27.191217
>
> ## 95% Credible Intervals...
> ci <- apply(th,1,quantile,prob=c(.025,.975))
> ci[,2:3] <- exp(ci[,2:3]) ;ci[,3] <- ci[,3]+1
> colnames(ci) <- c("mu","sig","df")
> ci
            mu      sig         df
2.5%  50.85020 0.893452    3.09977
97.5% 51.48983 1.484936 1766.29037
```