Metropolis and Metropolis-Hastings I

Dr. Olanrewaju Michael Akande

April 3, 2020

Announcements

- Reminder: let the instructor know if you plan to request a letter grade.

Outline

- Bayesian model selection and averaging
  - Recap
  - Model selection and averaging for linear regression models
  - Example
- Metropolis algorithm
  - Introduction and intuition
  - Algorithm
  - Illustration

Bayesian model selection and model averaging

Recap

- General setting:

1. Define a list of models. That is, let Γ be the "finite" set of different possible models.
2. Each model γ is in Γ, including the "true" model. Also, let θ_γ represent the parameters in model γ.
3. Put a prior over the set Γ. Let $\Pi_\gamma = \Pr[\gamma]$, for all $\gamma \in \Gamma$. Most common choice is the uniform prior, that is, $\Pi_\gamma = \frac{1}{\#\Gamma}$, for all $\gamma \in \Gamma$, where $\#\Gamma$ is the total number of models in Γ.
4. Put a prior on the parameters in each model, that is, each $\pi(\theta_\gamma)$.
5. Compute marginal posterior probabilities $\Pr[\gamma|Y]$ for each model.

Recap

- For each model $\gamma \in \Gamma$, need to compute $\Pr[\gamma|Y]$.
- Let $\mathcal{L}_\gamma(Y)$ denote the marginal likelihood of the data under model γ,then

$$\begin{aligned}\hat{\Pi}_\gamma = \Pr[\gamma|Y] &= \frac{\mathcal{L}_\gamma(Y)\Pi_\gamma}{\sum_{\gamma^\star\in\Gamma}\mathcal{L}_{\gamma^\star}(Y)\Pi_{\gamma^\star}} \\ &= \frac{\Pi_\gamma \cdot \left[\int_{\Theta_\gamma}\mathcal{L}_\gamma(Y|\theta_\gamma)\cdot\pi(\theta_\gamma)\mathrm{d}\theta_\gamma\right]}{\sum_{\gamma^\star\in\Gamma}\mathcal{L}_{\gamma^\star}(Y)\Pi_{\gamma^\star}}.\end{aligned}$$

- If we assume a uniform prior on Γ, that is, $\Pi_\gamma = \frac{1}{\#\Gamma}$, for all $\gamma \in \Gamma$, then

$$\begin{aligned}\hat{\Pi}_\gamma &= \frac{\mathcal{L}_\gamma(Y)}{\sum_{\gamma^\star\in\Gamma}\mathcal{L}_{\gamma^\star}(Y)} \\ &= \frac{\left[\int_{\Theta_\gamma}\mathcal{L}_\gamma(Y|\theta_\gamma)\cdot\pi(\theta_\gamma)\mathrm{d}\theta_\gamma\right]}{\sum_{\gamma^\star\in\Gamma}\mathcal{L}_{\gamma^\star}(Y)}.\end{aligned}$$

Recap

- How should we choose the Bayes optimal model?
- If loss function is

$$L(\hat{\gamma}, \gamma) = \mathbf{1}(\hat{\gamma} \neq \gamma),$$

that is,

1. Loss equals zero if the correct model is chosen; and
2. Loss equals one if incorrect model is chosen.

- Then, selecting the model with the largest posterior probability minimizes the corresponding Bayes risk.
- If goal is prediction, then

$$p(y_{n+1}|Y = (y_1, \ldots, y_n)) = \sum_{\gamma \in \Gamma} \hat{\Pi}_\gamma \cdot p(y_{n+1}|Y, \gamma),$$

which is known as Bayesian model averaging (BMA).

Back to Bayesian linear regression

- So what does this mean specifically in the context of linear regression?
- First, recall that for model γ, the posterior probability that the model is the right model is

$$\hat{\Pi}_\gamma = \frac{\Pi_\gamma \mathcal{L}_\gamma(Y)}{\sum_{\gamma^\star \in \Gamma} \Pi_{\gamma^\star} \mathcal{L}_{\gamma^\star}(Y)}.$$

- *Practical issues*
  - We need to calculate marginal likelihoods for ALL models in Γ.
  - In general for, we cannot calculate the marginal likelihoods unless we have a proper or conjugate priors.
  - For linear regression, that would mean looking to priors like Zellner's g-prior, the horseshoe prior you were introduced to in the lab, and so on.

Bayesian variable selection

- To explore Bayesian variable selection, rewrite each model $\gamma \in \Gamma$ as

$$\boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 \boldsymbol{I}_{n \times n}).$$

- γ represents the set of predictors we want to throw into our model.
- Using the notation as before, each $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_{p-1}) \in \{0, 1\}^p$, so that the cardinality of Γ is 2^p, that is, the number of models in Γ.
- That is,
  - $\gamma_j = 1$ means the j'th predictor is included in the model, but $\gamma_j = 0$ means it is not;
  - $\boldsymbol{X}_\gamma$ is the matrix of predictors with $\gamma_j = 1$;
  - $\boldsymbol{\beta}_\gamma$ is the corresponding vector of predictors with $\gamma_j = 1$.
- Set $p_\gamma = \sum_{j=1}^{p} \gamma_j$, so that p_γ is the number of predictors included in model γ, then $\boldsymbol{X}_\gamma$ is $n \times p_\gamma$ and $\boldsymbol{\beta}_\gamma$ is $p_\gamma \times 1$.

Bayesian variable selection

- Recall that we can also write each model as

$$Y_i = \boldsymbol{\beta}_\gamma^T \boldsymbol{x}_{i\gamma} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

- As an example, suppose we had data with 6 predictors including the intercept, so that each $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$.

- Then for model with $\gamma = (1, 1, 0, 0, 0, 0)$, $Y_i = \boldsymbol{\beta}_\gamma^T \boldsymbol{x}_{i\gamma} + \epsilon_i$

$$\implies Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

with $p_\gamma = 2$.

- Whereas for model with $\gamma = (1, 0, 0, 1, 1, 0)$, $Y_i = \boldsymbol{\beta}_\gamma^T \boldsymbol{x}_{i\gamma} + \epsilon_i$

$$\implies Y_i = \beta_0 + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2),$$

with $p_\gamma = 3$.

Bayesian variable selection

- The outline for variable selection would be as follows:

  1. Write down likelihood under model γ. That is,

$$p(\boldsymbol{y}|\boldsymbol{X},\gamma,\boldsymbol{\beta}_\gamma,\sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)^T(\boldsymbol{y}-\boldsymbol{X}_\gamma\boldsymbol{\beta}_\gamma)\right\}$$

  2. Define a prior for γ, $\Pi_\gamma = \Pr[\gamma]$. For example, (i) uniform over all 2^p possible models, or even (ii) beta prior (since each $\gamma_j \in \{0,1\}$).

  3. Put a prior on the parameters in each model. Using the g-prior, we have

$$\pi(\boldsymbol{\beta}_\gamma|\sigma^2) = \mathcal{N}_p\left(\boldsymbol{\beta}_{0\gamma} = \mathbf{0}, \Sigma_{0\gamma} = g\sigma^2\left[\boldsymbol{X}_\gamma^T\boldsymbol{X}_\gamma\right]^{-1}\right)$$
$$\pi(\sigma^2) = \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$$

Bayesian variable selection

- With those pieces, the conditional posteriors are straightforward.
- We can then compute marginal posterior probabilities $\Pr[\gamma|Y]$ for each model and select model with the highest posterior probability.
- We can also compute posterior $\Pr[\gamma_j|Y]$, the posterior probability of including the j'the predictor, often called marginal inclusion probability (MIP), allowing for uncertainty in the other predictors.
- Also straightforward to do model averaging once we have all posterior samples.
- The Hoff book works through one example and you can find the Gibbs sampler for doing inference there. I strongly recommend you go through it carefully!
- In class however, let's focus on using R packages for doing the same.

Example

- Health plans use many tools to try to control the cost of prescription medicines.
- For older drugs, generic substitutes that are the equivalent to name-brand drugs are available at considerable savings.
- Another tool that may lower costs is restricting drugs that the physician may prescribe.
- For example if three similar drugs for treating the same condition are available, a health plan may require the physician to prescribe only one of them, allowing the plan to negotiate discounts based on a higher volume of sales.
- We have data from 29 health plans can be used to explore the effectiveness of these two strategies in controlling drug costs.
- The response is COST, the average cost of the prescriptions to the plan per day (in dollars).

Example

- Potential explanatory variables are:
  - RXPM: Average number of prescriptions per member per year
  - GS: Percent generic substitute used by the plan
  - RI: Restrictiveness Index, from 0 (no restrictions) to 100 (total restrictions on the physician)
  - COPAY: Average member copay on prescriptions
  - AGE: Average member age
  - F: percent female members
  - MM: Member months, a measure of the size of the plan
  - ID: an identifier for the name of the plan
- Since we do not have so many data points, let's use Bayesian model selection and model averaging to explore the relationship of GS and RI to COST, adjusting for the other variables.
- The data is in the file `costs.txt` on Sakai.

In-class analysis: move to the R script here.

Metropolis algorithm

Introduction

- So far in this course, inference has been made "relatively" easy with **conjugate** and **semi-conjugate** priors.
- As we have seen, under conjugate or semi-conjugate priors, posteriors can be approximated with the Monte Carlo method or Gibbs sampler.
- However, sometimes a conjugate prior is unavailable or undesirable!
- In such cases, the full conditional distributions of parameters often have no standard form, and Gibbs sampling cannot be easily used.
- So what can we do?
- Metropolis and Metropolis-Hastings algorithms provide a generic method of approximating the posterior distribution corresponding to any combination of prior and data model.

Introduction

- As a refresher, suppose $Y \sim \pi(y|\theta)$ and suppose we specify a prior $\pi(\theta)$ on θ.
- Then as usual, we are interested in

$$\pi(\theta|y) = \frac{\pi(\theta)L(y;\theta)}{\mathcal{L}(y)}.$$

- As we already know, the challenge is that it is often difficult to compute $\mathcal{L}(y)$.
- Using the Monte Carlo method or Gibbs sampler, we have seen that we don't need to know $\mathcal{L}(y)$. As long as we have conjugate and semi-conjugate priors, we can generate samples directly from $\pi(\theta|y)$.
- So again, the question is, what happens if we cannot sample directly from $\pi(\theta|y)$?

Motivating example

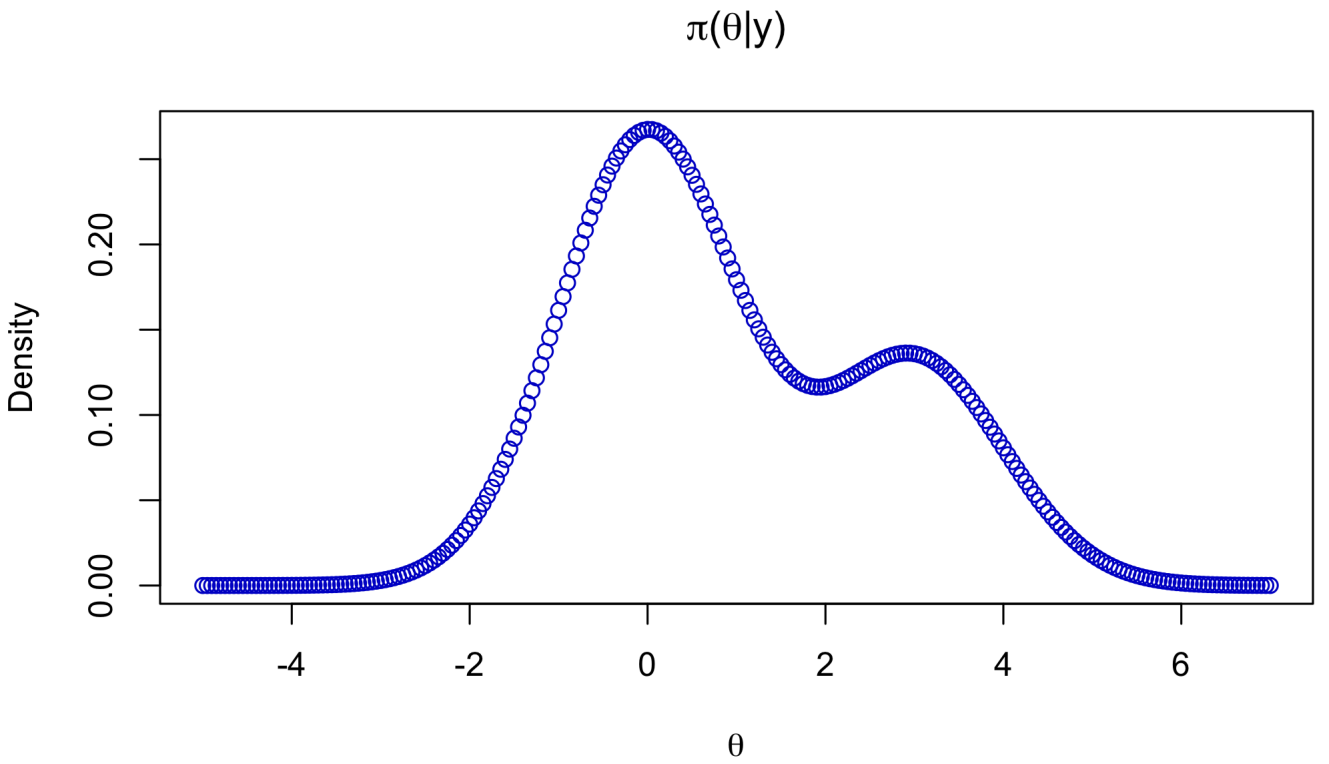
- To motivate our discussions on the Metropolis algorithm, let's explore a simple example (somewhat different from what we have seen so far).
- Suppose we wish to sample from the following density

$$\pi(\theta|y) \propto \exp^{-\frac{1}{2}\theta^2} + \frac{1}{2}\exp^{-\frac{1}{2}(\theta-3)^2}$$

- This is a *mixture of two normal densities*, one with mode near 0 and the other with mode near 3. Finite mixtures models remains the most likely topic we will cover next Friday.
- Anyway, let's use this density to explore the main ideas behind the Metropolis sampler.
- By the way, as you will see, we actually don't need to know the normalizing constant for Metropolis sampling but for this example, we will find it in class for practice.

Motivating example

- Let's take a look at the (normalized) density:



- There are other ways of sampling from this density, but let's focus specifically on the Metropolis algorithm here.

Metropolis algorithm

- From a sampling perspective, we need to have a large group of values, $\theta^{(1)}, \ldots, \theta^{(S)}$ from $\pi(\theta|y)$ whose empirical distribution approximates $\pi(\theta|y)$.
- That means that for any any two values θ_a and θ_b, we want

$$\frac{\#\theta^{(s)} = a}{S} \div \frac{\#\theta^{(s)} = b}{S} = \frac{\#\theta^{(s)} = a}{S} \times \frac{S}{\#\theta^{(s)} = b} = \frac{\#\theta^{(s)} = a}{\#\theta^{(s)} = b} \approx \frac{\pi(\theta_a|y)}{\pi(\theta_b|y)}$$

- Basically, we want to make sure that if θ_a and θ_b are in $\pi(\theta|y)$, the ratio of the number of the $\theta^{(1)}, \ldots, \theta^{(S)}$ values equal to them properly approximates $\frac{\pi(\theta_a|y)}{\pi(\theta_b|y)}$.
- How might we construct a group like this?

Metropolis algorithm

- Suppose we have a working group $\theta^{(1)}, \ldots, \theta^{(s)}$ at iteration s, and need to add a new value $\theta^{(s+1)}$.
- Consider a candidate value $\theta^\star$ *(we will get to how to generate the candidate value in a minute)* that is close to $\theta^{(s)}$. Should we set $\theta^{(s+1)} = \theta^\star$ or not?
- Well, we should probably compute $\pi(\theta^\star|y)$ and see if $\pi(\theta^\star|y) > \pi(\theta^{(s)}|y)$. Equivalently, look at $r = \dfrac{\pi(\theta^\star|y)}{\pi(\theta^{(s)}|y)}$.
- By the way, notice that

$$r = \frac{\pi(\theta^\star|y)}{\pi(\theta^{(s)}|y)} = \frac{\mathcal{L}(y|\theta^\star)\pi(\theta^\star)}{\mathcal{L}(y)} \div \frac{\mathcal{L}(y|\theta^{(s)})\pi(\theta^{(s)})}{\mathcal{L}(y)}$$

$$= \frac{\mathcal{L}(y|\theta^\star)\pi(\theta^\star)}{\mathcal{L}(y)} \times \frac{\mathcal{L}(y)}{\mathcal{L}(y|\theta^{(s)})\pi(\theta^{(s)})} = \frac{\mathcal{L}(y|\theta^\star)\pi(\theta^\star)}{\mathcal{L}(y|\theta^{(s)})\pi(\theta^{(s)})},$$

which does not depend on the marginal likelihood we don't know!

Metropolis algorithm

- If $r > 1$
  - Intuition: $\theta^{(s)}$ is already a part of the density we desire and the density at $\theta^\star$ is even higher than the density at $\theta^{(s)}$.
  - Action: set $\theta^{(s+1)} = \theta^\star$
- If $r < 1$,
  - Intuition: relative frequency of values on our group $\theta^{(1)}, \ldots, \theta^{(s)}$ equal to $\theta^\star$ should be $\approx r = \dfrac{\pi(\theta^\star|y)}{\pi(\theta^{(s)}|y)}$. For every $\theta^{(s)}$, include only a fraction of an instance of $\theta^\star$.
  - Action: set $\theta^{(s+1)} = \theta^\star$ with probability r and $\theta^{(s+1)} = \theta^{(s)}$ with probability $1 - r$.

Metropolis algorithm

- This is the basic intuition behind the Metropolis algorithm.
- Where should the proposed value $\theta^\star$ come from?
- Sample $\theta^\star$ close to the current value $\theta^{(s)}$ using a symmetric proposal distribution $g[\theta^\star|\theta^{(s)}]$. g is actually a "family of proposal distributions", indexed by the specific value of $\theta^{(s)}$.
- Here, symmetric means that $g[\theta^\star|\theta^{(s)}] = g[\theta^{(s)}|\theta^\star]$.
- The symmetric proposal is usually very simple with density concentrated near $\theta^{(s)}$, for example, $\mathcal{N}(\theta^\star; \theta^{(s)}, \delta^2)$ or $\text{Unif}(\theta^\star; \theta^{(s)} - \delta, \theta^{(s)} + \delta)$.
- After obtaining $\theta^\star$, either add it or add a copy of $\theta^{(s)}$ to our current set of values, depending on the value of r.

Metropolis algorithm

- The algorithm proceeds as follows:

  1. Given $\theta^{(1)}, \ldots, \theta^{(s)}$, generate a *candidate* value $\theta^\star \sim g[\theta^\star | \theta^{(s)}]$.

  2. Compute the acceptance ratio

  $$r = \frac{\pi(\theta^\star | y)}{\pi(\theta^{(s)} | y)} = \frac{\mathcal{L}(y|\theta^\star)\pi(\theta^\star)}{\mathcal{L}(y|\theta^{(s)})\pi(\theta^{(s)})},$$

  3. Set

  $$\theta^{(s+1)} = \begin{cases} \theta^\star & \text{with probability} \quad \min(r, 1) \\ \theta^{(s)} & \text{with probability} \quad 1 - \min(r, 1) \end{cases}$$

  which can be accomplished by sampling $u \sim U(0, 1)$ independently and setting

  $$\theta^{(s+1)} = \begin{cases} \theta^\star & \text{if} \quad u < r \\ \theta^{(s)} & \text{if} \quad \text{otherwise} \end{cases}.$$

Metropolis algorithm

- Once we obtain the samples, then we are back to using Monte Carlo approximations for quantities of interest.
- That is, we can again approximate posterior means, quantiles, and other quantities of interest using the empirical distribution of our sampled values.
- *Some notes:*
  - The Metropolis chain ALWAYS moves to the proposed $\theta^\star$ at iteration $s+1$ if $\theta^\star$ has higher target density than the current $\theta^{(s)}$.
  - Sometimes, it also moves to a $\theta^\star$ value with lower density in proportion to the density value itself.
  - This leads to a random, Markov process than naturally explores the space according to the probability defined by $\pi(\theta|y)$, and hence generates a sequence that, while dependent, eventually represents draws from $\pi(\theta|y)$.

Metropolis algorithm: convergence

- We will not cover the convergence theory behind Metropolis chains in detail, but below are a few notes for those interested:
  - The Markov process generated under this condition is ergodic and has a limiting distribution.
  - Here, think of ergodicity as meaning that the chain can move anywhere at each step, which is ensured, for example, if $g[\theta^\star|\theta^{(s)}] > 0$ everywhere!
  - By construction, it turns out that the Metropolis chains are reversible, so that convergence to $\pi(\theta|y)$ is assured.
  - Think of reversibility as being equivalent to symmetry of the joint density of two consecutive $\theta^{(s)}$ and $\theta^{(s+1)}$ in the stationary process, which we do have by using a symmetric proposal distribution.
- If you want to learn more about convergence of MCMC chains, consider taking one of the courses on stochastic processes, or Markov chain theory.

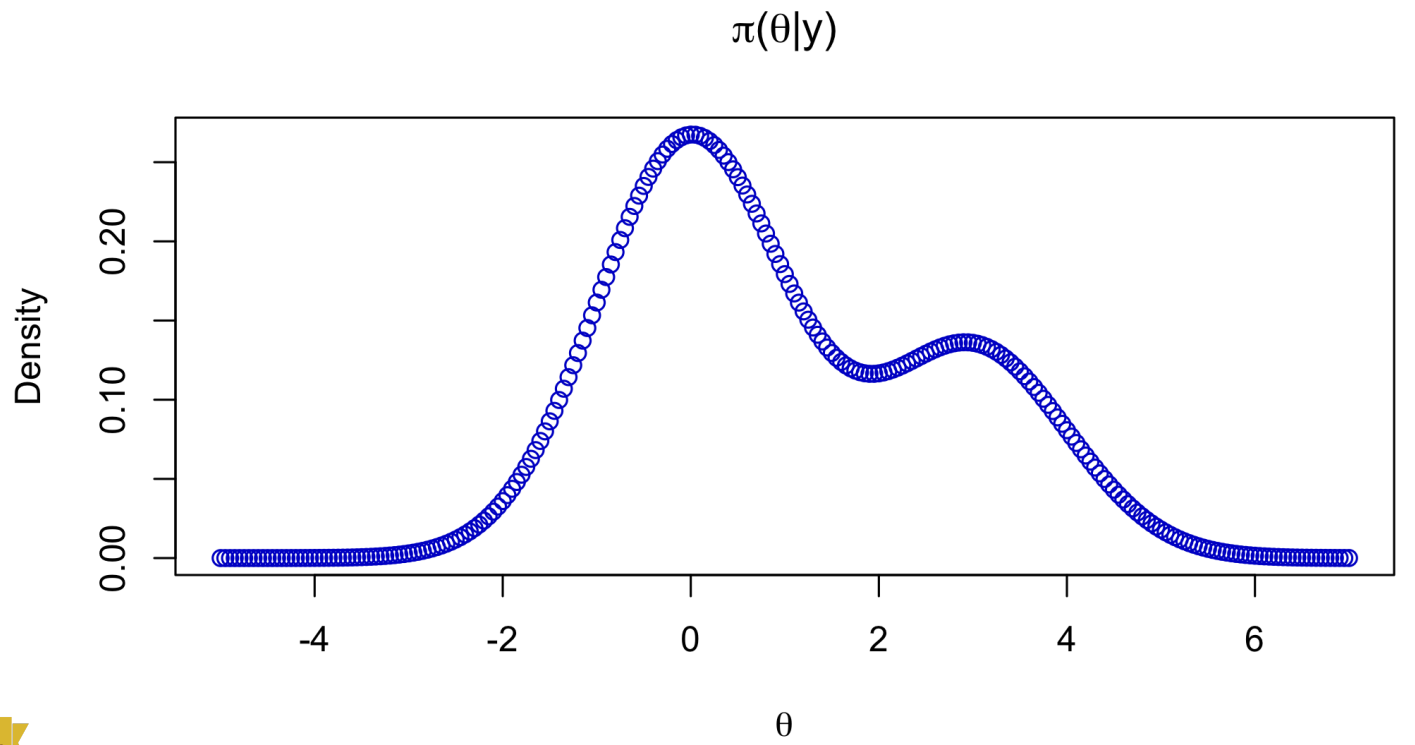
Metropolis algorithm: tuning

- Correlation between samples can be adjusted by selecting optimal δ (i.e., spread of the distribution) in the proposal distribution
- Decreasing correlation increases the effective sample size, increasing rate of convergence, and improving the Monte Carlo approximation to the posterior.
- However,
  - δ too small leads to $r \approx 1$ for most proposed values, a high acceptance rate, but very small moves, leading to highly correlated chain.
  - δ too large can get "stuck" at the posterior mode(s) because $\theta^\star$ can get very far away from the mode, leading to a very low acceptance rate and again high correlation in the Markov chain.
- Thus, good to implement several short runs of the algorithm varying δ and settle on one that yields acceptance rate in the range of 25-50%.
- Burn-in is even more important here!

Metropolis in action

Back to our example with

$$\pi(\theta|y) \propto \exp^{-\frac{1}{2}\theta^2} + \frac{1}{2}\exp^{-\frac{1}{2}(\theta-3)^2}$$



In-class analysis: move to the R script here.

Poisson regression

Count data

- In the next class, we will use the Metropolis sampler on count data with predictors, so let's first do some general review.
- Suppose you have count data (non-negative integers) as your response variable.
- For example, we may want to explain the number of c-sections carried out in hospitals using potential predictors such as hospital type, (that is, private vs public), location, size of the hospital, etc.
- The models we have covered so far are not (completely) adequate for count data with predictors.
- Of course there are instances where linear regression, with some transformations (especially taking logs) on the response variable, might still work reasonably well for count data.
- That's not the focus here, so we won't cover that.

Poisson regression

- As we have seen so far, a good distribution for modeling count data with no limit on the total number of counts is the Poisson distribution.
- As a reminder, the Poisson distribution is parameterized by λ and the pmf is given by

$$\Pr[Y = y] = \frac{\lambda^y e^{-\lambda}}{y!}; \quad y = 0, 1, 2, \ldots; \quad \lambda > 0.$$

- Remember that

$$\mathbb{E}[Y = y] = \mathbb{V}[Y = y] = \lambda.$$

- When our data fails this assumption, we may have what is known as over-dispersion and may want to consider the Negative Binomial distribution instead (actually easy to fit within the Bayesian framework!).
- With predictors, we want to index λ with i, where each λ_i is a function of $\boldsymbol{X}$. We can therefore write the random component of this glm as

$$y_i \sim \text{Poisson}(\lambda_i); \quad i = 1, \ldots, n.$$

Poisson regression

- We must ensure that $\lambda_i > 0$ at any value of $\boldsymbol{X}$, therefore, we need a link function that enforces this. A natural choice is the natural logarithm, so that we have

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}.$$

- Combining these pieces give us our full mathematical representation for the Poisson regression.
- For the frequentist version, in R, use the `glm` command but set the option `family = "poisson"`.
- Clearly, λ_i has a natural interpretation as the "expected count", and

$$\lambda_i = e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}}$$

  means that we can interpret the e^{β_j}'s as multiplicative effects on the expected counts.

Poisson regression

- For predictions, we can look at the expected counts, that is,

$$\hat{\lambda}_i = e^{\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \ldots + \hat{\beta}_p x_{ip}}$$

- Interpretation of e^{β_j}:
  - For continuous x_j: the expected count of Y increases by a multiplicative factor of e^{β_j} when increasing x_j by one unit.
  - For binary x_j: the expected count of Y increases by a multiplicative factor of e^{β_j} for the group with $x_j = 1$ in comparison to the group with $x_j = 0$.
- For example, suppose
  - Suppose the response variable is the number of mating for elephants, and let x_1 represent the age of the elephants
  - Also suppose $\hat{\beta}_j = 0.069$, so that $e^{\hat{\beta}_j} = e^{0.069} = 1.0714$.
  - Then, an increase in age of one year increases the expected number of mating for elephants by 7 percent.