Advanced Regression Methods for Independent Data

STAT/BIOST 570, 2020

Introduction to Bayesian Inference

Mauricio Sadinle

Department of Biostatistics

UNIVERSITY of WASHINGTON

Philosophical Motivation

- Frequentist paradigm: model parameters are unknown *constants*, and are treated as such
- Bayesian paradigm: model parameters are *treated* as *random variables* to represent uncertainty about their true values
- *Note*: the fact that Bayesians *treat* parameters as random variables, doesn't mean they really think parameters are random
    - For example, the median household income in the USA at a given point in time is a fixed number, but a Bayesian treats this parameter as a random quantity to quantify uncertainty about its true value

Philosophical Motivation

- Bayesian analysis requires:
  - Likelihood function, coming from a model for the distribution of the data
  - *Prior distribution* on model parameters, coming from previous knowledge about the phenomenon under study
- Prior distribution is *updated* using the likelihood via *Bayes' theorem*, resulting in a *posterior distribution*
- Bayesian inferences are drawn from posterior distribution (updated knowledge)

Practical Motivation

Bayesian *machinery* might make this approach appealing

- Quantification of uncertainty in large discrete parameter spaces
  - Partitions in clustering problems
  - Graphs in graphical models
  - Binary vectors of variable inclusion in regression model selection
  - Bipartite matchings in record linkage
- Frequentist approach calls for constructing confidence sets, whereas Bayesian approach works with a posterior distribution from which we can sample to approximate summaries of interest
- Under some conditions, posteriors behave like sampling distribution of MLEs, but Bayesian machinery might be easier to implement than deriving estimate of asymptotic covariance matrix of MLE
- Convenient in hierarchical/multilevel models – priors just add another level to the hierarchy

The Likelihood Function

- $Z = (Z_1, \ldots, Z_K)$: generic vector of study variables
- We work under a *parametric model* for the distribution of Z

$$\{p(z \mid \theta)\}_\theta, \quad \theta = (\theta_1, \theta_2, \ldots, \theta_d)$$

- Data from random i.i.d. vectors $\{Z_i\}_{i=1}^n \equiv \boldsymbol{Z}$
- Under our parametric model, the joint distribution of $\{Z_i\}_{i=1}^n$ has a density function

$$p(\boldsymbol{z} \mid \theta) = \prod_{i=1}^n p(z_i \mid \theta)$$

- This, seen as a function of θ, is the likelihood function

$$L(\theta \mid \boldsymbol{z}) = \prod_{i=1}^n p(z_i \mid \theta)$$

The Prior Distribution

- Prior to observing the realizations of $\boldsymbol{Z} = \{Z_i\}_{i=1}^n$, do we have any information on the parameters θ?

- Represent this prior information in terms of a distribution

$$p(\theta)$$

The Posterior Distribution

Now, "*simply*" use Bayes' theorem

$$p(\theta \mid \boldsymbol{z}) = \frac{L(\theta \mid \boldsymbol{z})p(\theta)}{p(\boldsymbol{z})}$$

$$= \frac{L(\theta \mid \boldsymbol{z})p(\theta)}{\int L(\theta \mid \boldsymbol{z})p(\theta)d\theta}$$

$$\propto L(\theta \mid \boldsymbol{z})p(\theta)$$

"*That's it!*"

The Posterior Distribution

For simple problems, we typically have two ways of computing the posterior $p(\theta \mid \boldsymbol{z})$

- Compute the integral $\int L(\theta \mid \boldsymbol{z})p(\theta)d\theta$, and then compute

$$p(\theta \mid \boldsymbol{z}) = \frac{L(\theta \mid \boldsymbol{z})p(\theta)}{\int L(\theta \mid \boldsymbol{z})p(\theta)d\theta}$$

- Stare at / manipulate the expression $L(\theta \mid \boldsymbol{z})p(\theta)$ seen as a function of θ alone
  - If $L(\theta \mid \boldsymbol{z})p(\theta) = a(\theta, \boldsymbol{z})b(\boldsymbol{z})$, then $p(\theta \mid \boldsymbol{z}) \propto a(\theta, \boldsymbol{z})$
  - If $a(\theta, \boldsymbol{z})$ looks like a known distribution except for a constant, then we have identified the posterior

Example: Binomial Data, Beta Prior

Let $Z \mid \theta \sim \text{Binom}(n, \theta)$, and $\theta \sim \text{Beta}(a, b)$

- $L(\theta \mid z) = \binom{n}{z}\theta^z(1-\theta)^{n-z}$, $\quad p(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$

- The proportionality constant is

$$\begin{aligned}\int L(\theta \mid z)p(\theta)d\theta &= \frac{\binom{n}{z}}{B(a,b)} \int \theta^{z+a-1}(1-\theta)^{n-z+b-1}d\theta \\ &= \binom{n}{z}\frac{B(z+a, n-z+b)}{B(a,b)}\end{aligned}$$

- And the posterior is

$$\begin{aligned}p(\theta \mid z) &= \frac{L(\theta \mid z)p(\theta)}{\int L(\theta \mid z)p(\theta)d\theta} \\ &= \frac{1}{B(z+a, n-z+b)}\theta^{z+a-1}(1-\theta)^{n-z+b-1}\end{aligned}$$

- Therefore, $\theta \mid z \sim \text{Beta}(z+a, n-z+b)$

Example: Binomial Data, Beta Prior

Let $Z \mid \theta \sim \text{Binom}(n, \theta)$ and $\theta \sim \text{Beta}(a, b)$

- $L(\theta \mid z) = \binom{n}{z}\theta^z(1-\theta)^{n-z}$, $\quad p(\theta) = \frac{1}{B(a,b)}\theta^{a-1}(1-\theta)^{b-1}$
- We could also have noticed

$$\begin{aligned} p(\theta \mid z) &\propto L(\theta \mid z)p(\theta) \\ &\propto \theta^{z+a-1}(1-\theta)^{n-z+b-1} \end{aligned}$$

- This is the non-constant part (*the kernel*) of the density function of a beta random variable with parameters $z + a$ and $n - z + b$, therefore $\theta \mid z \sim \text{Beta}(z + a, n - z + b)$

Example: Binomial Data, Beta Prior

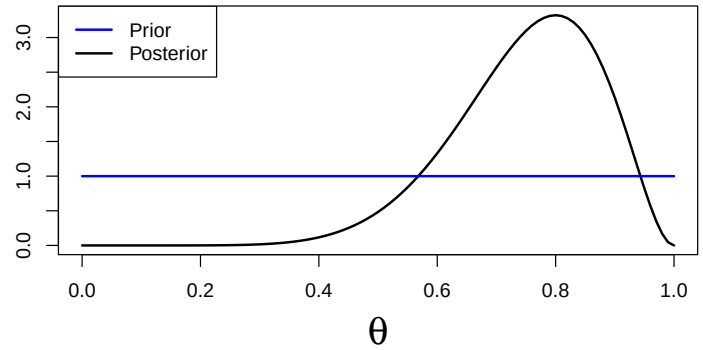
To illustrate the idea, say:

- Someone is flipping a coin $n = 10$ times in an independent and identical fashion
- Number of heads $Z \sim \text{Binomial}(n, \theta)$
- What is the value of θ?
- We use a $\text{Beta}(a, b)$ to express our *prior belief* on θ
- We observe $Z = 8$

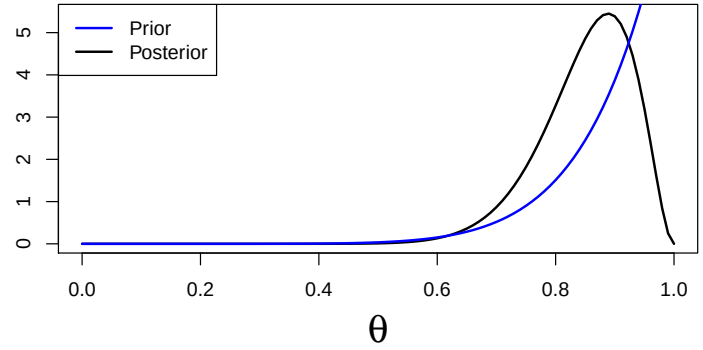
Example: Binomial Data, Beta Prior

Possible scenarios of prior information on θ; posteriors with $Z = 8$:

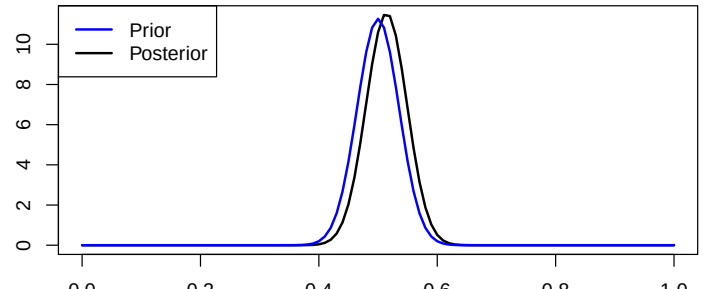
- No idea of what θ could be: Beta(1,1)



- The person flipping the coin looks like a trickster: Beta(9,1)



- Coin flipping usually has 50/50 chance of heads/tails: Beta(100,100)



Comments So Far

- Bayesian approach allows you to incorporate side information based on context
  - Do you know what "flipping a coin" means?
  - Is the person flipping the coin someone you trust?
  - You need to understand what θ represents

Example: Normal Data, Normal Prior

- Suppose we have

$$Z_i \mid \theta \overset{iid}{\sim} N(\theta, \sigma^2), \quad i = 1, \ldots, n,$$

with σ^2 assumed known and θ unknown.

- Say you express your prior in terms of a normal distribution

$$\theta \sim N(m, v)$$

where you choose m and v based on your knowledge of θ

- We can write

$$\begin{aligned} p(\theta \mid \boldsymbol{z}) &\propto L(\theta \mid \boldsymbol{z}) p(\theta) \\ &\propto \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(z_i-\theta)^2\right\} \frac{1}{(2\pi v)^{1/2}} \exp\left\{-\frac{1}{2v}(\theta-m)^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left[\frac{v+\sigma^2/n}{v\sigma^2/n}\theta^2 - 2\theta(n\bar{z}/\sigma^2 + m/v)\right]\right\}, \end{aligned}$$

and need to complete squares to identify the kernel of a normal...

Example: Normal Data, Normal Prior

- After completing squares, we find that

$$\theta \mid \boldsymbol{z} \sim N\left(\bar{z} \times w + m \times (1 - w), \frac{\sigma^2}{n} \times w\right),$$

where $w = \frac{v}{v + \sigma^2/n}$.

- Think about cases
  - $n = 0$: recover the prior
  - $n \to \infty$: $w \to 1$ and data dominates posterior, unless $v = 0$
  - $v \to 0$: posterior $\to$ prior
  - $v \to \infty$: $w \to 1$, improper prior, leads to

$$\theta \mid \boldsymbol{z} \sim N\left(\bar{z}, \frac{\sigma^2}{n}\right),$$

and recall that with normal data we have

$$\bar{Z} \sim N\left(\theta, \frac{\sigma^2}{n}\right),$$

so that frequentist and Bayesian estimates coincide in this case

Example: Multinomial Data, Dirichlet Prior

- Let $Z_i = (Z_{i1}, Z_{i2})$, $\quad Z_{i1}, Z_{i2} \in \{1, 2\}$, $\quad Z_i$'s are i.i.d.,

$$p(Z_{i1} = k, Z_{i2} = l \mid \theta) = \pi_{kl}$$

- $\theta = (\ldots, \pi_{kl}, \ldots)$, $\qquad W_{ikl} = I(Z_{i1} = k, Z_{i2} = l)$

- The likelihood of the study variables is

$$\begin{aligned} L(\theta \mid \boldsymbol{z}) &= \prod_i \left[\prod_{k,l} \pi_{kl}^{W_{ikl}} \right] \\ &= \prod_{k,l} \pi_{kl}^{n_{kl}} \end{aligned}$$

where

$$n_{kl} = \sum_i W_{ikl}, \quad k, l \in \{1, 2\}$$

Example: Multinomial Data, Dirichlet Prior

- Inference on multinomial parameters is convenient using the *Dirichlet* prior

- $\theta = (\ldots, \pi_{kl}, \ldots) \sim \text{Dirichlet}(\alpha), \quad \alpha = (\ldots, \alpha_{kl}, \ldots),$

$$p(\theta) = \frac{\Gamma(\sum \alpha_{kl})}{\prod_{k,l} \Gamma(\alpha_{kl})} \prod_{k,l} \pi_{kl}^{\alpha_{kl}-1}$$

- The posterior is given by

$$\begin{aligned} p(\theta \mid \boldsymbol{z}) &\propto L(\theta \mid \boldsymbol{z}) p(\theta) \\ &\propto \prod_{k,l} \pi_{kl}^{n_{kl}+\alpha_{kl}-1} \end{aligned}$$

- Therefore, $\theta \mid \boldsymbol{z} \sim \text{Dirichlet}(\alpha'), \quad \alpha' = (\ldots, \alpha_{kl} + n_{kl}, \ldots)$

Comments on Priors

- There is no guarantee that the combination of arbitrary likelihoods and priors will lead to posteriors that are easy to work with
- *Conjugate priors* are distributions that lead to posteriors in the same family, as in the previous examples – typically easier to work with, but not always available
- Non-conjugate priors can be used, but we require more advanced techniques for handling them
- Lists of conjugate priors are available in multiple books and online resources

Example: Multivariate Normal

- Distribution of the data

$$\boldsymbol{Z} = \{Z_i\}_{i=1}^n \mid \mu, \Lambda \overset{i.i.d.}{\sim} \text{Normal}(\mu, \Lambda^{-1})$$

where $Z_i \in \mathbb{R}^K$, μ is the vector of means, Λ^{-1} is the covariance matrix, and Λ is the inverse covariance matrix (the *precision matrix*)

- Conjugate prior is constructed in two steps

$$\mu \mid \Lambda \sim \text{Normal}(\mu_0, (\kappa_0 \Lambda)^{-1})$$
$$\Lambda \sim \text{Wishart}(\upsilon_0, W_0)$$

Joint distribution of (μ, Λ) is called *Normal-Wishart*. The parameterization is such that $E(\Lambda) = \upsilon_0 W_0$.

Example: Multivariate Normal

- Posterior is also Normal-Wishart

$$\mu \mid \Lambda, \boldsymbol{z} \sim \text{Normal}(\mu', (\kappa' \Lambda)^{-1})$$
$$\Lambda \mid \boldsymbol{z} \sim \text{Wishart}(\upsilon', W')$$

where

$$\mu' = (\kappa_0 \mu_0 + n\bar{z})/\kappa'$$
$$\kappa' = \kappa_0 + n$$
$$\upsilon' = \upsilon_0 + n$$
$$W' = \{W_0^{-1} + n[\hat{\Sigma} + \frac{\kappa_0}{\kappa'}(\bar{z} - \mu_0)(\bar{z} - \mu_0)^T]\}^{-1}$$
$$\bar{z} = \sum_{i=1}^{n} z_i / n$$
$$\hat{\Sigma} = \sum_{i=1}^{n} (z_i - \bar{z})(z_i - \bar{z})^T / n$$

- HW: think about what happens to $\text{E}(\mu \mid \boldsymbol{z})$ when $\kappa_0 \to 0$ and when $\kappa_0 \to \infty$

Comments So Far

- Expressing prior information works nicely with some parametric families
- Quantification of prior information can be tricky, especially for complicated models
- Example: Say $Y > 0$, $\quad x_1, x_2 \in \{0, 1\}$

  $$\log \mathsf{E}(Y \mid x_1, x_2) = \beta_{11}x_1x_2 + \beta_{10}x_1(1-x_2) + \beta_{01}(1-x_1)x_2 + \beta_{00}(1-x_1)(1-x_2)$$

  then

  $$\beta_{x_1x_2} = \log \mathsf{E}(Y \mid x_1, x_2),$$

  or

  $$\exp \beta_{x_1x_2} = \mathsf{E}(Y \mid x_1, x_2),$$

  so if you have prior information for $\mathsf{E}(Y \mid x_1, x_2)$ then you can transfer that to $\beta_{x_1x_2}$

Comments So Far

- Example (cont'd): If you parameterize the above model as

$$\log \mathsf{E}(Y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

then we have

$$\exp \beta_0 = \mathsf{E}(Y \mid x_1 = 0, x_2 = 0),$$

$$\exp \beta_1 = \frac{\mathsf{E}(Y \mid x_1 = 1, x_2 = 0)}{\mathsf{E}(Y \mid x_1 = 0, x_2 = 0)},$$

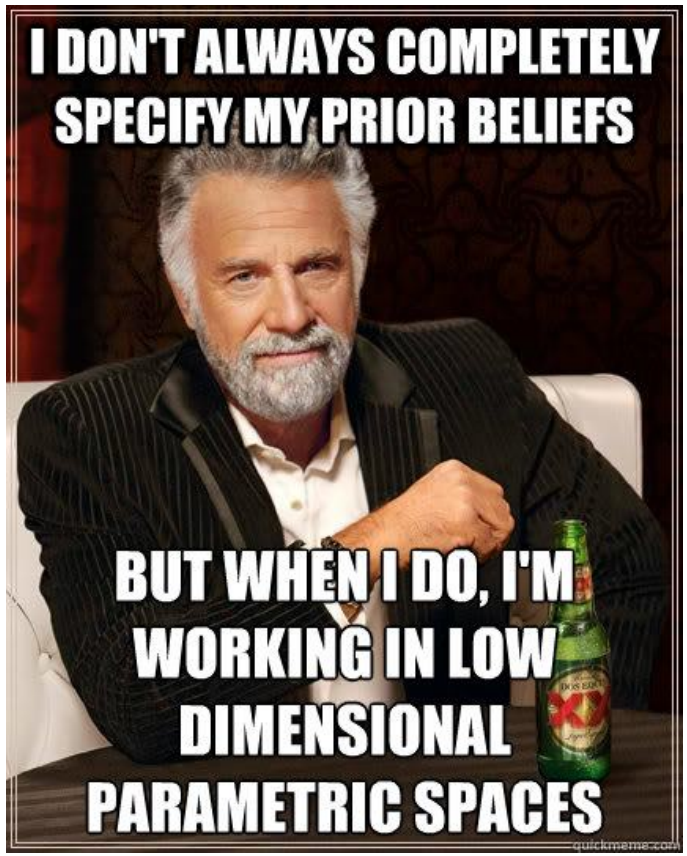
$$\exp \beta_2 = \frac{\mathsf{E}(Y \mid x_1 = 0, x_2 = 1)}{\mathsf{E}(Y \mid x_1 = 0, x_2 = 0)},$$

$$\exp \beta_{12} = \frac{\mathsf{E}(Y \mid x_1 = 1, x_2 = 1)\mathsf{E}(Y \mid x_1 = 0, x_2 = 0)}{\mathsf{E}(Y \mid x_1 = 1, x_2 = 0)\mathsf{E}(Y \mid x_1 = 0, x_2 = 1)}$$

- Understanding what a parameter means is important for prior specification!

Comments So Far

Quantification of prior information can be tricky, especially for complicated models



With complicated models, Bayesians often default to convenient and/or vague priors

Bayesian Point Estimation

- $\mathcal{L}(\theta, \theta')$: loss of estimating a parameter to be θ' when the true value is θ
- Bayes estimator minimizes the expected posterior loss

$$\hat{\theta}_{\text{Bayes}} = \arg\min_{\theta'} \int \mathcal{L}(\theta, \theta') p(\theta \mid \boldsymbol{z}) d\theta$$

- For univariate θ it is common to choose
  - $\mathcal{L}(\theta, \theta') = (\theta - \theta')^2 \implies \hat{\theta}_{\text{Bayes}}$ is posterior mean
  - $\mathcal{L}(\theta, \theta') = |\theta - \theta'| \implies \hat{\theta}_{\text{Bayes}}$ is posterior median
  - $\mathcal{L}(\theta, \theta') = I(\theta \neq \theta') \implies \hat{\theta}_{\text{Bayes}}$ is posterior mode

Bayesian Credible Sets/Intervals

- C is a $(1-\alpha)100\%$ *credible set* if

$$\int_C p(\theta \mid \boldsymbol{z}) d\theta \geq 1-\alpha$$

- For univariate θ, define the $(1-\alpha)100\%$ *credible interval* C as the interval within the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior $p(\theta \mid \boldsymbol{z})$

Asymptotic Behavior of Posteriors: Bernstein - von Mises

- Under some conditions, the posterior distribution asymptotically behaves like the sampling distribution of the MLE

- Heuristically, the Bernstein - von Mises theorem says[1]

$$p(\theta \mid \boldsymbol{z}) \approx N(\hat{\theta}_{\text{MLE}}, \mathcal{I}_n(\hat{\theta}_{\text{MLE}})^{-1})$$

- Therefore, for well-behaved models and with a good amount of data, Bayesian and frequentist inferences will be similar

- Important conditions:
  - The prior should not exclude any region of the parameter space
  - The prior should not be data-dependent

[1] See lecture notes of Richard Nickl: `http://www.statslab.cam.ac.uk/~nickl/Site/__files/stat2013.pdf` or Ferguson (1996), *A Course in Large Sample Theory*

Summary So Far

Bayesian inference offers alternative framework for deriving inferences from data

- Philosophical motivation: inclusion of prior belief or knowledge, uncertainty quantification in terms of distributions for parameters
- Practical motivation: convenient in some problems, might lead to good frequentist performance
- Complex problems are computationally challenging – posterior needs to be approximated (e.g., Markov chain Monte Carlo)

Monte Carlo Approximations for Bayesian Inference

For Bayesian inference we work with the posterior

$$p(\theta \mid \boldsymbol{z}) = \frac{L(\theta \mid \boldsymbol{z})p(\theta)}{\int L(\theta \mid \boldsymbol{z})p(\theta)d\theta}$$

- This expression might not be available in closed form
- Computing functionals of interest $E[h(\theta) \mid \boldsymbol{z}]$ might be complicated
- Idea: sample from $p(\theta \mid \boldsymbol{z})$ and approximate functionals of interest via Monte Carlo

Monte Carlo Approximations for Bayesian Inference

- Say you are able to obtain independent samples

$$\{\theta^{(t)}\}_{t=1}^{m} \overset{i.i.d.}{\sim} p(\theta \mid \boldsymbol{z})$$

- We can use these to approximate

$$\mathrm{E}[h(\theta) \mid \boldsymbol{z}] \approx \frac{1}{m} \sum_{t=1}^{m} h(\theta^{(t)})$$

- The Strong Law of Large Numbers tells us:

$$\frac{1}{m} \sum_{t=1}^{m} h(\theta^{(t)}) \to_{a.s.} \mathrm{E}[h(\theta) \mid \boldsymbol{z}]$$

- The Central Limit Theorem tells us:

$$\sqrt{m} \left[\frac{1}{m} \sum_{t=1}^{m} h(\theta^{(t)}) - \mathrm{E}[h(\theta) \mid \boldsymbol{z}] \right] \to_{d} \mathrm{N}(0, \mathrm{var}[h(\theta) \mid \boldsymbol{z}])$$

- How are we supposed to sample from $p(\theta \mid \boldsymbol{z})$ if this is an unknown or complicated distribution?

Rejection Sampling

In simple problems, one option for sampling from $p(\theta \mid \boldsymbol{z})$ is *rejection sampling*:

- Say you want to sample from a distribution with density function

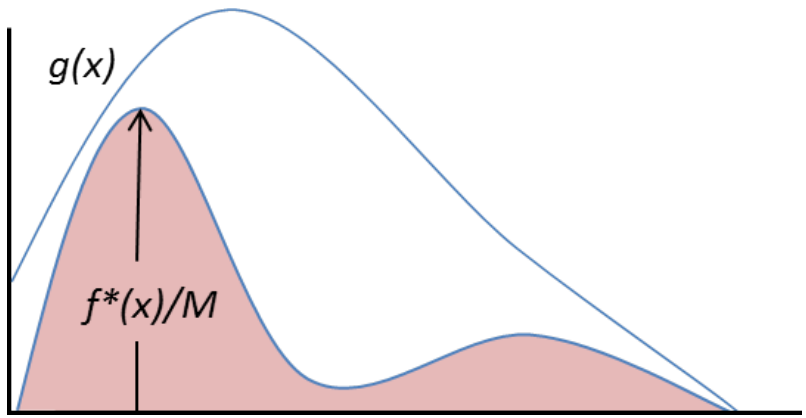
$$f(x) = f^*(x)/c$$

where you know $f^*(x)$ but don't know $c = \int f^*(x)dx$

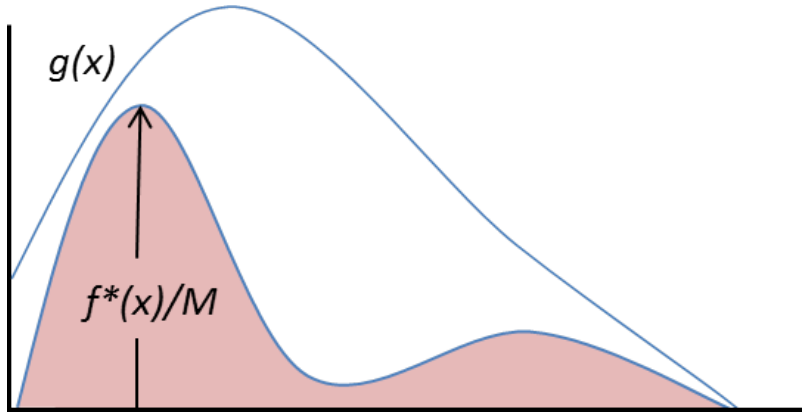
- Say you can sample from a distribution with density $g(x)$ such that

$$g(x) \geq f^*(x)/M$$

for some constant M and for all x



Rejection Sampling



The *rejection sampling* algorithm to sample from a distribution with density $f(x) = f^*(x)/c$:

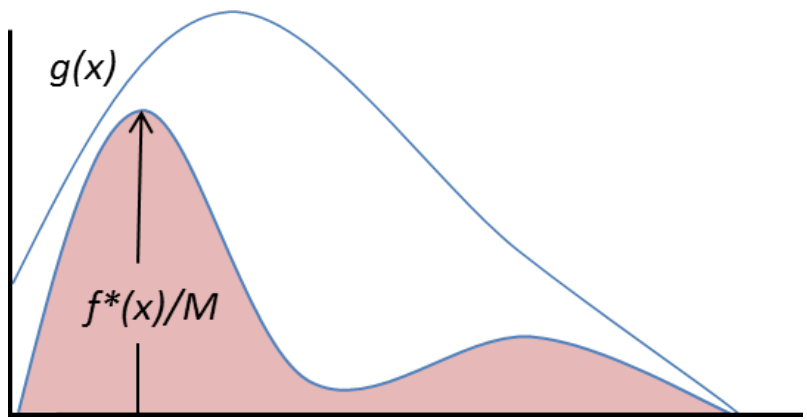
1. Generate $X \sim g(\cdot)$ and $V \mid X \sim U[0, g(X)]$.
2. Keep X if
$$V \leq f^*(X)/M,$$
otherwise reject X and return to step 1.

Rejection Sampling

- Let $X \sim g(\cdot)$, $A = 1$ if the draw is accepted, $Y \sim f(\cdot)$. We can show that

$$(X \mid A = 1) \stackrel{d}{=} Y$$

- This algorithm can be quite inefficient if the sampling density does not closely follow the target density (think about the area in between $g(x)$ and $f^*(x)/M$ in the plot below)



Importance Sampling

Rather than rejecting/accepting draws, *importance sampling* uses weights for the draws generated from some density $g(\cdot)$

- We have

$$\mathrm{E}_f[h(X)] = \int h(x)f(x)dx = \int \frac{h(x)f(x)}{g(x)}g(x)dx = \mathrm{E}_g[h(X)w(X)],$$

where $w(x) = f(x)/g(x)$

- Taking $X_i \overset{i.i.d.}{\sim} g(\cdot)$, the *importance sampling* estimator is

$$\hat{\mathrm{E}}_f[h(X)] = \frac{1}{m}\sum_{i=1}^{m} h(X_i)w(X_i)$$

- If $f(\cdot)$ and/or $g(\cdot)$ are known up to a proportionality constant, the *self-normalized importance sampling* estimator can be used instead

$$\hat{\mathrm{E}}_f[h(X)] = \frac{\sum_{i=1}^{m} h(X_i)w(X_i)}{\sum_{i=1}^{m} w(X_i)}, \qquad X_i \overset{i.i.d.}{\sim} g(\cdot).$$

Comments on Posterior Approximations

- Rejection and importance sampling offer workable solutions mostly for problems with a small number of parameters
- Non-MC alternatives include *Laplace approximations* to posterior expectations (Tierney and Kadane, 1986), and a more modern approach called *Integrated Nested Laplace Approximation* – INLA (Rue, Martino, and Chopin, 2009)
- For problems with multiple parameters, however, the most common way of approximating the posterior is *Markov Chain Monte Carlo* (for an intro reading see Wakefield (2013, sec. 3.8))

Markov Chain Monte Carlo

Big picture:

- If we can construct a Markov chain $\boldsymbol{X}^{(0)}, \boldsymbol{X}^{(1)}, \ldots$ with invariant distribution $\pi(\cdot)$, then we know that after some t_0, $\boldsymbol{X}^{(t)} \sim \pi(\cdot)$ for $t > t_0$
- In Bayesian inference we take $\pi(\cdot)$ to be $p(\theta \mid \boldsymbol{z})$, but MCMC has broader applications
- Under some conditions on the Markov chain, we'll be able to approximate expectations with respect to $\pi(\cdot)$ as

$$\mathrm{E}[h(\boldsymbol{x})] = \int h(\boldsymbol{x})\pi(\boldsymbol{x})\, d\boldsymbol{x} \approx \frac{1}{m}\sum_{t=1}^{m} h(\boldsymbol{x}^{(t)})$$

  where $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}$ represents the sample path of the Markov chain
- *Ergodic theorems* exist which provide analogs to the CLT and SLLN in the i.i.d. case
- These theorems apply to the chains obtained with the MCMC methods presented here

Markov Chain Monte Carlo

- Consider a random variable $\boldsymbol{X}$ with support $\mathbb{R}^p$ and density $\pi(\cdot)$
- A sequence of random variables $\boldsymbol{X}^{(0)}, \boldsymbol{X}^{(1)}, \ldots$ is called a *Markov chain* if

$$\Pr\left(\boldsymbol{X}^{(t+1)} \in A \mid \boldsymbol{X}^{(t)}, \boldsymbol{X}^{(t-1)}, \ldots, \boldsymbol{X}^{(0)}\right) = \Pr\left(\boldsymbol{X}^{(t+1)} \in A \mid \boldsymbol{X}^{(t)}\right)$$

  for all t and for all measurable sets A, so that the probability of moving to any set A at time $t+1$ only depends on where we are at time t

- Furthermore, for a *homogeneous* Markov chain:

$$\Pr\left(\boldsymbol{X}^{(t+1)} \in A \mid \boldsymbol{X}^{(t)} = \boldsymbol{x}\right) = \Pr\left(\boldsymbol{X}^{(1)} \in A \mid \boldsymbol{X}^{(0)} = \boldsymbol{x}\right).$$

Markov Chain Monte Carlo

- If there exists $k(\boldsymbol{x}, \boldsymbol{y})$ such that

$$\Pr(\boldsymbol{Y} \in A \mid \boldsymbol{x}) = \int_A k(\boldsymbol{x}, \boldsymbol{y}) \, d\boldsymbol{y},$$

  then $k(\boldsymbol{x}, \boldsymbol{y})$ is called the *transition kernel density*

- We could also denote $k(\boldsymbol{x}, \boldsymbol{y}) \equiv k(\boldsymbol{x} \to \boldsymbol{y}) \equiv p_{Y|X}(\boldsymbol{y} \mid \boldsymbol{x})$ but many sources use the $k(\boldsymbol{x}, \boldsymbol{y})$ notation

- A probability distribution $\pi(\cdot)$ is called an *invariant distribution* of a Markov chain with transition kernel density $k(\boldsymbol{x}, \boldsymbol{y})$ if so-called *global balance* holds:

$$\pi(\boldsymbol{y}) = \int \pi(\boldsymbol{x}) k(\boldsymbol{x}, \boldsymbol{y}) \, d\boldsymbol{x}.$$

- A Markov chain is called *reversible* if

$$\pi(\boldsymbol{x}) k(\boldsymbol{x}, \boldsymbol{y}) = \pi(\boldsymbol{y}) k(\boldsymbol{y}, \boldsymbol{x})$$

  for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, $\boldsymbol{x} \neq \boldsymbol{y}$. This condition is known as *detailed balance*

The Metropolis-Hastings Algorithm

- It can be shown that if reversibility holds for $\pi(\cdot)$ and $k(\cdot, \cdot)$ then $\pi(\cdot)$ is the invariant distribution

- The *Metropolis-Hastings* algorithm provides a very flexible method for defining a Markov chain that satisfies detailed balance for a distribution $\pi(\cdot)$ of interest

- Given $\boldsymbol{x}^{(t)}$, the following steps provide the new point $\boldsymbol{x}^{(t+1)}$:

  1. Sample a point $\boldsymbol{y}$ from a *proposal* distribution $q(\cdot \mid \boldsymbol{x}^{(t)})$.

  2. Calculate the acceptance probability:

$$\alpha(\boldsymbol{x}^{(t)}, \boldsymbol{y}) = \min\left[\frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x}^{(t)})} \times \frac{q(\boldsymbol{x}^{(t)} \mid \boldsymbol{y})}{q(\boldsymbol{y} \mid \boldsymbol{x}^{(t)})}, 1\right].$$

  3. Set

$$\boldsymbol{x}^{(t+1)} = \begin{cases} \boldsymbol{y} & \text{with probability } \alpha(\boldsymbol{x}^{(t)}, \boldsymbol{y}) \\ \boldsymbol{x}^{(t)} & \text{otherwise.} \end{cases}$$

The Metropolis-Hastings Algorithm

- In a Bayesian context, this algorithm is appealing because the ratio $\pi(\boldsymbol{y})/\pi(\boldsymbol{x}^{(t)})$ doesn't depend on the normalizing constant of $\pi(\cdot)$, where $\pi(\cdot)$ is taken as the posterior
- It is easy to verify that the acceptance probability $\alpha(\boldsymbol{x}^{(t)}, \boldsymbol{y})$ leads to detailed balance for $\pi(\cdot)$; what is the transition kernel here?
- The Metropolis-Hastings algorithm has different special cases of interest

The Metropolis Algorithm

- Suppose the proposal distribution is *symmetric* in the sense that

$$q(\boldsymbol{y} \mid \boldsymbol{x}^{(t)}) = q(\boldsymbol{x}^{(t)} \mid \boldsymbol{y}).$$

- For example, in the *random walk* Metropolis algorithm $q(\boldsymbol{y} \mid \boldsymbol{x}^{(t)}) = q(\ |\boldsymbol{y} - \boldsymbol{x}^{(t)}|\ )$, with common choices for $q(\cdot)$ being normal or uniform distributions.

- In this case the acceptance probability simplifies to

$$\alpha(\boldsymbol{x}^{(t)}, \boldsymbol{y}) = \min\left[\ \frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x})}, 1\ \right]$$

Comments on the Practice of Metropolis-Hastings

- In practice, there is a tradeoff between having high acceptance rates with small movement or low acceptance rates with large movement
- Rule of thumb: aim for 30% acceptance rate (optimal in many circumstances). This may be obtained by tuning the proposal density, e.g., the variance in a normal proposal

The Gibbs Sampler

- Another particularly important special case of Metropolis-Hastings is the *Gibbs sampler*
- We describe two flavors: the *sequential* Gibbs sampler and the *random scan* Gibbs sampler
- In the following, let $\boldsymbol{x}_{-i}$ represent the vector $\boldsymbol{x}$ with the i-th variable removed, i.e.

$$\boldsymbol{x}_{-i} = [x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p]$$

The Gibbs Sampler

The *sequential scan Gibbs sampling* algorithm:

- Start with some initial value $\boldsymbol{x}^{(0)}$
- At step t, given current point

$$\boldsymbol{x}^{(t)} = [x_1^{(t)}, \ldots, x_p^{(t)}],$$

produce a new point

$$\boldsymbol{x}^{(t+1)} = [x_1^{(t+1)}, \ldots, x_p^{(t+1)}]$$

with the following p steps

- Sample $x_1^{(t+1)} \sim \pi_1\left(x_1 \mid \boldsymbol{x}_{-1}^{(t)}\right)$
- Sample $x_2^{(t+1)} \sim \pi_2\left(x_2 \mid x_1^{(t+1)}, x_3^{(t)}, \ldots, x_p^{(t)}\right)$
- Sample $x_3^{(t+1)} \sim \pi_3\left(x_3 \mid x_1^{(t+1)}, x_2^{(t+1)}, x_4^{(t)}, \ldots, x_p^{(t)}\right)$

$$\vdots$$

- Sample $x_p^{(t+1)} \sim \pi_p\left(x_p \mid \boldsymbol{x}_{-p}^{(t+1)}\right)$.

The Gibbs Sampler

- The advantage of Gibbs sampling comes from being able to generate the p-dimensional vector $\boldsymbol{X}^{(t+1)}$ by generating its entries individually from conditional distributions

- In many cases, these conditionals are known distributions; for example *conditional conjugacy* can be exploited:

  - *Conditional conjugacy*: when the conditional posterior

    $$p(\theta_j \mid \theta_{-j}, \boldsymbol{z}) \propto L(\theta \mid \boldsymbol{z}) p(\theta_j \mid \theta_{-j})$$

    is in the same family as the conditional prior $p(\theta_j \mid \theta_{-j})$.

The Gibbs Sampler

Gibbs sampling is a particular case of Metropolis-Hastings:

- Consider a single component move in the Gibbs sampler:
  - Current point: $\boldsymbol{x}^{(t)}$
  - Proposed point:

    $$\boldsymbol{y} = [x_1^{(t)}, \ldots, x_{i-1}^{(t)}, x_i^*, x_{i+1}^{(t)}, \ldots, x_p^{(t)}],$$

    obtained by replacing the i-th component in $\boldsymbol{x}^{(t)}$ with a draw x_i^* from the conditional $\pi_i\left(x_i \mid \boldsymbol{x}_{-i}^{(t)}\right)$.
- Taking $\boldsymbol{y}$ as a proposal in Metropolis-Hastings, we obtain

$$q(\boldsymbol{y} \mid \boldsymbol{x}^{(t)}) = \pi_i\left(x_i^* \mid \boldsymbol{x}_{-i}^{(t)}\right)$$

The Gibbs Sampler

Then the Metropolis-Hastings acceptance ratio becomes

$$\begin{aligned}
\alpha(\boldsymbol{x}^{(t)}, \boldsymbol{y}) &= \min\left[\frac{\pi(\boldsymbol{y})}{\pi(\boldsymbol{x}^{(t)})} \times \frac{q(\boldsymbol{x}^{(t)} \mid \boldsymbol{y})}{q(\boldsymbol{y} \mid \boldsymbol{x}^{(t)})}, 1\right] \\
&= \min\left[\frac{\pi\left(x_i^*, \boldsymbol{x}_{-i}^{(t)}\right)}{\pi\left(x_i^{(t)}, \boldsymbol{x}_{-i}^{(t)}\right)} \frac{\pi_i\left(x_i^{(t)} \mid \boldsymbol{x}_{-i}^{(t)}\right)}{\pi_i\left(x_i^* \mid \boldsymbol{x}_{-i}^{(t)}\right)}, 1\right] \\
&= \min\left[\frac{\pi\left(\boldsymbol{x}_{-i}^{(t)}\right)}{\pi\left(\boldsymbol{x}_{-i}^{(t)}\right)}, 1\right] = 1
\end{aligned}$$

because

$$\pi\left(x_i^\star \mid \boldsymbol{x}_{-i}^{(t)}\right) = \pi\left(x_i^*, \boldsymbol{x}_{-i}^{(t)}\right) / \pi\left(\boldsymbol{x}_{-i}^{(t)}\right).$$

- We always accept when using conditionals as our proposals in Metropolis-Hastings!
- This means that Gibbs sampling produces a Markov chain with stationary distribution $\pi(\cdot)$!

The Gibbs Sampler

- Updating each component sequentially is not the only way to execute Gibbs sampling (though it is the easiest to implement and the most common)
- We can also randomly select a component to update
- *Random scan* Gibbs sampling:
  - Sample a component i with probability $\alpha_i > 0$, $\sum_{i=1}^{p} \alpha_i = 1$
  - Sample $x_i^{(t+1)} \sim \pi_i \left(x_i \mid \boldsymbol{x}_{-i}^{(t)} \right)$

The Gibbs Sampler

- Sometimes it is easy to sample from the conditional distributions for groups of variables

- In a *blocked Gibbs sampler* we use

$$\boldsymbol{x} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_K],$$

where $\boldsymbol{x}_k$ is a group of variables, and we obtain $\boldsymbol{x}^{(t+1)}$ with the following K steps

  - Sample $\boldsymbol{x}_1^{(t+1)} \sim \pi\left(\boldsymbol{x}_1 \mid \boldsymbol{x}_2^{(t)}, \ldots, \boldsymbol{x}_K^{(t)}\right)$
  - Sample $\boldsymbol{x}_2^{(t+1)} \sim \pi\left(\boldsymbol{x}_2 \mid \boldsymbol{x}_1^{(t+1)}, \boldsymbol{x}_3^{(t)}, \ldots, \boldsymbol{x}_K^{(t)}\right)$

$$\vdots$$

  - Sample $\boldsymbol{x}_K^{(t+1)} \sim \pi\left(\boldsymbol{x}_K \mid \boldsymbol{x}_1^{(t+1)}, \boldsymbol{x}_2^{(t+1)}, \ldots, \boldsymbol{x}_{K-1}^{(t+1)}\right)$.

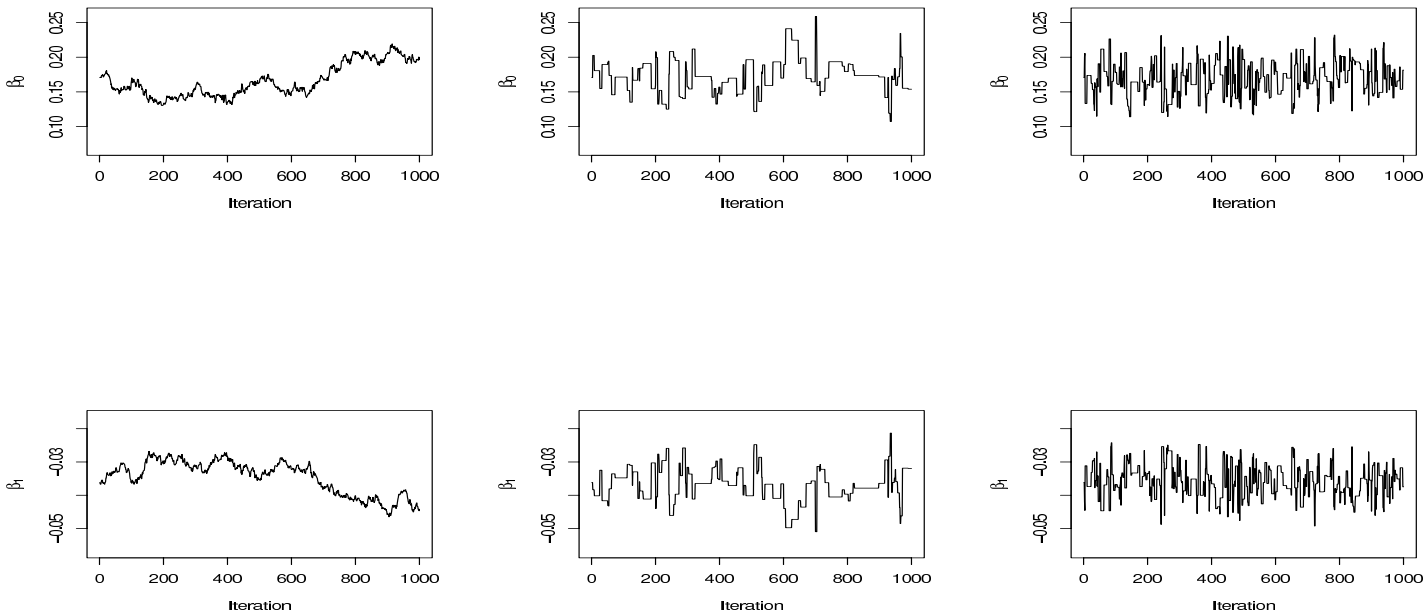
Combining Markov Kernels: Hybrid Schemes

- Say we can construct multiple transition kernels, each with invariant distribution $\pi(\cdot)$
- We can then construct a Markov chain, where at each step we sequentially generate new states from the different kernels in a predetermined order
- One popular example is *Metropolis within Gibbs* in which all conditionals are sampled with Gibbs steps for the recognizable conditionals and Metropolis-Hastings for the remainder

Diagnostics

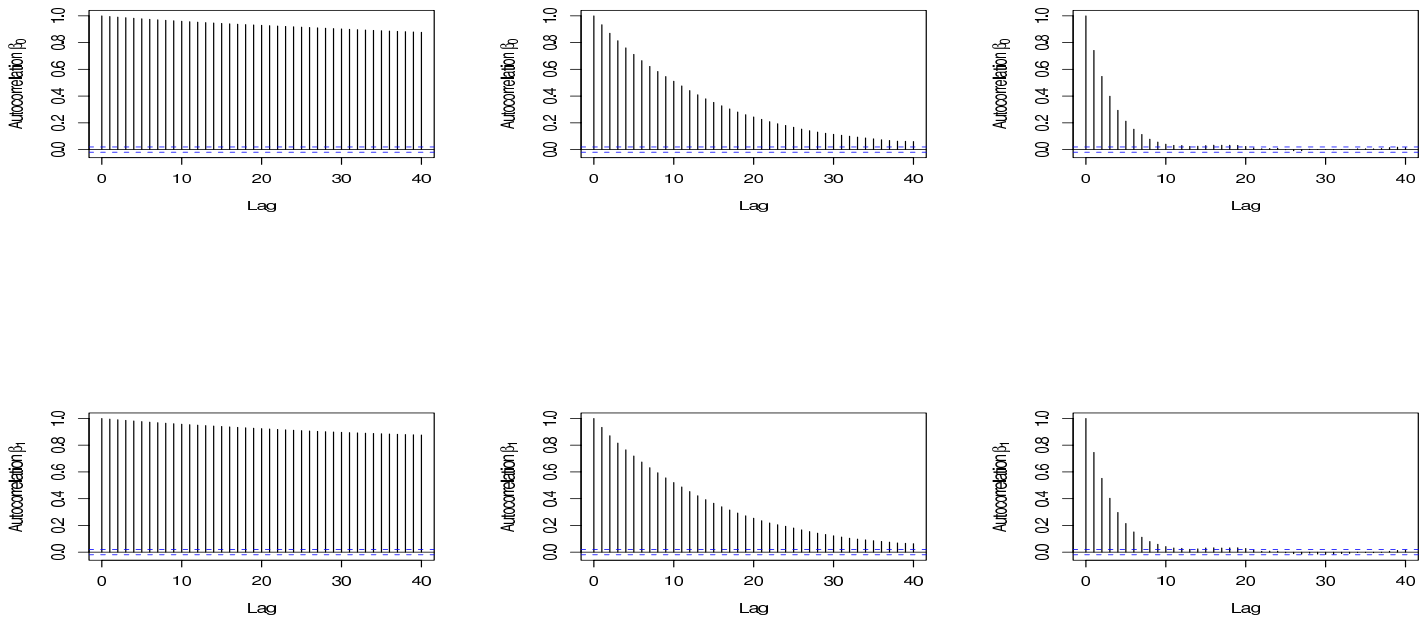
- Trace plots provide a useful method for detecting problems with MCMC convergence and mixing
- *Mixing* of the Markov chain: how well it moves through its sample space
- Ideally, trace plots of unnormalized log posterior and model parameters should look like stationary time series
- Slowly mixing Markov chains produce trace plots with high autocorrelation, which can be further visualized by autocorrelation plots at different lags
- Slow mixing does not imply lack of convergence, however, but that more samples will be required for accurate inference

Example



Example of traceplots from Metropolis algorithms for a regression problem presented in Section 3.8.6 of Wakefield (2013): β_0 in top row and β_1 in bottom row; left column: univariate proposals with small variance; center column: univariate proposals with large variance; right column: bivariate proposals.

Example



Autocorrelation functions for β_0 (top row) and β_1 (bottom row); left column: univariate proposals with small variance; center column: univariate proposals with large variance; right column: bivariate proposals.

Implementation Details

- How long to run the chain in order to obtain reliable Monte Carlo estimates of expectations?
  - Some chains may have very slow mixing and an examination of autocorrelation aids in deciding on the number of iterations required or on whether to re-design the Markov chain
  - The Markov chain will display better mixing properties if the parameters are approximately independent in the posterior
  - Dependence in the Markov chain may be greatly reduced by sampling simultaneously variables that are highly dependent, a strategy known as *blocking*
  - Reparameterization may also be helpful in this regard
- If storage of samples is an issue, or if one wants to retain a subchain with little autocorrelation, then one may decide to *thin* the chain by only collecting samples at equally spaced intervals