Chapter 9 Bayesian Computation Methods

C. Devon Lin

Queen's University, Nov 14, 2019

Outline

- 9.1 Background and concepts
- 9.2 Monte Carlo method for computing integrals
- 9.3 Rejection sampling
- 9.4 Importance sampling
- 9.5 Metropolis-Hastings algorithm
- 9.6 Gibbs sampling

Reference: Course notes,Chapters 1-3, 6, 7 of *Monte Carlo Statistical Methods* by Robert and Casella,Chapters 10 - 13 of *Bayesian Data Analysis* by Gelman.

Background and concepts

(a) **Prior**: In Bayesian statistics, the parameter θ is considered to be a quantity whose variation can be described by a probability distribution. This distribution is called *prior distribution* $\pi(\theta)$.

(b) **Posterior**: A sample is taken from a population indexed by θ and the prior distribution is updated with this sample information. The updated prior is called the *posterior distribution* $\pi(\theta|x)$.

(c) **Bayes' Rule**

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)} = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\theta)\pi(\theta)d(\theta)}$$

An example

Example

Let $X_1, \ldots, X_n$ be n i.i.d random variables with mean zero and unknown variance σ^2. The likelihood function is then give by

$$L(X|\sigma^2) \propto (\sigma^2)^{-n/2} \exp\{-\sum_{i=1}^{n} X_i^2/(2\sigma^2)\}.$$

Suppose the prior distribution for σ^2 is noninformative, that is, $\pi(\sigma^2) \propto 1/\sigma^2$ Then the posterior density of σ^2 is

$$\pi(\sigma^2|X_1, \ldots, X_n) = \frac{\pi(\sigma^2) f(X_1, \ldots, X_n|\sigma^2)}{\int \pi(\sigma^2) f(X_1, \ldots, X_n|\sigma^2) d(\sigma^2)} \propto (\sigma^2)^{-n/2-1} \exp\{-\sum_{i=1}^{n} X_i^2/(2\sigma^2)\}.$$

It can be shown σ^2 follows scaled inverse chi-squared distribution.

Conjugate priors

- Conjugate prior: belonging to a specific distributional family $\pi(\theta)$, with the likelihood $f(x|\theta)$, it leads to a posterior distribution $p(\theta|x)$ belong to the same distribution family as the prior.

An example of conjugate priors

Example

Suppose X is the number of pregnant women arriving at a particular hospital to deliver their babies during a given month. The discrete count nature of the data plus its natural interpretation as an arrival rate suggest adopting a Poisson likelihood,

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!}, x = 0, 1, \ldots, \theta > 0.$$

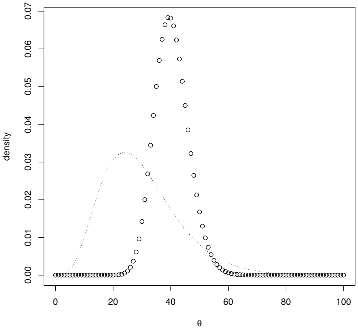
Suppose the prior is a gamma distribution,

$$\pi(\theta) = \frac{\theta^{\alpha-1}e^{-\theta/\beta}}{\Gamma(\alpha)\beta^\alpha}, \theta > 0, \alpha > 0, \beta > 0.$$

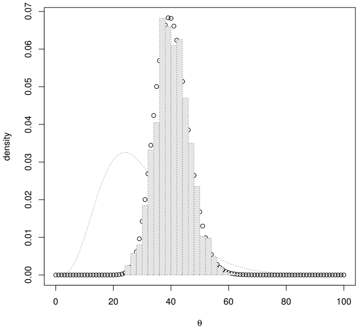
The posterior distribution would be

$$p(\theta|x) \propto \theta^{x+\alpha-1}e^{-\theta(1+1/\beta)}.$$

Gamma prior and posterior



Posterior draws of Gamma



Other conjugate priors

- Beta
- Gamma
- Dirichlet
- Gaussian
- Inverse Gamma
- Wishart
- Inverse-Wishart

Non-informative priors

- Non-informative priors: a prior that contains no information about the parameter θ, that is, the prior is "flat" relative to the likelihood function.
- Other names: vague, diffuse, and flat prior

Examples of non-informative priors

- If $0 \leq \theta \leq 1$, *Uniform*$(0, 1)$ is a non-informative prior for θ.
- If $-\infty < \theta < \infty$, $N(\theta_0, \sigma_0^2)$ and $\sigma_0^2 \to \infty$ forms a non-informative prior.
- If $-\infty < \theta < \infty$, $\pi(\theta) = c$ where c is a constant
- Jeffery's prior

Jeffreys' priors

- Jeffery's Rule: a rule for the choice of a non-informative prior
- Jeffreys' priors: the prior is given by

$$\pi(\theta) \propto |I(\theta)|^{1/2},$$

where $I(\theta)$ is the expected Fisher information.

Examples of Jeffreys' priors

Example

Suppose $y_1, \ldots, y_n \overset{iid}{\sim}$ Binomial$(1, \theta)$. Derive Jeffreys' prior.

$$
\begin{aligned}
I(\theta) &= E\left(-\frac{\partial^2 log f(y|\theta)}{\partial\theta^2}\right) \\
&= E\left(\frac{y_i}{\theta^2} + \frac{1-y_i}{(1-\theta)^2}\right) \\
&= \frac{1}{\theta} + \frac{1}{1-\theta} \\
&= \frac{1}{\theta(1-\theta)}
\end{aligned}
$$

$$\pi(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} = \theta^{\frac{1}{2}-1}(1-\theta)^{\frac{1}{2}-1}.$$

Improper priors

- Improper priors: $\int \pi(\theta)d(\theta) = \infty$
- Improper priors can lead to proper or improper posterior.
- Example: $y_1, \ldots, y_n \overset{iid}{\sim} N(\theta, 1)$ and $\pi(\theta) \propto 1$. Drive the posterior distribution of θ.

$$
\begin{aligned}
\pi(\theta|y_1, \ldots, y_n) &\propto f(y_1, \ldots, y_n|\theta)\pi(\theta) \\
&\propto (\frac{1}{\sqrt{2\pi}})^2 exp\{-\frac{\sum_{i=1}^n (y_i - \theta)^2}{2}\} \\
&\propto exp\{-\frac{\sum_{i=1}^n (\theta^2 + y_i^2 - 2\theta y_i)}{2}\} \\
&\propto exp\{-\frac{n\theta^2 - 2\theta \sum_{i=1}^n y_i}{2}\} \\
&\propto exp\{-\frac{(\theta - \frac{\sum_{i=1}^n y_i}{n})^2}{2\frac{1}{n}}\}
\end{aligned}
$$

Informative priors

- An informative prior is a type of prior that is not dominated by the likelihood function and has an impact on the posterior.
- For example, $y_1, \ldots, y_n \overset{iid}{\sim} N(\theta, 5)$, $\theta \sim N(0, 1)$.
- Some choices of informative priors
  - $\theta \in R$: normal distribution or t distribution
  - $\theta > 0$: gamma, inverse gamma, lognnormal
  - $\theta \in (0, 1)$: Beta distribution

Hierarchical priors

Example

An insect lays a large number of eggs, each surviving with probability p. On average, how many eggs will survive? Let X be the number of survivors and Y be the number of eggs laid. We have $X|Y \sim binomial(Y, p)$ and $Y \sim Poisson(\lambda)$. Thus,

$$E(X) = E(E(X|Y)) = E(pY) = p\lambda.$$

Hierarchical priors

Example

Consider a generalization Example 4, where instead of one mother insect there are a large number of mothers, and one mother is chosen at random. Let X be the number of survivors, then $X|Y \sim binomial(Y, p)$, $Y|\Lambda \sim Poisson(\Lambda)$, and $\Lambda \sim exponential(\beta)$. Thus,

$$E(X) = E(E(X|Y)) = E(pY) = E(E(pY|\Lambda)) = E(p\Lambda) = p\beta.$$

Computing integral

Suppose we wish to compute a complex integral,

$$\int_a^b w(x)d(x).$$

If we can decompose $w(x)$ into the product of a function $h(x)$ and a probability density function $f(x)$ defined over the interval (a, b), we then have

$$\int_a^b w(x)d(x) = \int_a^b h(x)f(x)d(x) = \mathrm{E}_{f(x)}[h(x)].$$

Monte Carlo integration draws a large number of $x_1, \ldots, x_n$ of random variables from $f(x)$, then

$$\int_a^b w(x)d(x) = \mathrm{E}_{f(x)}[h(x)] \simeq \frac{1}{n}\sum_{i=1}^{n} h(x_i)$$

MC integration

```
> # compute normal cdf by Monte Carlo integration
> t<-0
> n<-10000
> mean(rnorm(n)<t)
[1] 0.5021
> mean(rnorm(n)<1.96)
[1] 0.9749
```

MC integration

Obtain the integral of the following function

$$h(x) = [cos(50x) + sin(20x)]^2$$

Monte Carlo method for estimating the mean of $h(\theta)$

Suppose θ has a posterior density $\pi(\theta|x)$ and we are interested in the mean of $h(\theta)$, given by

$$\mathrm{E}(h(\theta)|x) = \int h(\theta)\pi(\theta|x)d\theta.$$

To obtain a Monte Carlo estimate, we simulate an independent sample $\theta^1, \ldots, \theta^m$ from the posterior density $\pi(\theta|x)$. The Monte Carlo estimate is given by the sample mean

$$\bar{h} = \sum_{j=1}^{m} h(\theta^j)/m$$

and its associated simulation standard error is

$$se_{\bar{h}} = \sqrt{\sum_{j=1}^{m}(h(\theta^j) - \bar{h})^2/[(m-1)m]}.$$

An example

Let p be the proportion of the American college students who sleep at least eight hours. We are interested in estimating p. We now take a sample of 27 students. Among them, 11 has at least eight hours of sleep. If we regard a "success" as sleeping at least eight hours and we take a random sample with s successes and f failures, then the likelihood function is given by

$$L(p) \propto p^S (1-p)^f, 0 \leq p \leq 1.$$

The posterior density for p is $\pi(p|data) \propto \pi(p)L(p)$. Suppose that the prior distribution is chosen to be

$$\pi(p) \propto p^{a-1}(1-p)^{b-1}, 0 \leq p \leq 1.$$

The posterior density is

$$\pi(p|data) \propto p^{a+s-1}(1-p)^{b+f-1}, 0 \leq p \leq 1.$$

Estimating p^2

Now suppose $a = 3.26, b = 7.19$. If now we are interested in the posterior mean of p^2.

```
> ### estimating the mean of p^2
> p<- rbeta(1000,14.26,23.19)
> est<-mean(p^2)
> se<-sd(p^2)/sqrt(1000)
> c(est,se)
[1] 0.149490312 0.001850406
```

Rejection sampling

(a) Goal: generate random sample $x \sim f(x)$, where f is the pdf of X.

(b) Procedure:

Step 1: Independently generate y from the probability density g and $U \sim U(0, 1)$.

Step 2: Accept $x = y$ if $U \leq \frac{f(y)}{cg(y)}$, where $c = \max \frac{f(y)}{g(y)}$.

Step 3: Continue Steps 1 and 2 until one has collected a sufficient number of accepted x's.

Justifications

- The acceptance probability $P(U < \frac{f(y)}{cg(y)}) = \frac{1}{c}$.
- $P(X \leq x | \text{X is accepted}) = F(x)$

Example 1

Example

Generate a random variable with the p.d.f

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Consider the proposed Cauchy distribution, $g_X(x) = \frac{1}{\pi(1+x^2)}$. Now choose c such that

$$c = \max\left(\frac{f_X(x)}{g_X(x)}\right) = \sqrt{\frac{2\pi}{e}},$$

and thus

$$\frac{f_X(x)}{c g_X(x)} = e^{-\frac{x^2}{2}}(1 + x^2)\sqrt{e}/2.$$

Example 1

Algorithm:

1. generate $U \sim U(0,1)$ and $y \sim Cauchy(0,1)$.

2 let $x = y$ if $U \leq e^{-\frac{x^2}{2}}(1+x^2)\sqrt{e}/2$

3. repeat 1 and 2 many times.

Example 2

Example

Generate a random variable with the p.d.f

$$f_X(x) = \frac{2}{\pi r^2}\sqrt{r^2 - x^2}, -r \leq x \leq r.$$

Consider the proposed distribution, $g_X(x) = \frac{1}{2r}, -r \leq x \leq r$. Now choose c such that

$$c = \max(\frac{f_X(x)}{g_X(x)}) = \frac{4}{\pi},$$

and thus

$$\frac{f_X(x)}{cg_X(x)} = \frac{4\pi}{4\pi r}\sqrt{r^2 - x^2} = \frac{1}{r}\sqrt{r^2 - x^2}.$$

Example 2

Algorithm:

1. generate $U \sim U(0,1)$ and $y \sim U(-r, r)$.

2 let $x = y$ if $U \leq \frac{1}{r}\sqrt{r^2 - y^2}$

3. repeat 1 and 2 many times.

Importance sampling

- Motivation
  Certain values of the random variable have more impact than others when computing $E_{f(x)}[h(x)]$. If we sample more frequently these "important" values then the variance of the estimator can be reduced.

- Method
  Suppose we can generate a sample $x_1, \ldots, x_n$ from a given distribution g. Then we approximate $E_{f(x)}[h(x)]$ in this way

$$E_{f(x)}[h(x)] \simeq \frac{1}{n} \sum_{i=1}^{n} \frac{f(x_i)}{g(x_i)} h(x_i)$$

where $w(x_i) = \frac{f(x_i)}{g(x_i)}$ is referred to as weights.

Properties of importance sampling

- The expected values of the weights is $E_g(\frac{f(X)}{g(X)}) = 1$.
- Bias and variance of $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(x_i)}{g(x_i)} h(x_i)$

$$E_g(\hat{\mu}) = \mu = \mathrm{E}_{f(x)}[h(x)]$$

$$Var_g(\hat{\mu}) = \frac{Var_g(\frac{f(X)}{g(X)} h(X))}{n}$$

An example

Example

Importance sampling for normal tails

We wish to estimate $\theta = P(X > c)$ where $X \sim N(0, \sigma^2)$ and $c > 3\sigma$. We use the following distribution for the proposed distribution,

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2}).$$

Thus the weights will be

$$\frac{f(x)}{g(x)} = \exp(-\frac{x^2 - (x-\mu)^2}{2\sigma^2}) = \exp(\frac{\mu(\mu - 2x)}{2\sigma^2}).$$

The importance sampling estimator for θ is

$$\hat{\theta} = \frac{1}{n} \sum_{j=1}^{n} I(x_j > c) \frac{f(x_j)}{g(x_j)} = \frac{1}{n} \sum_{j=1}^{n} I(x_j > c) \exp(\frac{\mu(\mu - 2x_j)}{2\sigma^2})$$

An example

The results for $c = 3$, $\sigma = 1$

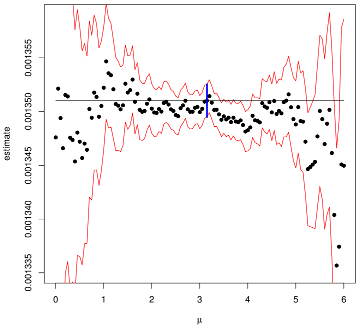
μ	estimate	standard deviation	L95	U95
0	1.347600e-03	11.60079e-06	1.324863e-03	1.370337e-03
1	1.352221e-03	2.906081e-06	1.346525e-03	1.357916e-03
2	1.350277e-03	1.175928e-06	1.347973e-03	1.352582e-03
3	1.349946e-03	7.855588e-07	1.348407e-03	1.351486e-03
3.1	1.350928e-03	7.805866e-07	1.349398e-03	1.352458e-03
3.15	1.351018e-03	7.798168e-07	1.349490e-03	1.352546e-03
3.20	1.351434e-03	7.804941e-07	1.349905e-03	1.352964e-03
4.	1.348260e-03	9.766230e-07	1.346346e-03	1.350174e-03

An example

The results for $c = 4.5$, $\sigma = 1$

μ	estimate	standard deviation	L95	U95
0.0	3.300000e-06	5.744553e-07	2.174088e-06	4.425912e-06
4.5	3.397519e-06	2.423692e-09	3.392769e-06	3.402270e-06
4.6	3.400929e-06	2.417336e-09	3.396191e-06	3.405667e-06
4.7	3.402426e-06	2.424090e-09	3.397675e-06	3.407178e-06

An example


0.001335
0.001340
0.001345
0.001350
0.001355
estimate
0
1
2
3
4
5
6
μ

Choice of $g(x)$

We wish to choose $g(x)$ such that the estimator obtained by importance sampling has finite variance.
Sufficient conditions: $f(x) < Mg(x)$ and $Var_f(h(x)) < \infty$.
If $f(x)$ has heavier tails than $g(x)$, the corresponding estimator could have infinite variance.
Theorem (for choosing optimal $g(x)$)
The proposed distribution $g(x)$ that minimizes the variance of $\hat{\mu}$ is

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(t)|f(t)dt}.$$

Remark: The theorem has little practical use. Choose $g(x)$ such that it is close to $|h(x)|f(x)$.

An example

Example

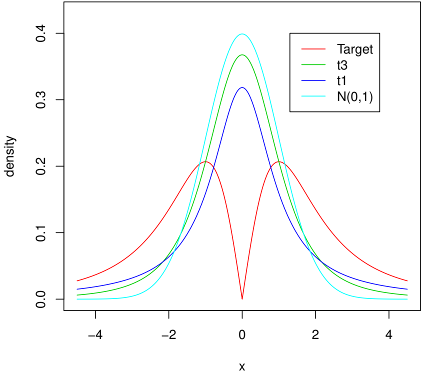
Compute $E_f(|X|)$ where $X \sim t_3$. Consider the following three sampling methods

(i) Directly sampling from t_3

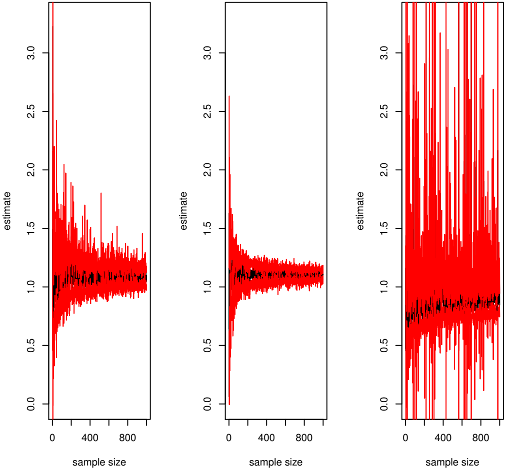
(ii) Use t_1 as the proposed sampling

(iii) Use $N(0,1)$ as the proposed sampling

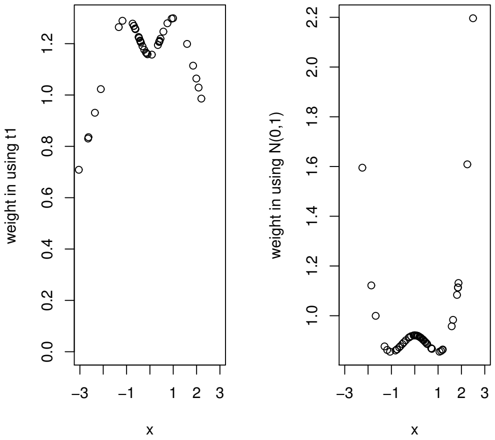
An example



An example


estimate
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
400
800
sample size
estimate
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
400
800
sample size
estimate
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
400
800
sample size

An example


weight in using t1
0.0
0.2
0.4
0.6
0.8
1.0
1.2
−3
−1
1
2
3
x
weight in using N(0,1)
1.0
1.2
1.4
1.6
1.8
2.0
2.2
−3
−1
1
2
3
x

Self-normalized importance sampling

Consider the estimator:

$$\hat{\mu} = \frac{\sum_{i=1}^{n} w(x_i)h(x_i)}{\sum_{i=1}^{n} w(x_i)}.$$

Properties: consistency, biased but asymptotically unbiased

Markov Chain Monte Carlo

Construct a Markov chain with stationary probability being the desired posterior distribution $\pi(\theta|X)$. Sample from such a Markov chain $\theta_1, \theta_2, \ldots, \theta_n$ that are converged. Then estimate $E(h(\theta)|X)$ using

$$\frac{1}{n}\sum_{i=1}^{n} h(\theta_i).$$

Ergodic Theorem tells us for $\theta_1, \theta_2, \ldots, \theta_n$ from a Markov chain that is aperiodic, irreducible, and positive recurrent, with probability 1,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} h(\theta_i) = E(h(\theta)|X).$$

Definitions and Concepts

Definition

A *transition kernel* is a function P defined on $\chi \times B(\chi)$ such that

(i) $\forall x \in \chi$, $P(x, \cdot)$ is a probability measure;

(ii) $\forall A \in B(\chi)$, $P(\cdot, A)$ is measurable.

- When χ is discrete, the transition kernel is a transition matrix P with elements $P_{xy} = P(X_n = y | X_{n-1} = x), x, y \in \chi)$
- When χ is continuous, the kernel also denotes the conditional density $P(x, x')$ of the transition $P(x, \cdot)$; that is, $P(X \in A|x) = \int_A P(x, x')d(x')$.

Example

The president of the United States tells a person A his or her intention to run or not to run in the next election. Then the person A relays the news to B, who in turn relays the message to another person C, and so forth. Assume that there is a probability a that a person will change the answer from yes to no when transmitting it to the next person and a probability b that he or she will change it from no to yes. The transition matrix is

$$P = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}.$$

The initial state represents the president's choice.

Definition

Given a transition kernel P, a sequence $X_0, X_1, \ldots, X_n, \ldots$ of random variables is a *Markov chain*, denoted by (X_n), if, for any t, the conditional distribution of X_t given $x_{t-1}, x_{t-2}, \ldots, x_0$ is the same as the distribution of X_t given x_{t-1}; that is,

$$P(X_{k+1} \in A | x_0, x_1, \ldots, x_k) \quad = \quad P(X_{k+1} \in A | x_k).$$

Markov Chain

A random process where all information about the future is contained in the present state. Suppose that we generate a sequence of random variables, $\{X_0, X_1, \ldots\}$ such that at each time t, the next state X_{t+1} is sampled from a distribution $P(X_{t+1}|X_t)$. That is, given X_t, the next state X_{t+1} does not depend further on the history of the chain $\{X_0, X_1, \ldots, X_{t-1}\}$, this property is known as "Markov property". The sequence is called a *Markov chain*.

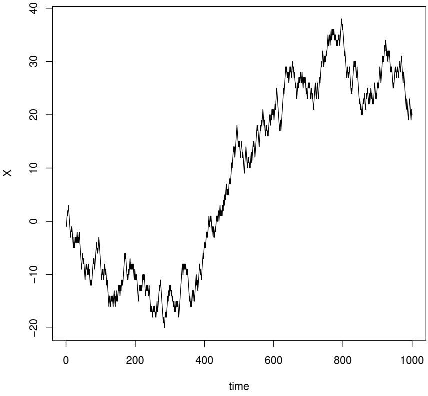
Definition

A sequence of random variables (X_n) is a *random walk* if it satisfies

$$X_{n+1} = X_n + \epsilon_n,$$

where ϵ_n is generated independent of $X_n, X_{n-1}, \ldots$. If the distribution of the ϵ_n is symmetric about zero, the sequence is called a *symmetric random walk*.

A simple random walk



Rational

If it is hard to generate an iid sample from the distribution $\pi(\theta|X)$ in Monte Carlo approach, we may look to generate a sequence from a Markov chain with limiting distribution π.

A Monte Carlo Markov Chain (MCMC) strategy

construct a Markov chain with stationary probability being our desired posterior distribution $\pi(\theta|X)$. Sample from such a Markov chain $\theta_1, \theta_2, \ldots, \theta_n$ that are converged. Then estimate $E(h(\theta)|X)$ using

$$\frac{1}{n}\sum_{i=1}^{n} h(\theta_i).$$

Ergodic Theorem tells us for $\theta_1, \theta_2, \ldots, \theta_n$ from a Markov chain that is aperiodic, irreducible, and positive recurrent, with probability 1,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} h(\theta_i) = E(h(\theta)|X).$$

Metropolis-Hastings algorithm

Suppose we would like to generate samples from a target density f. At each iteration t, do the following steps.

Step 1: Sample $\theta^* \sim q(\theta^*|\theta^{(t)})$, where $q(\cdot)$ is a **proposal distribution** and it is chosen so that $q(\theta^*|\theta^{(t)})$ is easy to sample from.

Step 2: With probability

$$\alpha(\theta^*|\theta^{(t)}) = \min\{1, \frac{f(\theta^*)q(\theta^{(t)}|\theta^*)}{f(\theta^{(t)})q(\theta^*|\theta^{(t)})}\},$$

set $\theta^{(t+1)} = \theta^*$ else set $\theta^{(t+1)} = \theta^{(t)}$.

Thus, in this way we generate a sequence of simulated values $\theta^{(1)}, \theta^{(2)}, \ldots,$ and this sequence converges to a random variable that has the distribution f.

Two special cases

- **Random walk**: $q(x, y) = q^*(y - x)$ for some distribution q^*.
- **Independence chain**: $q(x, y) = q(y)$. Similar to rejection sampling and the scale factor $q(x)/q(y)$ instead if $1/c$ and the rejected points are retained.

Convergence

The convergence requires the following conditions:

- Detailed balance
- Irreducible
- Aperiodic
- Positive recurrent

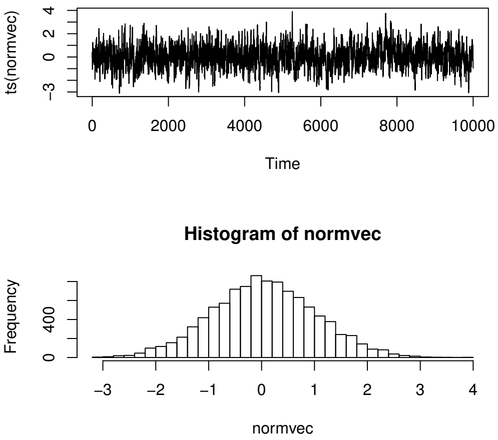
Random walk

Example

The goal is to generate samples from an $N(0,1)$. The proposed distribution is the uniform distribution.

```
norm = function (n, alpha){
        vec = vector("numeric", n)
        vec[1] = 0
        for (i in 2:n) {
                y = runif(1, -alpha+vec[i-1], alpha+vec[i-1])
                aprob = min(1, dnorm(y)/dnorm(vec[i-1]))
                u = runif(1)
                if (u < aprob)
                    vec[i] = y
                else
                    vec[i] = vec[i-1]
        }
        return(vec)
}
normvec<-norm(10000,1)
par(mfrow=c(2,1))
plot(ts(normvec))
hist(normvec,30)
```

Random walk MH



Example of MH

Example

Consider generating samples from a standard cauchy distribution using the normal distribution with the standard deviation being 2 as the proposal distribution in an MH algorithm. Plot histogram the 10,000 samples chosen after after throwing away the first 500 samples and compare the histogram with the true pdf.

Independence chain

Example

The goal is to generate samples from an $Gamma(\alpha, \beta)$. The proposed distribution is $N(\alpha/\beta, \alpha/\beta^2)$

```
gamm = function (n, alpha, beta)
{
   mu = alpha/beta
   sigma = sqrt(alpha/(beta^2))
   vec = vector("numeric", n)
   vec[1] = alpha/beta
   for (i in 2:n) {
   y <- rnorm(1, mu, sigma)
   aprob <- min(1, (dgamma(y, alpha, beta)/dgamma(vec[i-1],alpha, beta))
    /(dnorm(y, mu, sigma)/dnorm(vec[i-1],mu, sigma)))
      u <- runif(1)
      if (u < aprob)
          vec[i] = y
      else
          vec[i] = vec[i-1]
    }
    return(vec)
}
```

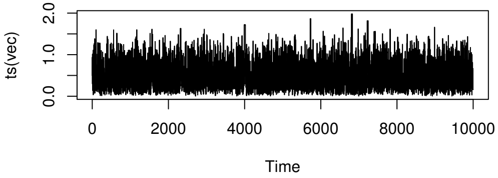
Independence MH

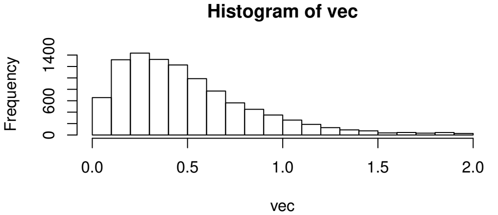
```
vec<-gamm(10000,2,4)
par(mfrow=c(2,1))
plot(ts(vec))
hist(vec,20)

vec<-gamm(10000,1,3)
par(mfrow=c(2,1))
plot(ts(vec))
hist(vec,20)

vec<-gamm(10000,4,1)
par(mfrow=c(2,1))
plot(ts(vec))
hist(vec,20)
```

Independence MH





Independence MH

Example

Consider generating samples from a standard cauchy distribution using the standard normal distribution with the standard deviation being 2 as the proposal distribution in an MH algorithm. Plot histogram the 10,000 samples chosen after after throwing away the first 500 samples and compare the histogram with the true pdf.

Gibbs sampling

Consider the parameter vector of interest $\theta = (\theta_1, \ldots, \theta_p)$. We have a joint distribution of $\theta_1, \ldots, \theta_p$. We wish to generate samples from a posterior distribution. The idea behind Gibbs sampling is that we can set up a Markov chain simulation algorithm from the joint posterior distribution by successfully simulating individual parameters from the set of p conditional distributions.

Two-stage Gibbs sampler

If two random variables X and Y have joint density $f(x, y)$ with the corresponding conditional densities $f_{Y|X}$ and $f_{X|Y}$, the two-stage Gibbs sampler generates a Markov chain (X_t, Y_t) according to the following steps: Take $X_0 = x_0$
For $t = 1, 2, \ldots,$ generate

1. $Y_t \sim f_{Y|X}(\cdot|x_{t-1})$;
2. $X_t \sim f_{X|Y}(\cdot|y_t)$.

An example

Example

Consider the bivariate normal model

$$(X, Y) \sim N_2(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}).$$

Then for given x_t, Gibbs sampler generates

$$Y_{t+1}|x_t \sim N(\rho x_t, 1 - \rho^2),$$

$$X_{t+1}|y_{t+1} \sim N(\rho y_{t+1}, 1 - \rho^2).$$

An example

Example

The use of Gibbs sampler in hierarchical model Consider the pair of the distributions

$$X|\theta \sim Bin(n, \theta), \theta \sim Beta(a, b).$$

The joint distribution is

$$f(x, \theta) = \binom{n}{x} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{x+a-1}(1-\theta)^{n-x+b-1}.$$

We have the conditional distributions,

$$X|\theta \sim Bin(n, \theta), \theta|x \sim Beta(x+1, n-x+b).$$

General Gibbs sampler

Let $[\theta_p|data]$ be the joint posterior distribution of θ. Define the set of conditional distributions

$$[\theta_1|\theta_2, \ldots, \theta_p, data]$$

$$[\theta_2|\theta_1, \ldots, \theta_p, data]$$

$$\vdots$$

$$[\theta_p|\theta_2, \ldots, \theta_{p-1}, data]$$

$[X|Y, Z]$ represents the distribution of X condition of the random variables Y and Z.

Gibbs sampling procedure

Gibbs sampling obtains samples in the following way, for
$t = 0, \dots,,$

$$
\begin{aligned}
\theta_1^{(t+1)} &\sim [\theta_1 | \theta_2^{(t)}, \theta_3^{(t)}, \dots, \theta_p^{(t)}, data] \\
\theta_2^{(t+1)} &\sim [\theta_2 | \theta_1^{(t+1)}, \theta_3^{(t)}, \dots, \theta_p^{(t)}, data] \\
\theta_3^{(t+1)} &\sim [\theta_3 | \theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_4^{(t)}, \dots, \theta_p^{(t)}, data] \\
\vdots \quad & \quad \vdots \quad \vdots \\
\theta_p^{(t+1)} &\sim [\theta_p | \theta_1^{(t+1)}, \theta_2^{(t+1)}, \theta_3^{(t+1)}, \dots, \theta_{p-1}^{(t+1)}, data]
\end{aligned}
$$

An Example

Consider the data

t	94	16	63	126	5	31	1	1	2	10
y	5	1	5	14	3	19	1	1	4	22

The response y_i is the number of failure for a pump observed at time t_i in a nuclear plant. We wish to model the number of failures with a Poisson distribution with the expected number of failures being λ_i. Since t_i is different, we need to scale each λ_i by its observed time t_i. Thus the likelihood function is $\prod_{i=1}^{10} Poisson(\lambda_i t_i)$.

Prior choices
Prior for λ_i: $Gamma(\alpha, \beta)$ with $\alpha = 1.5$ where β has the prior $Gamma(\gamma, \delta)$ with $\gamma = 0.01$ and $\delta = 1$.

Our posterior is
$\pi(\lambda_1, \ldots, \lambda_{10}, \beta | y, t) \propto$
$\Big(\prod_{i=1}^{10} Poisson(\lambda_i t_i) \times Gamma(\alpha, \beta) \Big) \times Gamma(\gamma, \delta)$

Full conditionals

The full conditionals are

$$\pi(\lambda_i|\lambda_{-i}, \beta, y, t) \propto \lambda_i^{y_i+\alpha-1} e^{-(t_i+\beta)\lambda_i}$$

$$\pi(\beta|\lambda_1, \ldots, \lambda_{10}, y, t) \propto e^{-\beta(\delta+\sum_{i=1}^{10} \lambda_i)} \beta^{10\alpha+\gamma-1}$$

Steps of the Gibbs sampling:

Step 1. Choose the initial value for $\beta^{(0)}$.

Step 2. Based on the initial value of β, draw $(\lambda_1^{(1)}, \lambda_2^{(1)}, \dots, \lambda_{10}^{(1)})$ from its full conditional distribution.

```
lambda.draw <- function(alpha, beta, y, t)
 {
  rgamma(length(y), y + alpha, t + beta)
 }
```

Step 3. Based on $(\lambda_1^{(1)}, \lambda_2^{(1)}, \dots, \lambda_{10}^{(1)})$, draw $\beta^{(1)}$ from its full conditional distribution.

```
beta.draw <- function(alpha, gamma, delta, lambda, y)
{
  rgamma(1, length(y) * alpha + gamma, delta + sum(lamb
 }
```

Step 4. Repeat Steps 2 and 3 until we obtain the desired number of M draws.

Burn-in and thinning

- Burn-in: throw out the beginning part of the Markov chain
- Thinning: keep only every kth point after the burn-in period

Results

```
> posterior = gibbs(n.sims = 10000, beta.start = 1, alpha = 1.5,
+                      gamma = 0.01, delta = 1, y = y, t = t)
> round(colMeans(posterior$lambda.draws),3)
 [1] 0.068 0.139 0.100 0.121 0.649 0.624 0.868 0.875 1.406 1.966
> mean(posterior$beta.draws)
[1] 1.979316
> round(apply(posterior$lambda.draws, 2, sd),3)
 [1] 0.027 0.087 0.039 0.031 0.311 0.138 0.595 0.587 0.638 0.415
> round(sd(posterior$beta.draws),3)
[1] 0.621
```