Metropolis-Hastings算法的理论与应用

2022年概率基础课程期末报告

学号：2022103795

姓名：吕明倩

日期：2022年12月25日

算法原理简介

MCMC 方法介绍:

在贝叶斯统计中，通常希望得到一些后验量，这些后验量可以写成某函数 $f(x)$ 关于 $\pi(x)$ 的期望 $E_\pi f(X)=\int_X f(x)\pi(x)dx$，这里随机变量 X 具有密度函数 $\pi(x)$，如果密度函数为 $\pi(x)$ 的随机数 $X_1,X_2,\ldots,X_n$ 很容易通过 $Monte\ Carlo$ 方法产生，则根据大数定律，$E_\pi f(X)$ 可以直接通过 $\hat{E}_\pi f(X)=\frac{1}{n}\sum\limits_{i=1}^{n} f(X_i)$ 估计得到。但在实际问题中，后验分布 $\pi(x)$ 往往是复杂的、高维的、非标准形式的分布，难以直接抽样，所以提出 $MCMC$ 算法。

$Markov\ Chain\ Monte\ Carlo$ 就是在贝叶斯理论框架下发展起来的蒙特卡罗采样方法。MCMC 算法的基本思想是通过构建一个平稳分布为 $\pi(x)$ 的 $Markov$ 链 $X^{(1)},X^{(2)},\ldots,X^{(n)}$，在一定的条件下可得:

$$\hat{E}_\pi f=\frac{1}{n}\sum_{i=1}^{n} f(X^{(i)})\xrightarrow{P} E_\pi f=\int_X f(x)\pi(x)dx$$

Metropolis-Hastings 方法:

对于 MH 算法可以采用以下步骤获得来自后验分布 $\pi(\theta)$ 的样本:

- 确定一个初始值 θ_0
- 从提案分布 $g(\cdot|\theta)$ 中生成新候选值 θ^*
- 计算 M-H 比率 $R(\theta,\theta^*)=\frac{\pi(\theta^*)g(\theta|\theta^*)}{\pi(\theta)g(\theta^*|\theta)}$，接受概率 $\alpha(\theta,\theta^*)=\min\{R(\theta,\theta^*),1\}$
- 产生随机数 $u\sim U(0,1)$，令

$$\theta^{(t+1)}=\begin{cases}\theta^{(*)}, u\le\alpha(\theta,\theta^*)\\ \theta^{(t)}, u>\alpha(\theta,\theta^*)\end{cases}$$

- 重复步骤 2 至 4 得到采样序列 $\theta^{(1)},\ldots,\theta^{(n)}$，该序列可以看作来自后验分布 $\pi(\theta)$ 的样本。

Problem1

证明：M-H 算法最终是从函数 $\pi(\theta)$ 中提取样本。将采样序列看作一个 $Markov$ 链，则问题转化为证明 $\pi(\theta)$ 是 $Markov$ 链 $\{\theta^{(t)}\}$ 的平稳分布。可以证明 $Markov$ 链 $\{\theta^{(t)}\}$ 满足细致平衡条件 ($Detailed\ Balance\ Condition$)：$\pi(\theta_1)p(\theta_1,\theta_2)=\pi(\theta_2)p(\theta_2,\theta_1)$。

假设 $\theta^{(t)}=\theta_1$，$\theta^{(t+1)}=\theta_2$。若 $\theta_2=\theta_1$，即拒绝了新候选值，$\theta^{(t+1)}=\theta^{(t)}$，则上式 ($DBC$) 显然成立；若 $\theta_2\neq\theta_1$，即接受了新候选值，马氏链发生状态跳转，此时转移概率：

$$
\begin{aligned}
p\left(\theta_1,\theta_2\right) &= P\left(\theta^{(t+1)}=\theta_2|\theta^{(t)}=\theta_1\right)\\
&= P\left(\theta^{(t+1)}=\theta_2,u\leq\alpha\left(\theta^{(t)},\theta^{(t+1)}\right)|\theta^{(t)}=\theta_1\right)\\
&= P\left(\theta^{(t+1)}=\theta_2|\theta^{(t)}=\theta_1\right)P\left(u\leq\alpha\left(\theta^{(t)},\theta^{(t+1)}\right)|\theta^{(t+1)}=\theta_2,\theta^{(t)}=\theta_1\right)\\
&= g\left(\theta_2|\theta_1\right)P\left(u\leq\alpha\left(\theta_1,\theta_2\right)\right)\\
&= g\left(\theta_2|\theta_1\right)\min\{\alpha\left(\theta_1,\theta_2\right),1\}\\
&= \min\{g\left(\theta_2|\theta_1\right)\tfrac{\pi(\theta_2)g(\theta_1|\theta_2)}{\pi(\theta_1)g(\theta_2|\theta_1)},g\left(\theta_2|\theta_1\right)\tfrac{\pi(\theta_1)}{\pi(\theta_1)}\}\\
&= \tfrac{\min\{\pi(\theta_1)g(\theta_2|\theta_1),\pi(\theta_2)g(\theta_1|\theta_2)\}}{\pi(\theta_1)}
\end{aligned}
\tag{1}
$$

注意到转移概率矩阵不依赖时间 t，因此 $Markov$ 链 $\{\theta^{(t)}\}$ 是时间齐次的。同时根据对称性可得：$p\left(\theta_2,\theta_1\right)=\frac{\min\{\pi(\theta_1)g(\theta_2|\theta_1),\pi(\theta_2)g(\theta_1|\theta_2)\}}{\pi(\theta_2)}$。因此，$\pi(\theta_1)p(\theta_1,\theta_2)=\pi(\theta_2)p(\theta_2,\theta_2)$ 满足细致平衡条件。

所以，$\pi(\theta)$ 是 $Markov$ 链 $\{\theta^{(t)}\}$ 的平稳分布，并且当 $\{\theta^{(t)}\}$ 满足非周期、不可约条件后，$\{\theta^{(t)}\}$ 有唯一的极限平稳分布 $\pi(\theta)$。

遍历性定理： 若 $\{\theta^{(t)}\}$ 是非周期、不可约的 $Markov$ 链，并且存在唯一的平稳分布 $\pi(\theta)$，则 $\theta^{(t)}\xrightarrow{d}\pi(\theta)$，即 $\theta^{(t)}$ 依分布收敛到 $\pi(\theta)$。

所以根据遍历性定理，若 $E_\pi f(X)$ 存在，则状态空间平均期望可由时间上的平均期望近似，即存在 $T>0$，当 $t\geq T$ 时，$\theta^{(t)}\xrightarrow{d}\pi(\theta)$，因此，$\frac{1}{L}\sum\limits_{t=T}^{T+L}f(\theta^{(t)})\xrightarrow{a.s.}E_\pi f(X)$。

此外，若采用对称的提案分布，例如 $g\left(\theta^*\mid\theta\right)=\frac{1}{\sqrt{2\pi}\psi}e^{-(\theta^*-\theta)^2/2\psi^2}$，此时 $g(\theta^*|\theta)=g(\theta|\theta^*)$，所以接受率可以化简为 $\alpha(\theta,\theta^*)=\min\{\frac{\pi(\theta^*)}{\pi(\theta)},1\}$。

Problem2

Part1

样本 $X=(x_1,\dots x_n)$，$x_i \sim N\left(\mu=200,\phi=\frac{1}{2}\right)$，其中 $\phi=\frac{1}{\sigma^2}$，先验分布 $p(\mu,\phi)\propto\frac{1}{\phi}$，后验分布 $p(\mu,\phi\mid X)\propto\phi^{n/2-1}\exp\left(-\frac{\phi}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right)$。选择正态分布作为两个参数 μ,ϕ 的提案分布，即 $\mu^*\sim N(\mu,1)$，$\phi^*\sim N(\phi,0.1)$，*Metropolis-Hastings* 抽样函数的伪代码如下所示：

算法 1 *Metropolis-Hastings*

Input: 样本 x, 上一次采样值 μ,ϕ
Output: 新采样值 μ^*,ϕ^*

function METROPOLIS$(x,\mu,\phi,n=100)$
  $\mu^*=rnorm(1,\mu,1)$ //从提案分布 $N(\mu,1)$ 中抽取新候选值 μ^*
  $R(\mu,\mu^*)=\frac{exp\{-\frac{\phi}{2}\sum(x-\mu^*)^2\}}{exp\{-\frac{\phi}{2}\sum(x-\mu)^2\}}*\frac{dnorm(\mu,\mu^*,1)}{dnorm(\mu^*,\mu,1)}$
  if $runif(1)>R(\mu,\mu^*)$ **then**
    $\mu^*=\mu$
  end if
  $\phi^*=rnorm(1,\phi,0.1)$ //从提案分布 $N(\phi,0.1)$ 中抽取新候选值 ϕ^*
  $R(\phi,\phi^*)=(\frac{\phi^*}{\phi})^{(\frac{n}{2}-1)}exp(-\frac{\phi^*-\phi}{2}\sum(x-\mu^*)^2)*\frac{dnorm(\phi,\phi^*,0.1)}{dnorm(\phi^*,\phi,0.1)}$
  if $runif(1)>R(\phi,\phi^*)$ **then**
    $\phi^*=\phi$
  end if
end function

通过循环调用函数 *metropolis* 可以获得参数 μ,ϕ 的采样序列。

采样 2 *Metropolis-Hastings*

$para=matrix(0,N,2)$ $//para$ 存放每次采样结果
$para(1,:)=c(\mu_0,\phi_0)$
for i *in* 2:N **do**
  $para(i,:)=metropolis(x,para(i-1,1),para(i-1,2))$
end for

关于 *Metropolis-Hastings* 算法，提案分布的选取对于抽样结果至关重要，选择一个适合的提案分布可以使 *Markov* 链在短时间内达到平稳分布，但如果选择的提案分布欠佳，*Markov* 链恐难以收敛。

Gibbs 抽样是一种单元素 *Metropolis-Hastings* 算法的特殊情况，当参数的满条件分布具有显式形式时，可以直接从满条件分布中获得参数的采样序列。对于 *Problem*2 中的后验分布，可以求得各参数的显式满条件分布，考虑采用 *Metropolis-Hastings-Gibbs* 算法

实现采样，伪代码如算法 3 所示。其中，参数 μ, ϕ 的满条件分布依次为：

$$\mu^* \sim N(\frac{\sum x}{n}, \frac{1}{\sqrt{n\phi}}), \quad \phi^* \sim Ga(\frac{n}{2}, \frac{\sum(x-\mu)^2}{2})$$

算法 3 *Metropolis-Hastings-Gibbs*

```
//从满条件分布中直接抽样
for i in 2:N do
    para(i,1) = rnorm(1, Σx/n, 1/sqrt(n*para(i-1,2)))
    para(i,2) = rgamma(1, n/2, Σ(x-para(i,1))^2/2)
end for
```

Part2

初始化 $\mu_0 = 0$ 和 $\phi_0 = 5$，采集样本总数 $N = 3000$，预烧值 $m = 500$。

首先运用 *Metropolis-Hastings* 算法采样，获得参数 μ 的完整轨迹以及预烧后轨迹（图 1），参数 ϕ 的完整轨迹以及预烧后轨迹（图 2）。

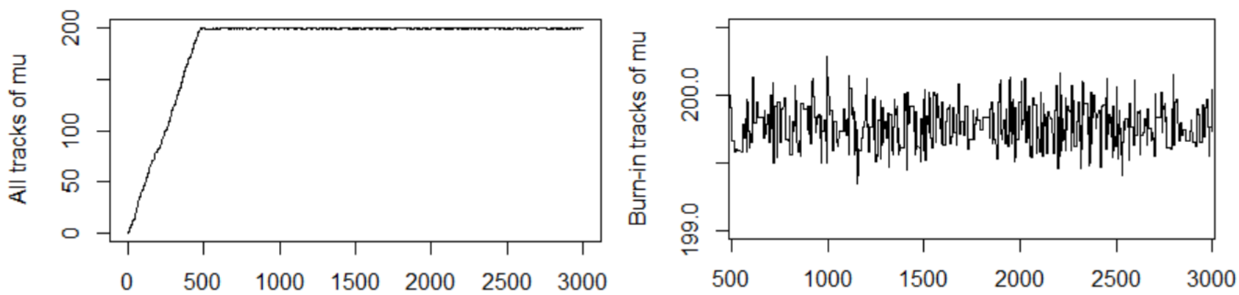


图 1：MH 完整（左）和预烧后（右）的 μ 的采样轨迹图

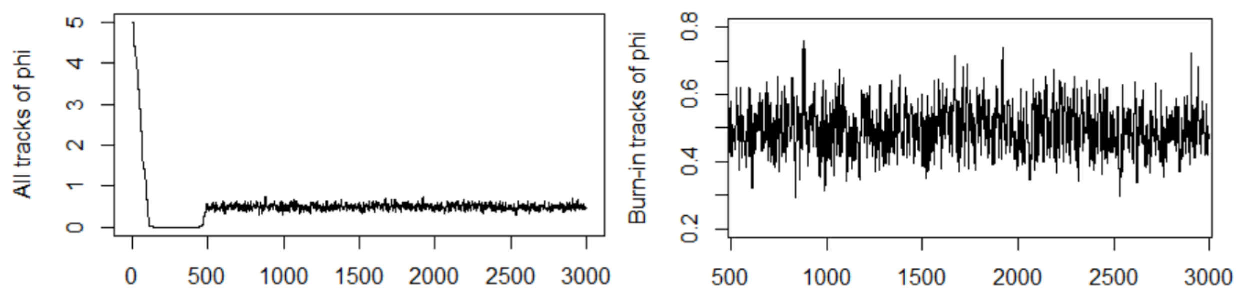


图 2：MH 完整（左）和预烧后（右）的 ϕ 的采样轨迹图

可以观察到，当预烧 500 次后，参数 μ 和 ϕ 的 *Markov* 链收敛到平稳分布，此时预烧后参数的边缘分布直方图如图 3 显示，参数 μ 和 ϕ 的采样均值分别为 199.782 和 0.498，候选值的接受率分别为 0.255 和 0.541。

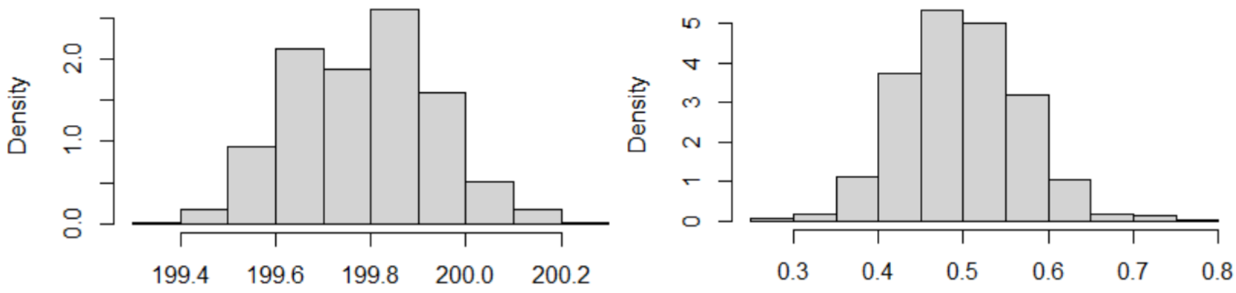


图 3：MH 预烧后参数的边缘分布直方图（左：μ，右：ϕ）

其次考虑运用 *Metropolis-Hastings-Gibbs* 算法采样，获得参数 μ 的完整轨迹以及预烧后轨迹（图 4），参数 ϕ 的完整轨迹以及预烧后轨迹（图 5）。

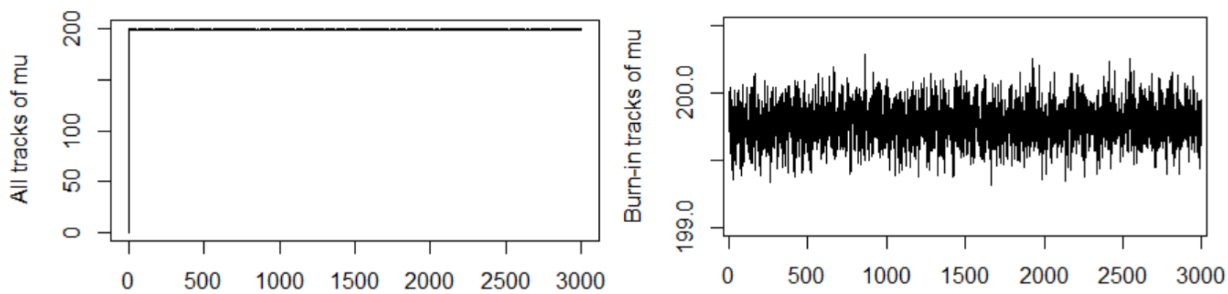


图 4：MHG 完整（左）和预烧后（右）的 μ 的采样轨迹图

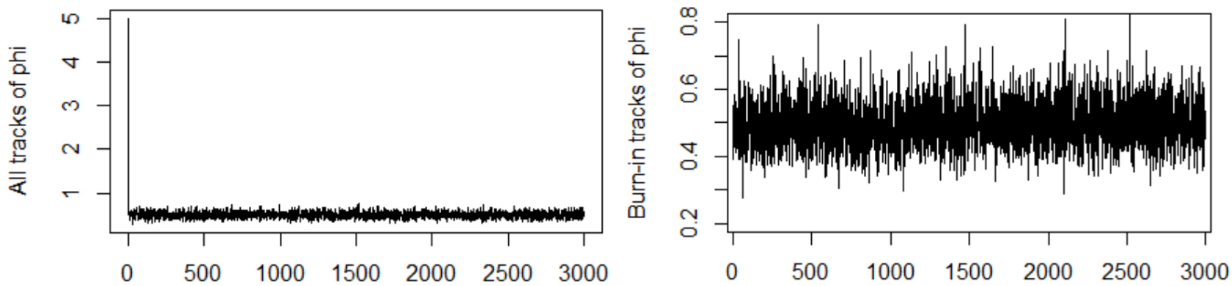
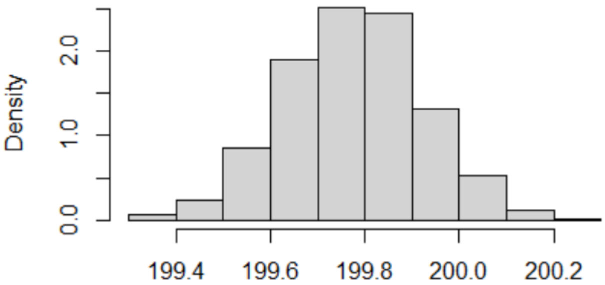


图 5：MHG 完整（左）和预烧后（右）的 ϕ 的采样轨迹图

与 *Metropolis-Hastings* 算法相比，*Metropolis-Hastings-Gibbs* 算法每一步都会更新采样值，可以更快达到平稳分布，此时预烧后参数的边缘分布直方图如图 6 显示，参数 μ 和 ϕ 的采样均值分别为 199.777 和 0.496。



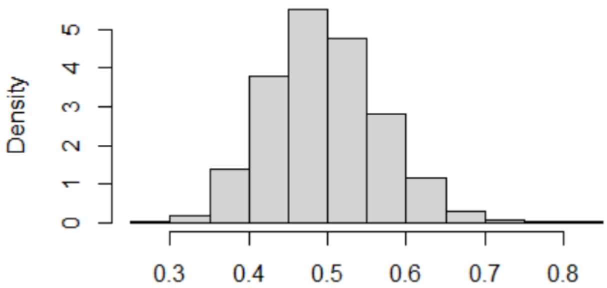


图 6：MHG 预烧后参数的边缘分布直方图（左：μ，右：ϕ）

R 语言代码

```r
N = 3000  #总抽样次数
n = 100   #样本个数
set.seed(1234)
x = rnorm(n, 200, sqrt(2))  #从给定分布中抽取100个样本

#初始值
mu0 = 0
phi0 = 5

#输出结果
para = matrix(0,N,2)
para[1, ]=c(mu0, phi0)

#Metropolis-Hastings算法
metropolis<-function(x,mu,phi,n=100){
  mustar = rnorm(1, mu, 1)
  rat = exp((-phi/2)*(n*(mustar^2-mu^2)-2*sum(x)*(mustar-mu))) *
        (dnorm(mu,mustar,1)/dnorm(mustar,mu,1))
  if (runif(1) > rat)  mustar = mu
  phistar = rnorm(1, phi, 0.1)
  rat = ((phistar/phi)^(n/2-1)*exp((-(phistar-phi)/2)*sum((x-mustar)^2))) *
        (dnorm(phi,phistar,0.1)/dnorm(phistar,phi,0.1))
  if (runif(1) > rat)  phistar = phi
  return (c(mustar,phistar))
}
for (i in 2:N){
  para[i, ] = metropolis(x,para[i-1,1],para[i-1,2])
}

#预烧次数
m = 500
```

```r
#图形展示
plot(para[,1], type='l', xlab = NULL, ylab = "All tracks of mu")
plot(para[,1], type='l', xlim = c(580,2920), ylim = c(199,200.5),
     ylab = "Burn-in tracks of mu", xaxt = 'n')
axis(1,c(500,1000,1500,2000,2500,3000))
plot(para[,2], type='l', xlab = NULL, ylab = "All tracks of phi")
plot(para[,2], type='l', xlim = c(580,2920), ylim = c(0.2,0.8),
     ylab = "Burn-in tracks of phi", xaxt = 'n')
axis(1,c(500,1000,1500,2000,2500,3000))

hist(para[m:N,1], prob=TRUE, xlab = "mu")
hist(para[m:N,2], prob=TRUE, xlab = "phi")
acceptmu = length(unique(para[,1]))/N  #mu的接受率=0.255
acceptphi = length(unique(para[,2]))/N #phi的接受率=0.541
mean(para[m:N,1])  #mu的采样均值=199.782
mean(para[m:N,2])  #phi的采样均值=0.498


#Metropolis-Hastings-Gibbs
for (i in 2:N){
  para[i,1] = rnorm(1, sum(x)/n, 1/sqrt(n*para[i-1,2]))
  para[i,2] = rgamma(1, n/2, sum((x-para[i,1])^2)/2)
}

#预烧次数
m = 10
```

```r
#图形展示
plot(para[,1], type='l', xlab = NULL, ylab = "All tracks of mu")
plot(para[2:3000,1], type='l', xlim = c(80,2920), ylim = c(199,200.5),
     ylab = "Burn-in tracks of mu", xaxt = 'n')
axis(1,c(0,500,1000,1500,2000,2500,3000))
plot(para[,2], type='l', xlab = NULL, ylab = "All tracks of phi")
plot(para[2:3000,2], type='l', xlim = c(80,2920), ylim = c(0.2,0.8),
     ylab = "Burn-in tracks of phi", xaxt = 'n')
axis(1,c(0,500,1000,1500,2000,2500,3000))

hist(para[2:N,1], prob=TRUE, xlab = "mu")
hist(para[2:N,2], prob=TRUE, xlab = "phi")
mean(para[2:N,1])  #mu的采样均值=199.777
mean(para[2:N,2])  #phi的采样均值=0.496
```