

---

# MCMC 讲义

@NWU

---

强喆

[zhe.qng@nwu.edu.cn](mailto:zhe.qng@nwu.edu.cn)

2019 年 3 月 20 日



# 目录

<b>1 马尔可夫链</b>	<b>1</b>
1.1 马尔可夫链的基本概念	1
1.2 马尔可夫链的可逆性和细致平衡条件	1
1.3 马尔可夫链蒙特卡罗模拟的原理	3
1.4 作业	3
<b>2 Borel-Cantelli Lemmas</b>	<b>4</b>
<b>3 Weak Laws of Large Numbers</b>	<b>10</b>
3.1 $L^2$ Weak Laws	10
3.2 The general form weak laws	10
<b>4 Strong Law of Large Numbers</b>	<b>13</b>
<b>5 Central Limit Theorems</b>	<b>15</b>
5.1 Weak convergence	15
5.2 Central Limit Theorems	15
<b>6 Accept-Reject Methods</b>	<b>16</b>
6.1 The Fundamental Theorem of Simulation	16
6.2 The Accept-Reject Algorithm	19
6.3 The Kiss Generator	20
<b>7 Introduction to Bayesian Computation</b>	<b>23</b>
7.1 Rejection sampling	23
<b>8 Metropolis Hastings 算法</b>	<b>25</b>
8.1 A generic Metropolis Hastings algorithm	25
8.2 The independent M-H algorithm	28
<b>9 Gibbs sampling</b>	<b>31</b>

## 目录

9.1	The two-stage Gibbs sampler . . . . .	31
9.2	Missing data and latent variables . . . . .	38
9.3	gibbs sampling on mixture model . . . . .	39
9.4	generate random number from truncated normal distribution . . . . .	45
9.4.1	The Inverse Transform . . . . .	45
9.4.2	generate random number from truncated normal distribution—by the inverse transform . . . . .	49
9.5	The slice sampler . . . . .	49
9.5.1	The fundamental theorem . . . . .	49
9.6	Back to the Gibbs Sampler . . . . .	54
9.7	The Hammersley-Clifford Theorem . . . . .	54
9.8	Examples . . . . .	55
9.9	importance sampling . . . . .	61
<b>10</b>	<b>chapter 11 Basics of Markov chain simulation</b>	<b>63</b>
10.1	11.1 Gibbs sampler/ alternating conditional sampling . . . . .	64
10.2	11.2 Metropolis and M-H 算法 . . . . .	64
10.2.1	Metropolis 算法 . . . . .	64
10.2.2	Metropolis-Hastings 算法 . . . . .	65
<b>11</b>	<b>Monitoring Convergence to the Stationary Distribution</b>	<b>66</b>
11.1	推断、评价收敛 . . . . .	66
11.1.1	simulation draws 的有效数目 . . . . .	68
11.2	Graphical diagnoses . . . . .	69
11.3	Nonparametric tests of stationarity . . . . .	70
11.4	A missing Mass . . . . .	70
11.5	Geweke.diag function . . . . .	71
11.6	Kolmoforov-Smirnov statistic . . . . .	71
11.7	Two-sample Kolmoforov-Smirnov test . . . . .	72
11.8	summary—convergence diagnosing . . . . .	73

# 1 马尔可夫链

## 1.1 马尔可夫链的基本概念

### 定义 (马尔可夫链)

称随机变量序列是马尔可夫链，如果这个序列满足：给定  $X_1, \dots, X_n$  时  $X_{n+1}$  的条件分布只依赖于  $X_n$ ，也就是：

$$P(X_{n+1}|X_1, X_2, \dots, X_n) = P(X_{n+1}|X_n)$$

由定义可以看出，马尔可夫链是由以下两个条件完全确定的：

- 初始分布；
- 转移概率分布：  $P(X_{n+1}|X_n)$ ，这里记作  $K(X_n, X_{n+1})$

### 定义 (平稳马尔可夫链)

进一步，如果这个马尔可夫链对于任意的  $i, k$ ，有  $(X_{n+1}, \dots, X_{n+k})$  的分布不依赖于  $n$ ，则称这个链是平稳的马尔可夫链。

### 定义 (平稳性)

对于  $\sigma$ -有限的测度  $\pi$ ，称它关于转移核  $K(\cdot, \cdot)$  是不变的，如果：

$$\pi(B) = \int_{\mathcal{X}} K(x, B) \pi(dx), \quad \forall B \in \mathcal{B}(X).$$

这样，如果  $X_0 \sim \pi$ ，那么  $X_n \sim \pi$ ， $\forall n$ ，因而把不变分布也称为平衡分布或者平稳分布。

## 1.2 马尔可夫链的可逆性和细致平衡条件

### 定义 (可逆性)

称马尔可夫链是可逆的，如果向前和向后的转移律相同，也就是

$$P(X_{i+1}, X_{i+2}, \dots, X_{i+k}) = P(X_{i+k}, \dots, X_{i+2}, X_{i+1}), \quad \forall i, k$$

当可逆性只考虑在前后两步转移时，称为细致平衡。

## 1 马尔可夫链

### 定义 (细致平衡条件)

称马尔可夫链的转移核  $K$  满足细致平衡条件 (detailed balance), 如果存在函数  $\pi$  满足:

$$K(y, x)\pi(y) = K(x, y)\pi(x), \forall x, y \in \mathcal{X}$$

下面的定理给出细致平衡条件、可逆性和平稳分布之间的关系。

### 定理

如果马尔可夫链的转移核  $K$  满足细致平衡条件, 则

(i) 密度  $\pi$  就是平稳分布的密度;

(ii) 链是可逆的。

**证明** (i) 对样本空间  $\mathcal{X}$  中的可测集  $B$ , 由细致平衡条件可知:

$$\begin{aligned} \int_{\mathcal{X}} K(y, B)\pi(y)dy &= \int_{\mathcal{X}} \int_B K(y, x)\pi(y)dx dy \\ &= \int_{\mathcal{X}} \int_B K(x, y)\pi(x)dx dy \\ &= \int_B \pi(x)dx \end{aligned}$$

因而, 不变分布存在, 其密度就是  $\pi$ 。

(ii) 如果链满足细节平衡条件, 那么对任意的  $i, k$ , 有

$$\begin{aligned} \pi(X_{i+1})K(X_{i+1}, X_{i+2}) &= \pi(X_{i+2})K(X_{i+2}, X_{i+1}) \\ \implies \pi(X_{i+1})K(X_{i+1}, X_{i+2})K(X_{i+2}, X_{i+3}) \\ &= \pi(X_{i+2})K(X_{i+2}, X_{i+3})K(X_{i+2}, X_{i+1}) \\ &= \pi(X_{i+3})K(X_{i+3}, X_{i+2})K(X_{i+2}, X_{i+1}) \\ &\vdots \\ \implies \pi(X_{i+1})K(X_{i+1}, X_{i+2}) \cdots K(X_{i+k-1}, X_{i+k}) \\ &= \pi(X_{i+k})K(X_{i+k}, X_{i+k-1}) \cdots K(X_{i+2}, X_{i+1}) \end{aligned}$$

因而链是可逆的。

取上式中  $k = 1$ , 即可得出链满足细致平衡条件。

通常在设计马尔可夫链的模拟算法时, 用细致平衡条件约束关系设计算法非常简单, 并且能够保证马尔可夫链的平稳分布存在, 而平稳分布存在又是算法收敛的必要条件。这是因为, 平稳马尔可夫链的平稳分布 (不变分布) 存在的充分不必要条件是, 马尔可夫链的可逆性或者要求细致平衡条件成立, 可以证明后二者是等价的。

## 1.3 马尔可夫链蒙特卡罗模拟的原理

MCMC 算法原理是用蒙特卡罗 (Monte Carlo, MC) 来模拟马尔可夫链。样本生成过程是：从  $X^{(t)}$  根据马尔可夫转移核生成  $X^{(t+1)}$ ，算法结束就得到一系列的样本  $(X^{(0)}, X^{(1)}, \dots, X^{(t)}, \dots, X^{(T)})$ 。因而这些样本实际上形成了一条马尔可夫链。由马尔可夫链的遍历性可以得到，对任何特定的函数  $h(x)$ ，有如下结论：

$$\frac{1}{T} \sum_{t=1}^T h(X^{(t)}) \rightarrow \int_{\mathcal{X}} h(x) \pi(x) dx$$

这意味这马尔可夫链的状态概率分布可以收敛到平稳分布。这一结果和经典的蒙特卡罗 (Monte Carlo, MC) 模拟是一致的：经典的蒙特卡罗模拟是利用独立同分布的随机变量  $X_1, X_2, \dots$ ，服从相同的分布  $\pi(x)$ ，由强大数定律 (the strong law of large numbers, SLLN)，以及满足控制收敛定理的某个连续函数  $h(x)$ ：

$$\frac{1}{n} \sum_{i=1}^n h(X_i) \rightarrow_{a.s.} \int_{\mathcal{X}} h(x) \pi(x) dx$$

这样通过蒙特卡罗模拟马尔可夫链产生的  $(X^{(0)}, \dots, X^{(t)}, \dots)$  理论上等同于用不变分布  $\pi$  产生的独立同分布 (independent identically distribution, iid) 的样本，这样马尔可夫链蒙特卡罗模拟产生的序列  $(X^{(t)})$  就可以作为 i.i.d 的样本使用。

实际上，MCMC 产生的样本是非独立的，但只要  $cov(X^{(t)}, X^{(t+k)})$  随着  $k$  的增大而减小， $(X^{(k)}, X^{(2*k)}, \dots, X^{(t*k)}, \dots)$  就是拟独立的样本，因而可以近似地替代 i.i.d 的样本来使用。

## 1.4 作业

- 强大数定律、弱大数定律是什么
- 安装 Rstudio [www.rstudio.com](http://www.rstudio.com), 简单学习 r 语言
- 产生随机数的算法

## 2 Borel-Cantelli Lemmas

Convergence in probability ( $X_n \xrightarrow{P} X$ , as  $n \rightarrow \infty$ ): for any  $\varepsilon > 0$

$$\lim_n P(|X_n - X| > \varepsilon) = 0$$

Convergence almost surely ( $X_n \xrightarrow{a.s.} X$ , as  $n \rightarrow \infty$ )

$$P(\lim X_n = X) = 1$$

or

$$P(w : \lim X_n(w) = X(w)) = 1$$

If  $A_n$  is a sequence of subsets of  $\Omega$ , we let

$$\begin{aligned} \limsup A_n &= \lim_{m \rightarrow \infty} \bigcup_{n=m}^{\infty} A_n = \{w \text{ that are in infinitely many } A_n\} \\ &= \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} A_n \\ &= (A_1 \cup A_2 \cup A_3 \cup \dots) \\ &\quad \cap (A_2 \cup A_3 \cup \dots) \\ &\quad \cap (A_3 \cup \dots) \\ &\quad \cap \dots \end{aligned}$$

the limit exists since the sequence is decreasing in  $m$  and let

$$\begin{aligned} \liminf A_n &= \lim_{m \rightarrow \infty} \bigcap_{n=m}^{\infty} A_n = \{w \text{ that are in all but finitely many } A_n\} \\ &= \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} A_n \\ &= (A_1 \cap A_2 \cap A_3 \cap \dots) \\ &\quad \cup (A_2 \cap A_3 \cap \dots) \\ &\quad \cup (A_3 \cap \dots) \\ &\quad \cup \dots \end{aligned}$$



- It is common to write  $\limsup A_n = \{w : w \in A_n i.o.\}$ , where i.o. stands for infinitely often. Let  $N(w) = \sum_{n=1}^{\infty} \mathbb{I}_{A_n}(w)$ , then  $w \in \{A_n i.o.\}$  means  $N(w) = \infty$ .
- $X_n \rightarrow X$  a.s. if and only if for all  $\varepsilon > 0$ ,  $P(|X_n - X| > \varepsilon i.o.) = 0$ .

*Proof.*

$$\begin{aligned}
X_n &\xrightarrow{a.s.} X \Leftrightarrow P(\lim_n X_n = X) = 1 \\
&\Leftrightarrow \forall \varepsilon > 0, P(\exists N > 0, \forall n > N, |X_n - X| < \varepsilon) = 1 \\
&\Leftrightarrow \forall \varepsilon > 0, P\left(\bigcup_{N=1}^{\infty} \bigcap_{n=N}^{\infty} |X_n - X| < \varepsilon\right) = 1 \\
&\Leftrightarrow \forall \varepsilon > 0, P(\liminf |X_n - X| < \varepsilon) = 1 \\
&\Leftrightarrow \forall \varepsilon > 0, P(\limsup |X_n - X| \geq \varepsilon) = 0 \\
&\Leftrightarrow \forall \varepsilon > 0, P(|X_n - X| \geq \varepsilon i.o.) = 0
\end{aligned}$$

□

### 定理 (Borel-Cantelli lemma)

if  $\sum_{n=1}^{\infty} P(A_n) < \infty$ , then

$$P(A_n i.o.) = 0$$

*Proof.* Fubini's theorem implies  $EN = \sum_k P(A_k) < \infty$ , so we must have  $N < \infty$  a.s. □

### 定理 (The second Borel-Cantelli lemma)

If the events  $A_n$  are independent then  $\sum P(A_n) = \infty$  if and only if  $P(A_n i.o.) = 1$ .

*Proof.*  $\Rightarrow$  Let  $M < N < \infty$ . Independence and  $1 - x \leq e^{-x}$  imply

$$\begin{aligned}
P(\cap_{n=M}^N A_n^c) &= \prod_{n=M}^N (1 - P(A_n)) \leq \prod_{n=M}^N \exp(-P(A_n)) \\
&= \exp\left(-\sum_{n=M}^N P(A_n)\right) \rightarrow 0 \text{ as } N \rightarrow \infty
\end{aligned}$$

So  $P(\cup_{n=M}^{\infty} A_n) = 1$  for all M, and since  $\cup_{n=M}^{\infty} A_n \downarrow \limsup A_n$  it follows that  $P(\limsup A_n) = 1$ .  $\Leftarrow$  This is proof by contradiction with Borel-Cantelli lemma. □

### 例子 (convergence in probability: Archer)

Suppose a person takes a bow and starts shooting arrows at a target. Let  $X_n$  be his score in n-th shot. Initially he will be very likely to score zeros, but as the time goes and his archery skill increases, he will become more and more likely to hit the bullseye and score 10

## 2 Borel-Cantelli Lemmas

points. After years of practice the probability that he hit anything but 10 will be getting increasingly smaller and smaller and will converge to 0. Thus, the sequence  $\{X_n\}$  converges in probability to  $X = 10$ . Note that  $X_n$  does not converge almost surely however. No matter how professional the archer becomes, there will always be a small probability of making an error. Thus the sequence  $\{X_n\}$  will never turn stationary: there will always be non-perfect scores in it, even if they are becoming increasingly less frequent.

### 例子

Let  $\Omega = [0, 1)$ ,

$$X_{2^{n-1}+k-1}(w) = \begin{cases} 1 & \text{if } w \in [\frac{k-1}{2^{n-1}}, \frac{k}{2^{n-1}}) \\ 0 & \text{otherwise} \end{cases}$$

For any  $\varepsilon > 0$ , we have

$$P(|X_n - 0| > \varepsilon) = \frac{1}{2^{n-1}}, \quad n \geq 1, 2^{n-1} \leq m < 2^n$$

that is  $X_n \xrightarrow{P} 0$ . However, for any  $w \in \Omega$ ,

$$\liminf X_n(w) = 0 < 1 = \limsup X_n(w)$$

then  $X_n$  does not converge to 0 almost surely.

### 例子

Consider the following random experiment: a fair coin is tossed once. Here, the sample space has only two elements  $S = \{H, T\}$ . We define a sequence of random variables  $X_1, X_2, X_3, \dots$  on this sample space as follows:

$$X_n(s) = \begin{cases} \frac{n}{n+1} & \text{if } s = H \\ (-1)^n & \text{if } s = T \end{cases}$$

- For each of the possible outcomes (H or T), determine whether the resulting sequence of real numbers converges or not.
- Find

$$P\left(\left\{s_i \in S : \lim_{n \rightarrow \infty} X_n(s_i) = 1\right\}\right)$$

**Solution** a. If the outcome is H, then we have  $X_n(H) = \frac{n}{n+1}$ , so we obtain the following sequence:

$$\frac{1}{2}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \dots$$

This sequence converges to 1 as  $n$  goes to infinity. If the outcome is T, then we have  $X_n(T) = (-1)^n$ , so we obtain the following sequence

$$-1, 1, -1, 1, \dots$$

This sequence does not converge as it oscillates between -1 and 1 forever.

- b. By part (a), the event  $\{s_i \in S : \lim_{n \rightarrow \infty} X_n(s_i) = 1\}$  happens if and only if the outcome is H,

$$P\left(\left\{s_i \in S : \lim_{n \rightarrow \infty} X_n(s_i) = 1\right\}\right) = P(H) = \frac{1}{2}$$

### 例子

Consider the sample space  $S = [0, 1]$  with a probability measure that is uniform on this space, i.e.,

$$P([a, b]) = b - a, \quad \forall 0 \leq a \leq b \leq 1$$

Define the sequence  $\{X_n, n = 1, 2, \dots\}$  as follows:

$$X_n(s) = \begin{cases} 1 & 0 \leq s < \frac{n+1}{2n} \\ 0 & \text{otherwise} \end{cases}$$

Also, define the random variable  $X$  on this sample space as follows:

$$X(s) = \begin{cases} 1 & 0 \leq s < \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

Show that  $X_n \xrightarrow{a.s.} X$ .

### Solution

Define the set  $A$  as follows:

$$A = \left\{s \in S : \lim_{n \rightarrow \infty} X_n(s) = X(s)\right\}.$$

We need to prove that  $P(A) = 1$ . Let's first find  $A$ . Note that  $\frac{n+1}{2n} > \frac{1}{2}$ , so for any  $s \in [0, \frac{1}{2})$ , we have

$$X_n(s) = X(s) = 1.$$

Therefore, we conclude that  $[0, 0.5) \subseteq A$ . For all  $s > \frac{1}{2}$ , we have

$$\lim_{n \rightarrow \infty} X_n(s) = 0 = X(s)$$

we conclude  $(\frac{1}{2}, 1] \subseteq A$ . You can check that  $s = \frac{1}{2} \notin A$ , since

$$X_n\left(\frac{1}{2}\right) = 1, \quad \forall n,$$

while  $X(\frac{1}{2}) = 0$ . We conclude

$$A = [0, \frac{1}{2}) \cup (\frac{1}{2}, 1] = S \setminus \{\frac{1}{2}\}.$$

Since  $P(A) = 1$ , we conclude  $X_n \xrightarrow{a.s.} X$ .

In some problems, proving almost sure convergence directly can be difficult. Thus, it is desirable to know some sufficient conditions for almost sure convergence.

## 2 Borel-Cantelli Lemmas

### 定理

If for all  $\varepsilon > 0$ , we have

$$\sum_{n=1}^{\infty} P(|X_n - X| > \varepsilon) < \infty,$$

then  $X_n \xrightarrow{a.s.} X$ .

*Proof.*

$$EN = E \sum_{n=1}^{\infty} \mathbb{I}_{|X_n - X| > \varepsilon} = \sum_{n=1}^{\infty} P(|X_n - X| > \varepsilon) < \infty$$

By the Borel-Cantelli lemma,  $P(|X_n - X| > \varepsilon \text{ i.o.}) = 0$ , that is  $P(\limsup\{|X_n - X| > \varepsilon\}) = 0$ , if and only if  $X_n \xrightarrow{a.s.} X$ .  $\square$

### 例子

Consider the sequence:

$$X_n(s) = \begin{cases} -\frac{1}{n} & \text{with probability } \frac{1}{2} \\ \frac{1}{n} & \text{with probability } \frac{1}{2} \end{cases}$$

Show that  $X_n \xrightarrow{a.s.} 0$ .

### Solution

$$\sum_{n=1}^{\infty} P(|X_n| > \varepsilon) = \sum_{n \leq \frac{1}{\varepsilon}} P(|X_n| > \varepsilon) = \frac{1}{[\varepsilon]} < \infty$$

This theorem provides only sufficient condition for almost sure convergence. In particular, if we obtain

$$\sum_{n=1}^{\infty} P(|X_n - X| > \varepsilon) = \infty$$

then we still don't know whether the  $X'_n$ s converge to  $X$  almost surely or not. Here, we provide a condition that is both necessary and sufficient.

### 定理

$X_n \xrightarrow{a.s.} X$  if and only if for any  $\varepsilon > 0$ , we have

$$\lim_{m \rightarrow \infty} P\left(\bigcap_{n=m}^{\infty} |X_n - X| < \varepsilon\right) = 1$$

*Proof.* Since the sequence  $\bigcap_{n=m}^{\infty} |X_n - X| < \varepsilon$  is increasing in  $m$ , the limit exists and

$$\lim_{m \rightarrow \infty} P\left(\bigcap_{n=m}^{\infty} |X_n - X| < \varepsilon\right) = P\left(\bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} |X_n - X| < \varepsilon\right) = P(\liminf |X_n - X| < \varepsilon) = 1$$

if and only if  $P(|X_n - X| > \varepsilon \text{ i.o.}) = 0$ , that is  $X_n \xrightarrow{a.s.} X$ .  $\square$

### 例子

Let  $X_n$  be iid r.v., where  $X_n \sim \mathcal{B}(1/n)$  for  $n = 2, 3, \dots$ , that is

$$X_n(s) = \begin{cases} 1 & \text{with probability } \frac{1}{n} \\ 0 & \text{otherwise} \end{cases}$$

The goal here is to check whether  $X_n \xrightarrow{a.s.} 0$ .

1. Check that  $\sum_{n=1}^{\infty} P(|X_n - X| > \varepsilon) = \infty$ ;
2. Show that the sequence  $X_1, X_2, \dots$  does not converge to 0 almost surely using Theorem above.

**Solution** 1.

$$\begin{aligned} \sum_{n=1}^{\infty} P(|X_n| > \varepsilon) &= \sum_{n=1}^{\infty} P(X_n = 1) \\ &= \sum_{n=1}^{\infty} \frac{1}{n} = \infty \end{aligned}$$

2. for all  $0 < \varepsilon < 1$ , we have

$$\begin{aligned} P\left(\bigcap_{n=m}^{\infty} |X_n - X| < \varepsilon\right) &= P\left(\bigcap_{n=m}^{\infty} \{X_n = 0\}\right) \\ &\leq P\left(\bigcap_{n=m}^N \{X_n = 0\}\right) \\ &= P(X_m = 0)P(X_{m+1} = 0) \cdots P(X_N = 0) \\ &= \frac{m-1}{m} \frac{m}{m+1} \cdots \frac{N-1}{N} \\ &= \frac{m-1}{N} \end{aligned}$$

thus, by choosing  $N$  large enough, we can prove  $\lim_{m \rightarrow \infty} P(\bigcap_{n=m}^{\infty} |X_n - X| < \varepsilon) = 0$ .  
Thus, the sequence does not converge to 0 almost surely.

## 3 Weak Laws of Large Numbers

### 3.1 $L^2$ Weak Laws

#### 定义 (uncorrelation)

Consider  $X_1, X_2, \dots$  as a family of r.v. with  $EX_i^2 < \infty$ , the sequence is said to be uncorrelation if we have

$$EX_i X_j = EX_i EX_j, \text{ whenever } i \neq j$$

#### 定理 ( $L^2$ weak laws)

Let  $X_1, X_2, \dots$  be uncorrelated random variables with  $EX_i = \mu$  and  $\text{var}(X_i) \leq C < \infty$ . If  $S_n = X_1 + \dots + X_n$  then as  $n \rightarrow \infty$ ,  $S_n/n \xrightarrow{L^2} \mu$  and  $S_n/n \xrightarrow{P} \mu$

*Proof.* To prove  $L^2$  convergence, observe that  $E(S_n/n) = \mu$ , so

$$E(S_n/n - \mu)^2 = \text{var}(S_n/n) = \frac{1}{n^2}(\text{var}(X_1) + \dots + \text{var}(X_n)) \leq \frac{Cn}{n^2} \rightarrow 0$$

To conclude there is also convergence in probability, we apply Chebyshev's inequality:

$$P(|S_n/n - \mu| > \varepsilon) \leq \varepsilon^{-2} \int |S_n/n - \mu|^2 dP \rightarrow 0$$

□

### 3.2 The general form weak laws

#### 引理

If  $Y \geq 0$  and  $p > 0$  then  $E(Y^p) = \int_0^\infty py^{p-1}P(Y > y)dy$ .

*Proof.* Using the definition of expected value, Fubini's theorem (for nonnegative random variables), and then calculating the resulting integrals gives

$$\begin{aligned} \int_0^\infty py^{p-1}P(Y > y)dy &= \int_0^\infty \int_\Omega py^{p-1}\mathbb{I}_{\{Y>y\}}dPdy \\ &= \int_\Omega \int_0^\infty py^{p-1}\mathbb{I}_{\{Y>y\}}dPdy \\ &= \int_\Omega \int_0^Y py^{p-1}dPdy = \int_\Omega Y^p dP = EY^p \end{aligned}$$

□

**定理 (weak law for truncation)**

For  $X_n$  iid r.v., let  $\bar{X}_k = X_k \mathbb{I}_{\{|X_k| < n\}}$ . Suppose that as  $n \rightarrow \infty$

1.  $\sum_{k=1}^n P(|X_k| > n) \rightarrow 0$ , and
2.  $n^{-2} \sum_{k=1}^n E\bar{X}_k^2 \rightarrow 0$ .

If we let  $S_n = X_1 + X_2 + \cdots + X_n$ , then

$$\frac{S_n}{n} - E\bar{X}_1 \xrightarrow{P} 0$$

*Proof.* Let  $\bar{S}_n = \bar{X}_1 + \cdots + \bar{X}_n$ . then

$$P\left(\left|\frac{S_n}{n} - E\bar{X}_1\right| > \varepsilon\right) \leq P(S_n \neq \bar{S}_n) + P\left(\left|\frac{\bar{S}_n}{n} - E\bar{X}_1\right| > \varepsilon\right)$$

for the first term,

$$P(S_n \neq \bar{S}_n) \leq P(\cup_{k=1}^n \{\bar{X}_k \neq X_k\}) \leq \sum_{k=1}^n P(|X_k| > n) \rightarrow 0$$

for the second term, by the Chebyshev's inequality:

$$\begin{aligned} P\left(\left|\frac{\bar{S}_n}{n} - E\bar{X}_1\right| > \varepsilon\right) &\leq \varepsilon^{-2} E\left|\frac{\bar{S}_n}{n} - E\bar{X}_1\right|^2 = \varepsilon^{-2} n^{-2} \text{var}(\bar{S}_n) \\ &= (n\varepsilon)^{-2} \sum_{k=1}^n \text{var}(\bar{X}_k) \leq (n\varepsilon)^{-2} \sum_{k=1}^n E(\bar{X}_k)^2 \rightarrow 0 \end{aligned}$$

and the proof is complete. □

**定理 (Weak law of large numbers)**

Let  $X_1, X_2, \dots$  be i.i.d r.v., with

$$xP(|X_i| > x) \rightarrow 0, \text{ as } x \rightarrow \infty$$

Let  $S_n = X_1 + \cdots + X_n$ , and let  $\mu_n = E(X_1 \mathbb{I}_{\{|X_1| \leq n\}})$ . Then  $S_n/n - \mu_n \xrightarrow{P} 0$ .

*Proof.* We need only to show  $n^{-1} E\bar{X}_1^2 \rightarrow 0$ . According to lemma 3.2.1, we have:

$$E(\bar{X}_1) = \int_0^\infty 2yP(|\bar{X}_1| > y) dy \leq \int_0^n 2yP(|X_1| > y) dy$$

since  $P(|\bar{X}_1| > y) = 0$  for  $y \geq n$  and  $P(|\bar{X}_1| > y) = P(|X_1| > y) - P(|X_1| > n)$  for  $y \leq n$ .

We claim that  $yP(|X_1| > y) \rightarrow 0$  implies

$$E(\bar{X}_1)/n \leq \frac{1}{n} \int_0^n 2yP(|X_1| > y) dy \rightarrow 0, \text{ as } n \rightarrow \infty$$

### 3 Weak Laws of Large Numbers

To spell out the details, denote  $g(y) = 2yP(|X_1| > y)$ , this is the average of  $g(y)$  over  $[0, n]$ .  $0 \leq g(y) \leq 2y$  and  $g(y) \rightarrow 0$  as  $y \rightarrow \infty$ . Devide  $[0, n]$  into  $[0, K] \cup [K, n]$ , let  $\epsilon_K = \sup\{g(y) : y > K\}$ , the integrals become:

$$\frac{1}{n} \int_0^n 2yP(|X_1| > y)dy \leq \frac{1}{n}(K * 2K + (n - K)\epsilon_K)$$

thus

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \int_0^n 2yP(|X_1| > y)dy \leq \epsilon_K$$

Since  $K$  is arbitrary and  $\epsilon_K \rightarrow 0$  as  $K \rightarrow \infty$ , the desired result follows.  $\square$

#### 定理 (Weak law of large numbers in general form)

Let  $X_1, X_2, \dots$  be iid r.v. with  $E|X_i| < \infty$ . Let  $S_n = X_1 + X_2 + \dots + X_n$  and let  $\mu = EX_1$ . Then  $S_n/n \xrightarrow{P} \mu$ .

*Proof.* Two applications of the dominated convergence theorem imply

$$\begin{aligned} xP(|X_1| > x) &\leq E(|X_1|\mathbb{I}_{\{|X_1| > x\}}) \rightarrow 0 \text{ as } x \rightarrow \infty \\ \mu_n = E(X_1\mathbb{I}_{\{|X_1| \leq n\}}) &\rightarrow E(X_1) = \mu \text{ as } n \rightarrow \infty \end{aligned}$$

Using Theorem 3.2.3, we see that if  $\varepsilon > 0$  then  $P(|S_n/n - \mu_n| > \varepsilon/2) \rightarrow 0$ . Since  $\mu_n \rightarrow \mu$ , it follows that  $P(|S_n/n - \mu| > \varepsilon) \rightarrow 0$ .  $\square$



## 4 Strong Law of Large Numbers

### 定理 (Strong law of large numbers)

Let  $X_1, X_2, \dots$  be pairwise iid r.v. with  $E|X_i| < \infty$ . Let  $EX_i = \mu$  and  $S_n = X_1 + \dots + X_n$ . Then  $S_n/n - \mu \xrightarrow{a.s.} 0$  as  $n \rightarrow \infty$ .

*Proof.*  $\sum_{k=1}^{\infty} P(|X_k| > k) \leq \int_0^{\infty} P(|X_1| > t) dt = E|X_1| < \infty$ , so  $P(X_k \neq Y_k i.o.) = 0$ . And

$$|S_n/n - \mu| \leq |(S_n - T_n)/n| + |T_n/n - \mu|$$

$$P(S_n \neq T_n) = P(\cup_{k=1}^n X_k \neq Y_k) = \sum_k P(X_k \neq Y_k) = \sum_{k=1}^n P(|X_k| > k) = E(|X_1|) < \infty$$

□

### 引理

$$\sum_{k=1}^{\infty} \text{var}(Y_k)/k^2 \leq 4E|X_1|.$$

*Proof.* Since

$$\text{var}(Y_k) \leq E(Y_k^2) = \int_0^{\infty} 2yP(|Y_k| > y) dy \leq \int_0^k 2yP(|X_1| > y) dy$$

so using Fubini's theorem:

$$\begin{aligned} \sum_{k=1}^{\infty} E(Y_k^2)/k^2 &\leq \sum_{k=1}^{\infty} k^{-2} \int_0^{\infty} \mathbb{I}_{\{y < k\}} 2yP(|X_1| > y) dy \\ &= \int_0^{\infty} \left\{ \sum_{k=1}^{\infty} k^{-2} \mathbb{I}_{\{y < k\}} \right\} 2yP(|X_1| > y) dy \end{aligned}$$

Since  $E|X_1| = \int_0^{\infty} P(|X_1| > y) dy$ , we can complete the proof by showing:

□

### 引理

If  $y \geq 0$ , then  $2y \sum_{k>y} k^{-2} \leq 4$

#### 4 Strong Law of Large Numbers

*Proof.* For  $m \geq 2$ , we have:

$$\sum_{k \geq m} k^{-2} \leq \int_{m-1}^{\infty} x^{-2} dx = (m-1)^{-1}$$

When  $y \geq 1$ , the sum starts with  $k = \lfloor y \rfloor + 1 \geq 2$ , so

$$2y \sum_{k > y} k^{-2} \leq 2y/\lfloor y \rfloor \leq 4$$

since  $y/\lfloor y \rfloor \leq 2$  for  $y \geq 1$ .

For  $0 \leq y \leq 1$ , we have

$$2y \sum_{k > y} k^{-2} \leq 2 \left( 1 + \sum_{k=2}^{\infty} k^{-2} \right) \leq 4$$

□

*Proof.* Since  $|X_n| = X_n^+ - X_n^-$ , we can without loss of generality suppose  $X_n \geq 0$ . Let  $\alpha \geq 1$ ,  $k(n) = \lfloor \alpha^n \rfloor$ , Chebyshev's inequality implies that if  $\varepsilon > 0$

$$\begin{aligned} \sum_{n=1}^{\infty} P(|T_{k(n)} - ET_{k(n)}| > \varepsilon k(n)) &\leq \varepsilon^{-2} \sum_{k=1}^{\infty} \text{var}(T_{k(n)})/k(n)^2 \\ &= \varepsilon^{-2} \sum_{k=1}^{\infty} k(n)^{-2} \sum_{m=1}^{k(n)} \text{var}(Y_m) = \varepsilon^{-2} \sum_{m=1}^{\infty} \text{var}(Y_m) \sum_{n: k(n) \geq m} k(n)^{-2} \end{aligned}$$

(By Fubini's theorem to interchange the two summations of nonnegative terms.) Since  $k(n) = \lfloor \alpha^n \rfloor$ ,  $\lfloor \alpha^n \rfloor \geq \alpha^n/2$ , we have

$$\sum_{n: k(n) \geq m} k(n)^{-2} = \sum_{n: \alpha^n \geq 2m} \lfloor \alpha^n \rfloor^{-2} \leq 4 \sum_{n: \alpha^n \geq m} \alpha^{-2n} \leq 4(1 - \alpha^{-2})^{-1} m^{-2}$$

Combining our computations shows

$$\sum_{k=1}^{\infty} P(|T_{k(n)} - ET_{k(n)}| > \varepsilon k(n)) \leq 4(1 - \alpha^{-2})^{-1} \varepsilon^{-2} \sum_{m=1}^{\infty} \text{var}(Y_m) m^{-2} < \infty$$

By lemma 4.0.2. Then  $(T_{k(n)} - ET_{k(n)})/k(n) \xrightarrow{a.s.} 0$ . The dominated convergence theorem implies  $EY_k \rightarrow EX_1$ , so  $ET_{k(n)}/k(n) \rightarrow EX_1$ , then  $T_{k(n)}/k(n) \xrightarrow{a.s.} EX_1$ . To handle the intermediate values, we observe that if  $k(n) \leq m < k(n+1)$

$$\frac{T_{k(n)}}{k(n+1)} \leq \frac{T_m}{m} \leq \frac{T_{k(n+1)}}{k(n)}$$

since  $k(n+1)/k(n) \rightarrow \alpha$

$$\frac{1}{\alpha} EX_1 \leq \liminf_{n \rightarrow \infty} \frac{T_m}{m} \leq \limsup_{m \rightarrow \infty} \frac{T_m}{m} \leq \alpha EX_1$$

Since  $\alpha > 1$  is arbitrary, the proof is complete.

□

# 5 Central Limit Theorems

## 5.1 Weak convergence

### 定义 (converge weakly)

A sequence of distribution functions is said to converge weakly to a limit  $F$  (written  $F_n \Rightarrow F$ ) if  $F_n(y) \rightarrow F(y)$  for all  $y$  that are continuity points of  $F$ .

### 定义 (converge weakly or converge in distribution)

A sequence of random variables  $X_n$  is said to converge weakly or converge in distribution to a limit  $X_\infty$  (written  $X_n \Rightarrow X_\infty$ ) if their distribution function  $F_n(x) = P(X_n \leq x)$  converge weakly.

Examples.....

## 5.2 Central Limit Theorems

### 定理

Let  $X_1, X_2 \dots$  be iid r.v. with  $EX_i = \mu, \text{var}(X_i) = \sigma^2 \in (0, \infty)$ . If  $S_n = X_1 + X_2 + \dots + X_n$   
Then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \Rightarrow \chi$$

where  $\chi$  has the standard normal distribution.

# 6 Accept-Reject Methods

## 6.1 The Fundamental Theorem of Simulation

定理

Simulating

$$X \sim f(x)$$

is equivalent to Simulating

$$(X, U) \sim \mathcal{U}\{(x, u) : 0 < u < f(x)\}.$$

The solution is to simulate the entire pair  $(X, U)$  at once in a bigger set, where simulation is easier, and take the pair if the constraint is satisfied!

- use that

$$\int_a^b f(x)dx = 1$$

and  $f$  is bounded by  $m$ .

- we can simulate the random pair  $(X, U) \sim \mathcal{U}(0 < u < m)$  by simulating:

—

$$Y \sim \mathcal{U}(a, b)$$

—

$$U|Y = y \sim \mathcal{U}(0, m)$$

- take the pair only if the further constraint  $0 < u < f(y)$  is satisfied.

$$\begin{aligned} P(X \leq x) &= P(Y \leq x | U < f(Y)) \\ &= \frac{\int_a^x \int_0^{f(y)} du dy}{\int_a^b \int_0^{f(y)} du dy} = \int_a^x f(y) dy \end{aligned}$$

**例子 (Beta simulation)**

To generate  $X \sim \mathcal{B}(\alpha, \beta)$ , we take  $Y \sim \mathcal{U}(0, 1)$  and  $U \sim \mathcal{U}(0, m)$ . Where  $m$  is the maximum of the Beta density. For  $\alpha = 2.7$  and  $\beta = 6.3$ , plot the results of generating 1000 pairs  $(Y, U)$ . The probability of acceptance of a given simulation in the box  $[a, b] \times [0, m]$  is given by

$$P(\text{Accept}) = P(U < f(Y)) = \frac{1}{m} \int_0^1 \dots du dy = \frac{1}{m}$$

language R

```
optimize(f=function(x){dbeta(x,2.7,6.3)},
         interval=c(0,1), maximum=TRUE)$objective
Nsim=2500
a = 2.7;b=6.3
M=2.67
u=runif(Nsim,max=M)
y=runif(Nsim)
x=y[u<dbeta(y,a,b)]
xu=u[u<dbeta(y,a,b)]
```

- We can generalize this formular to the situation where the larger set is not a box any longer! With the larger set is of the form:

$$\mathcal{L} = \{(y, u) : 0 < u < m(y)\}$$

the constraints are thus  $m(x) \geq f(x)$ .

- Efficiency dictates that  $m$  be as close as possible to  $f$  in order to avoid wasting simulatings.
- $m(x)$  cannot be a probability density, since  $m(x) \geq f(x)$ . And we Write

$$m(x) = Mg(x) \text{ where } \int_X m(x)dx = \int_X Mg(x)dx = M$$

To simulate  $Y \sim g$  and  $U|Y = y \sim \mathcal{U}(0, Mg(y))$ . If we only accept the  $y$ 's s.t. the constraint  $u < f(y)$  is satisfied, we have

$$\begin{aligned} P(X \in \mathcal{A}) &= P(Y \in \mathcal{A} | U < f(Y)) \\ &= \frac{\int_{\mathcal{A}} \int_0^{f(y)} \frac{du}{Mg(y)} g(y) dy}{\int_{\mathcal{X}} \int_0^{f(y)} \frac{du}{Mg(y)} g(y) dy} = \int_{\mathcal{A}} f(y) dy \end{aligned}$$

for every measurable set  $\mathcal{A}$ , and the accepted  $X$ 's are indeed distributed from  $f$ .

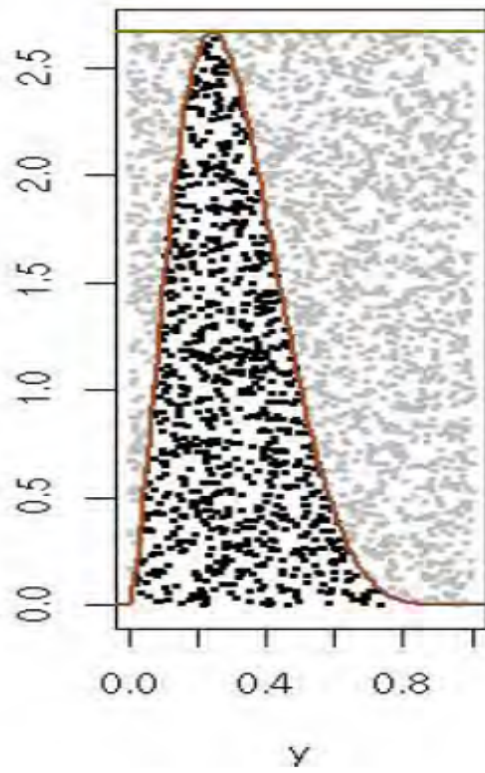


图 1:

**例子**

Simulate  $X \sim f(x)$ , Where

$$f(x) \propto \exp(-x^2/2)(\sin(6x)^2 + 3\cos(x)^2 \sin(4x)^2 + 1)$$

with upper bound the normal density

$$g(x) = \exp(-x^2/2)/\sqrt{2\pi}$$

**6.2 The Accept-Reject Algorithm**

The Accept-Reject Method

1. Generate  $X \sim g, U \sim \mathcal{U}(0, 1)$
2. Accept  $Y = X$  if  $U \leq f(X)/Mg(X)$ ;
3. Return to 1. otherwise.

Note that for  $f/g$  to remain bounded, it is necessary for  $g$  to have tails thicker than those of  $f$ . It is therefore impossible for instance to use A-R to simulate a Cauchy distribution  $f$  using a normal distribution  $g$ ; however the reverse works quite well.

**例子 (Normals from double exponentials)**

Consider generate a  $\mathcal{N}(0, 1)$  using a double exponential distribution  $\mathcal{L}(\alpha)$ , with density

$$g(x|\alpha) = (\alpha/2) \exp(-\alpha|x|).$$

since

$$\frac{f(x)}{g(x|\alpha)} \leq \sqrt{2/\pi} \alpha^{-1} e^{\alpha^2/2}$$

and the minimum of this bound is attained for  $\alpha = 1$ . The probability of acceptance is then  $\sqrt{\pi/2e} = 0.76$ , which shows that to produce one normal random variable, this A-R algorithm requires on the average  $1/0.76 \approx 1.3$  uniform variables.

**例子 (Gamma A-R)**

Simulate  $\mathcal{G}a(\alpha, \beta)$  with  $\mathcal{G}a(a, b)$ , where  $a, b$  are both intergers.  $a = [\alpha](\alpha \geq 1)$ , suppose  $\beta = 1$ . The ratio  $f/g$  is  $b^{-a} x^{\alpha-a} \exp\{-(1-b)x\}$ , up to a normalizing constant, and the bound :

$$M = b^{-a} \left( \frac{\alpha - a}{(1-b)e} \right)^{\alpha-a}, \quad b < 1$$

Since the maximum of  $b^{-a}(1-b)^{\alpha-a}$  is attained at  $b = a/\alpha$ , the optimal choice of  $b$  for simulating  $\mathcal{G}a(\alpha, 1)$  is  $\mathcal{G}a(a, a/\alpha)$ .

**例子 (Truncated normal distributions)**

Truncated normal distributions, with constraints  $x \geq \underline{\mu}$  produce densities proportional to

$$e^{-(x-\underline{\mu})^2/2\sigma^2} I_{x \geq \underline{\mu}}$$

for a bound  $\underline{\mu}$  large compared with  $\mu$ .

1. The naive method:  $X \sim \mathcal{N}(\mu, \sigma^2)$ , only accept  $X$  that is larger than  $\underline{\mu}$ . This approach requires an average number of  $\frac{1}{\Phi(\frac{\mu-\underline{\mu}}{\sigma})}$ .
2. Consider  $\mu = 0, \sigma = 1$ . The translated exponential distribution  $\mathcal{Exp}(\alpha, \underline{\mu})$ , with density

$$g_\alpha(z) = \alpha e^{-\alpha(z-\underline{\mu})} I_{z \geq \underline{\mu}}$$

The ratio  $f/g_\alpha(z) = e^{-\alpha(z-\underline{\mu})} e^{-z^2/2}$ , and is bounded by:

$$\begin{cases} \frac{1}{\alpha} \exp(\alpha^2/2 - \alpha\underline{\mu}) & \text{if } \alpha > \underline{\mu} \\ \frac{1}{\alpha} \exp(-\underline{\mu}^2/2) & \text{otherwise} \end{cases}$$

The first expression is minimized by

$$\alpha^* = \underline{\mu} + \frac{1}{2} \sqrt{\underline{\mu}^2 + 4}$$

the second expression is minimized by  $\alpha' = \underline{\mu}$ . The optimal choice of  $\alpha$  is  $\alpha^*$ .

**6.3 The Kiss Generator****定义 (Period)**

The period,  $T_0$ , of a generator is the smallest integer  $T$  s.t.  $u_{i+T} = u_i, \forall i$ ; that is, s.t.  $D^T$  is equal to the identity function.

- A generator of the form  $X_{n+1} = f(X_n)$  has a period no greater than  $M+1$  (for float type number of C language,  $2^{32}$ )
- In order to overcome this bound, a generator must utilize several sequences  $X_n^i$  simultaneously
- or must involve  $X_{n-1}, X_{n-2}, \dots$  in addition to  $X_n$
- or must use other methods such as start-up tables.

Kiss simultaneously use two generation techniques: *congruential generation* + *shift register generation*.



**定义 (congruential generator)**

A *congruential generator* on  $\{0, 1, \dots, M\}$  is defined by the function:

$$D(x) = (ax + b) \bmod (M + 1)$$

**定义 ( shift register generator)**

For a given  $k \times k$  matrix  $T$ , whose entries are either 0 or 1, the associated *shift register generator* is given by the Transformation

$$x_{n+1} = Tx_n$$

where  $x_n$  is represented as a vector of binary coordinates  $e_{ni}$ , that is to say

$$x_n = \sum_{i=0}^{k-1} e_{ni} 2^i$$

with  $e_{ni}$  equal to 0 or 1.

The generators used by *Kiss* are based on the matrices

$$T_L = \begin{pmatrix} 1 & 1 & & \\ & \ddots & & 0 \\ & & \ddots & 1 \\ & 0 & & 1 \end{pmatrix} T_R = \begin{pmatrix} 1 & & & \\ 1 & \ddots & & 0 \\ & & \ddots & \\ & 0 & 1 & 1 \end{pmatrix} \quad (6.1)$$

$$R(e_1, \dots, e_k)^T = (0, e_1, \dots, e_{k-1})^T$$

$$L(e_1, \dots, e_k)^T = (e_2, e_3, \dots, e_k, 0)^T$$

and  $T_R = (I + R)$ ,  $T_L = (I + L)$ .

The Kiss generator include one congruential generatir:

$$I_{n+1} = (69069 \times I_n + 23606797) \bmod 2^{32}$$

and two shift register generators:

$$J_{n+1} = (I + L^{15})(I + R^{17})J_n \bmod 2^{32}$$

$$K_{n+1} = (I + R^{13})(I + L^{18})K_n \bmod 2^{31}$$

These are then combined to produce:

$$X_{n+1} = (I_{n+1} + J_{n+1} + K_{n+1}) \bmod 2^{32}$$

The Kiss algorithm

```

long int kiss (i,j,k)
unsigned long *i,*j,*k
{
*j = *j ^ (*j<<17);
*k = (*k ^ (*k<<18))&0X7FFFFFFF;
return ((*i = 69069 * (*i) + 23606797) +
(*j ^ = (*j>> 15)) + (*k ^ = (*k >>13)) );
}

```

The period of Kiss is of order  $2^{95}!$  which is almost  $(2^{32})^3$

# 7 Introduction to Bayesian Computation

## 7.1 Rejection sampling

- The aim is to obtain a single random draw from a density  $p(\theta|y)$ , or perhaps an unnormalized density  $q(\theta|y)$  (with  $p(\theta|y) = \frac{q(\theta|y)}{\int q(\theta|y)d\theta}$ )
- Require a positive function  $g(\theta)$  defined on the support of  $p(\theta|y)$ , which has the following properties:
  - We can draw from the probability density proportional to  $g$ .  $g(\theta)$  must have a finite integral;
  - The importance ratio  $\frac{p(\theta|y)}{g(\theta)} \leq M$ , for all  $\theta$
- The algorithm is:
  1. sample  $\theta \sim g(\theta)$ ;
  2. with probability  $\frac{p(\theta|y)}{Mg(\theta)}$ , accept  $\theta$ ; if the drawn is rejected, return to step 1.
- Remark:
  - a good approximate density  $g(\theta)$  should be roughly proportional to  $p(\theta|y)$  (considered as a function of  $\theta$ ). In which case, with a suitable value of  $M$ , we can accept every draw with probability 1.
  - when  $g$  is not nearly proportional to  $p$ , the bound  $M$  must be set so large that almost all draws obtained in step 1 will be rejected in step 2.
  - a virtue of rejection sampling is self-monitoring—if the method is not working efficiently, few simulated draws will be accepted.
  - $g(\theta)$  is chosen to approximate  $p(\theta|y)$  and so in general will depend on  $y$ ;
  - however, in practice we will be considering approximations to one posterior distribution at a time, and the functional dependence of  $g$  on  $y$  is not of interest;
  - Rejection sampling is used in some fast methods for sampling from standard univariate distribution;

## *7 Introduction to Bayesian Computation*

- Rejection sampling can also be used for generic truncated multivariate distributions.

## 8 Metropolis Hastings 算法

### 8.1 A generic Metropolis Hastings algorithm

**Algorithm 4 Metropolis–Hastings**

Given  $x^{(t)}$ ,

1. Generate  $Y_t \sim q(y|x^{(t)})$ .
2. Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with probability } \rho(x^{(t)}, Y_t), \\ x^{(t)} & \text{with probability } 1 - \rho(x^{(t)}, Y_t), \end{cases}$$

where

$$\rho(x, y) = \min \left\{ \frac{f(y)}{f(x)} \frac{q(x|y)}{q(y|x)}, 1 \right\}.$$

图 1:

**定理**

Let  $X^{(t)}$  be the chain produced by the above algorithm. For every conditional distribution  $q$  whose support includes  $\text{supp}(f)$

- (a) the kernel of the chain satisfies the detailed balance condition with  $f$ ;
- (b)  $f$  is a stationary distribution of the chain.

*Proof.* (a) The transition kernel is:

$$K(x, y) = \underbrace{\rho(x, y)q(y|x)}_{\text{accept proposal}} + \underbrace{\left(1 - \int \rho(x, y)q(y|x)dy\right)}_{\text{reject proposal and didn't provide proposal}} \delta_x(y)$$

Now we want to varify detailed balance equation:

$$f(x)K(x, y) = f(y)K(y, x)$$

This can be break into two part:

$$f(x)\rho(x, y)q(y|x) = f(y)\rho(y, x)q(x|y)$$

and

$$f(x)\left(1 - \int \rho(x, y)q(y|x)dy\right)\delta_x(y) = f(y)\left(1 - \int \rho(y, x)q(x|y)dx\right)\delta_y(x)$$

Since

$$\rho(x, y) = \min \left\{ 1, \frac{f(y)q(x|y)}{f(x)q(y|x)} \right\},$$

this completes the proof.

(b) 1

□

**定理**

Suppose that the Metrololis-Hastings Markov chain is  $f$ -irreducible.

- (i) if  $h \in L^1(f)$ , then

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T h(X^{(t)}) = \int h(x)f(x)dx \quad a.e.f.$$

- (ii) If, in addition,  $X^{(t)}$  is aperiodic, then

$$\lim_{n \rightarrow \infty} \left\| \int K^n(x, \cdot) \mu(dx) - f \right\|_{TV} = 0$$

for every initial distribution  $\mu$ , where  $K^n(x, \cdot)$  denotes the kernel for  $n$  transitions.

**Algorithm 5 Independent Metropolis–Hastings**Given  $x^{(t)}$ 

1. Generate  $Y_t \sim g(y)$ .
2. Take

$$X^{(t+1)} = \begin{cases} Y_t & \text{with probability } \min \left\{ \frac{f(Y_t) g(x^{(t)})}{f(x^{(t)}) g(Y_t)}, 1 \right\} \\ x^{(t)} & \text{otherwise.} \end{cases}$$

图 2:

**例子**

Generate the density  $f = Be(2.7, 6.3)$  with M-H algorithm, and the candidate  $q(y|x)$  is  $U[0, 1]$ , which means that it does not depend on the previous value of the chain. The corresponding R code is:

```
a = 2.7; b = 6.3; c = 2.669    #initial values
Nsim = 5000
X = rep( runif(1), Nsim) #initialize the chains
for (i in 2:Nsim){
  Y = runif(1)
  rho = dbeta(Y, a, b) / dbeta(X[i-1], a, b)
  X[i] = X[i-1] + (Y - X[i-1]) * (runif(1) < rho)
}
```

**8.2 The independent M-H algorithm**



**定理**

The algorithm produces a uniformly ergodic chain if there exists a constant  $M$  such that

$$f(x) \leq Mg(x), \quad \forall x \in \text{supp}(f)$$

In this case

$$\|K^n(x, \cdot) - f\|_{TV} \leq 2(1 - 1/M)^n$$

**定理**

If

$$f(x) \leq Mg(x), \quad \forall x \in \text{supp}(f)$$

holds, the expected acceptance probability associated with the algorithm is at least  $1/M$  when the chain is stationary.

**例子**

To generate a Cauchy random variable (that is  $f = \mathcal{C}(0, 1)$ ), it is possible to use a  $\mathcal{N}(0, 1)$  candidate and a  $\mathcal{C}(0, 1)$  candidate within a M-H algorithms. The following R code will do it:

```
Nsim = 10^4
X = c(rt(1,1)) # initialize the chain from the stationary
for (t in 2:Nsim){
  Y=rnorm(1) # candidate normal
  rho=dt(Y,1)*dnorm(X[t-1])/(dt(X[t-1],1)*dnorm(Y))
  X[t]=X[t-1] + (Y - X[t-1])*(runif(1)<rho)
}
```

However when the starting value of  $X^{(0)}$  is a large value, 12.788 say. In this case,  $\text{dnorm}(X[t-1])$  is equal to 0, and

```
pnorm(12.78, log=T, low=F)/log(10)
[1] -36.97455
```

that means the probability of exceeding 12.78 is  $10^{-37}$ , and the Markov chain remains constant for the  $10^4$  iterations. In addition, very large values of the sequence will be heavily weighted, resulting in long strings where the chain remains constant, the isolated peak in the histogram being representative of such an occurrence.

If instead we use for the independent proposal  $g$  a Student's  $t$  distribution with .5 degrees of freedom (that is replace  $Y = \text{rnorm}(1)$  with  $Y = \text{rt}(1, .5)$ ), the R code is as follows:

## 8 Metropolis Hastings 算法

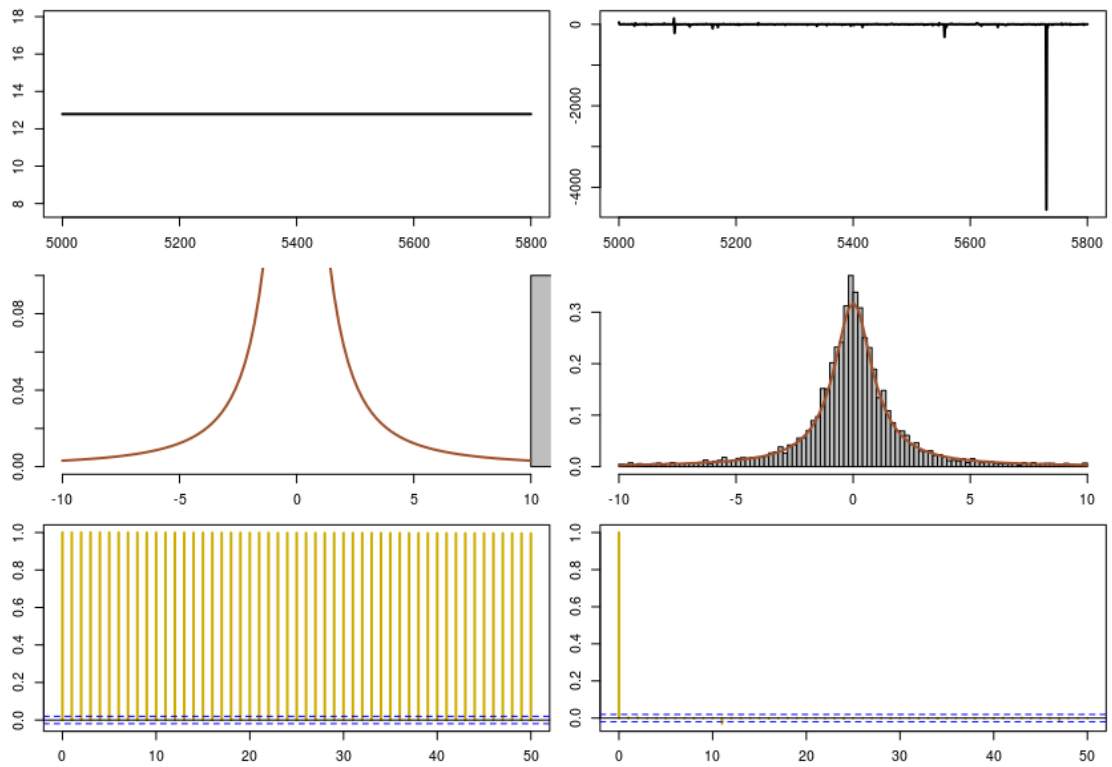


图 3:

```

Nsim = 10^4

Z = 12.788 # initialize the chain from the stationary
for (t in 2:Nsim){
  Y = rt(1, .5) # candidate normal
  rho = dt(Y, 1) * dt(Z[t-1], .5) / (dt(Z[t-1], 1) * dt(Y, .5))
  X[t] = X[t-1] + (Y - X[t-1]) * (runif(1) < rho)
}

```

## 9 Gibbs sampling

### 9.1 The two-stage Gibbs sampler

If the random variables  $X$  and  $Y$  have joint density  $f(x, y)$ , the two-stage Gibbs sampler generates a Markov chain  $(X_t, Y_t)$  according to the following steps: where  $f_{Y|X}$  and  $f_{X|Y}$

**Algorithm A.33 –Two-stage Gibbs sampler–**

```

Take  $X_0 = x_0$ 
For  $t = 1, 2, \dots$ , generate
  1.  $Y_t \sim f_{Y|X}(\cdot | x_{t-1})$ ;
  2.  $X_t \sim f_{X|Y}(\cdot | y_t)$ .

```

[A.33]

图 1:

are the conditional distributions associated with  $f$ :

$$f_Y(y) = \int f(x, y) dx, \quad f_{X|Y}(x|y) = \frac{f(x, y)}{f_Y(y)}$$

Note that not only is the sequence  $(X_t, Y_t)$  a Markov chain, but also each subsequence  $(X_t)$  and  $(Y_t)$  is a Markov chain.

*Proof.* •  $(X_t)$  is a Markov chain: the transition density:

$$K(x, x^*) = \int f_{Y|X}(y|x) f_{X|Y}(x^*|y) dy$$

which indeed depends on the past only through the last value of  $(X_t)$ .

In addition,  $f_X$  is the stationary distribution associated with this subchain, since

$$\begin{aligned}
 f_X(x') &= \int f_{X|Y}(x'|y) f_Y(y) dy \\
 &= \int f_{X|Y}(x'|y) \int f_{Y|X}(y|x) f_X(x) dx dy \\
 &= \int \left[ \int f_{X|Y}(x'|y) f_{Y|X}(y|x) dy \right] f_X(x) dx. \\
 &= \int K(x, x') f_X(x) dx
 \end{aligned}$$

## 9 Gibbs sampling

- $(X_t, Y_t)$  is a Markov chain: the transition kernel:

$$K(x, y; x^*, y^*) = f_{Y|X}(y^*|x^*)f_{X|Y}(x^*|y)$$

which indeed depends on the past only through the last value of  $(X_t, Y_t)$ .

In addition,  $f(x, y)$  is the stationary distribution associated with this subchain, since

$$\begin{aligned} & \int f(x, y)K(x, y; x^*, y^*)dxdy \\ &= \int f(x, y)f_{Y|X}(y^*|x^*)f_{X|Y}(x^*|y)dxdy \\ &= \int \left[ \int f(x, y)dx \right] f_{Y|X}(y^*|x^*)f_{X|Y}(x^*|y)dy \\ &= \left[ \int f_Y(y)f_{X|Y}(x^*|y)dy \right] f_{Y|X}(y^*|x^*) \\ &= \left[ \int f(x^*, y)dy \right] f_{Y|X}(y^*|x^*) \\ &= f_X(x^*)f_{Y|X}(y^*|x^*) \\ &= f(x^*, y^*) \end{aligned}$$

□

### 例子 (Normal bivariate Gibbs)

For the special case of the bivariate normal density,

$$(X, Y) \sim \mathcal{N}_2 \left( 0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

the gibbs sampler is : Given  $y_t$ , generates

$$\begin{aligned} X_{t+1}|y_t &\sim \mathcal{N}(\rho y_t, 1 - \rho^2) \\ Y_{t+1}|x_{t+1} &\sim \mathcal{N}(\rho x_{t+1}, 1 - \rho^2) \end{aligned}$$

*Proof.*

$$\begin{aligned}
f_{X|Y}(x|y) &= \frac{f(x, y)}{f_Y(y)} \\
f_Y(y) &= \int f(x, y) dx \\
f(x, y) &= \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{(x, y)\Sigma^{-1}(x, y)^T}{2}\right) \\
|\Sigma| &= \det\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = 1 - \rho^2 \\
\Sigma^{-1} &= \frac{1}{1 - \rho^2} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \\
f(x, y) &= \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left(-\frac{x^2 + y^2 - 2\rho xy}{2(1 - \rho^2)}\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
f_Y(y) &= \frac{1}{2\pi\sqrt{1 - \rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 + y^2 - 2\rho xy}{2(1 - \rho^2)}\right) dx \\
&= \frac{1}{2\pi\sqrt{1 - \rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2 - 2\rho xy + \rho^2 y^2 - \rho^2 y^2 + y^2}{2(1 - \rho^2)}\right) dx \\
&= \frac{1}{2\pi\sqrt{1 - \rho^2}} \int_{-\infty}^{\infty} \exp\left(-\frac{(x - \rho y)^2}{2(1 - \rho^2)}\right) dx \exp\left(-\frac{-\rho^2 y^2 + y^2}{2(1 - \rho^2)}\right) \\
&= \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{1 - \rho^2}} \exp\left(-\frac{(x - \rho y)^2}{2(1 - \rho^2)}\right) dx}_{=1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \\
&= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)
\end{aligned}$$

then

$$\begin{aligned}
f_{X|Y}(x|y) &= \frac{f(x, y)}{f_Y(y)} \\
&= \frac{\frac{1}{2\pi\sqrt{1 - \rho^2}} \exp\left(-\frac{x^2 + y^2 - 2\rho xy}{2(1 - \rho^2)}\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right)} \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1 - \rho^2}} \exp\left(-\frac{x^2 + y^2 - 2\rho xy - y^2(1 - \rho^2)}{2(1 - \rho^2)}\right) \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1 - \rho^2}} \exp\left(-\frac{x^2 - 2\rho xy + y^2\rho^2}{2(1 - \rho^2)}\right) \\
&= \frac{1}{\sqrt{2\pi}\sqrt{1 - \rho^2}} \exp\left(-\frac{(x - \rho y)^2}{2(1 - \rho^2)}\right) (\mathcal{N}(\rho y, 1 - \rho^2))
\end{aligned}$$

□

Note that the corresponding marginal Markov chain in  $X$  is defined by the AR(1) relation:

$$X_{t+1} = \rho^2 X_t + \sigma \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, 1)$$

*Proof.*

$$k(x, x^*) = \int f(x^*|y)f(y|x)dy = \int \mathcal{N}(y\rho, 1 - \rho^2)\mathcal{N}(x\rho, 1 - \rho^2)dy = \mathcal{N}(x\rho, \sigma^2)$$

In fact,

$$\begin{aligned} k(x, x^*) &= \int \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{(x^*-y\rho)^2}{2(1-\rho^2)}\right) \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{(y-x\rho)^2}{2(1-\rho^2)}\right) dy \\ &= \frac{1}{2\pi(1-\rho^2)} \int \exp\left(-\frac{(x^*)^2 - 2x^*y\rho + (y\rho)^2 + y^2 + (x\rho)^2 - 2x\rho y}{2(1-\rho^2)}\right) dy \\ &= \frac{1}{2\pi(1-\rho^2)} \int \exp\left(-\frac{y^2(\rho^2+1) - 2y(x^*\rho - x\rho) + (x^*)^2 + (x\rho)^2}{2(1-\rho^2)}\right) dy \\ &= \frac{1}{2\pi(1-\rho^2)} \int \exp\left(-\frac{(y\sqrt{1+\rho^2})^2 - 2y\sqrt{1+\rho^2}\frac{(x^*-x)\rho}{\sqrt{1+\rho^2}} + \frac{\rho^2(x^*-x)^2}{1+\rho^2}}{2(1-\rho^2)}\right) dy \\ &\quad \times \exp\left(-\frac{-\frac{\rho^2(x^*-x)^2}{1+\rho^2} + (x^*)^2 + (x\rho)^2}{2(1-\rho^2)}\right) \\ &= \frac{1}{2\pi(1-\rho^2)} \int \exp\left(-\frac{(y\sqrt{1+\rho^2} - \frac{(x^*-x)\rho}{\sqrt{1+\rho^2}})^2}{2(1-\rho^2)}\right) dy \\ &\quad \times \exp\left(-\frac{-\rho^2(x^*-x)^2 + (x^*)^2(1+\rho^2) + (x\rho)^2(1+\rho^2)}{2(1-\rho^2)(1+\rho^2)}\right) \\ &= \frac{1}{2\pi(1-\rho^2)} \int \exp\left(-\frac{(y - \frac{(x^*-x)\rho}{1+\rho^2})^2}{2(1-\rho^2)/(1+\rho^2)}\right) dy \\ &\quad \times \exp\left(-\frac{-\rho^2(x^*)^2 - \rho^2x^2 + 2\rho^2x^*x + (x^*)^2(1+\rho^2) + (x\rho)^2(1+\rho^2)}{2(1-\rho^2)(1+\rho^2)}\right) \\ &= \underbrace{\int \frac{1}{\sqrt{2\pi}\sqrt{\frac{1-\rho^2}{1+\rho^2}}} \exp\left(-\frac{(y - \frac{(x^*-x)\rho}{1+\rho^2})^2}{2(1-\rho^2)/(1+\rho^2)}\right) dy}_{=1} \\ &\quad \times \frac{1}{\sqrt{2\pi}\sqrt{(1-\rho^2)(1+\rho^2)}} \exp\left(-\frac{(x^*)^2 + 2\rho^2x^*x + x^2\rho^4}{2(1-\rho^4)}\right) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{(1-\rho^4)}} \exp\left(-\frac{(x^* - \rho^2x)^2}{2(1-\rho^4)}\right) (\text{Note : } \mathcal{N}(\rho^2x, 1 - \rho^4)) \end{aligned}$$

Actually, we can prove that if  $Z = a\epsilon + b\eta$ ,  $\epsilon, \eta \sim \mathcal{N}(0, 1)$ , then  $Z \sim \mathcal{N}(0, )$ .

$$\begin{aligned}
 P(Z \leq z) &= \frac{1}{ab} \int_{-\infty}^z ds \int_{-\infty}^{\infty} dx p(a\epsilon = x, b\eta = s - a\epsilon) \\
 &= \frac{1}{ab} \int_{-\infty}^z ds \int_{-\infty}^{\infty} dx f(\epsilon = \frac{x}{a}) f(\eta = \frac{s-x}{b}) \\
 &= \frac{1}{ab} \int_{-\infty}^z ds \int_{-\infty}^{\infty} dx \frac{1}{\sqrt{2\pi}} \exp(-\frac{(\frac{x}{a})^2}{2}) \frac{1}{\sqrt{2\pi}} \exp(-\frac{(\frac{s-x}{b})^2}{2}) \\
 &= \frac{1}{ab} \int_{-\infty}^z ds \int_{-\infty}^{\infty} dx \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2a^2}) \frac{1}{\sqrt{2\pi}} \exp(-\frac{(s-x)^2}{2b^2}) \\
 &= \frac{1}{ab} \int_{-\infty}^z ds \int_{-\infty}^{\infty} dx \frac{1}{2\pi} \exp(-\frac{x^2b^2 + (s-x)^2a^2}{2a^2b^2}) \\
 &= \frac{1}{ab} \int_{-\infty}^z ds \int_{-\infty}^{\infty} dx \frac{1}{2\pi} \exp(-\frac{x^2b^2 + (x^2 + s^2 - 2sx)a^2}{2a^2b^2}) \\
 &= \frac{1}{ab} \int_{-\infty}^z ds \int_{-\infty}^{\infty} dx \frac{1}{2\pi} \exp(-\frac{x^2(b^2 + a^2) - 2xsa^2 + s^2a^2}{2a^2b^2}) \\
 &= \frac{1}{ab} \int_{-\infty}^z ds \int_{-\infty}^{\infty} dx \frac{1}{2\pi} \exp(-\frac{(x\sqrt{b^2 + a^2})^2 - 2x\sqrt{a^2 + b^2} \frac{sa^2}{\sqrt{a^2 + b^2}} + (\frac{sa^2}{\sqrt{a^2 + b^2}})^2}{2a^2b^2}) \\
 &\quad \times \exp(-\frac{-(\frac{sa^2}{\sqrt{a^2 + b^2}})^2 + s^2a^2}{2a^2b^2}) \\
 &= \frac{1}{ab} \int_{-\infty}^z ds \int_{-\infty}^{\infty} dx \frac{1}{2\pi} \exp(-\frac{(x\sqrt{b^2 + a^2} - \frac{sa^2}{\sqrt{a^2 + b^2}})^2}{2a^2b^2}) \\
 &\quad \times \exp(-\frac{-(sa^2)^2 + s^2a^2(a^2 + b^2)}{2a^2b^2(a^2 + b^2)}) \\
 &= \frac{1}{ab} \int_{-\infty}^z ds \int_{-\infty}^{\infty} dx \frac{1}{2\pi} \exp(-\frac{(x - \frac{sa^2}{a^2 + b^2})^2}{2a^2b^2/(a^2 + b^2)}) \\
 &\quad \times \exp(-\frac{s^2}{2(a^2 + b^2)}) \\
 &= \frac{1}{ab} \int_{-\infty}^z ds \int_{-\infty}^{\infty} dx \frac{1}{\sqrt{2\pi} \sqrt{a^2b^2/(a^2 + b^2)}} \exp(-\frac{(x - \frac{sa^2}{a^2 + b^2})^2}{2a^2b^2/(a^2 + b^2)}) \\
 &\quad \times \frac{1}{\sqrt{2\pi}} \sqrt{a^2b^2/(a^2 + b^2)} \exp(-\frac{s^2}{2(a^2 + b^2)}) \\
 &= \int_{-\infty}^z ds \frac{1}{\sqrt{2\pi} \sqrt{(a^2 + b^2)}} \exp(-\frac{s^2}{2(a^2 + b^2)})
 \end{aligned}$$

Therefore,

$$f(Z = z) = \frac{1}{\sqrt{2\pi} \sqrt{(a^2 + b^2)}} \exp(-\frac{s^2}{2(a^2 + b^2)})$$

## 9 Gibbs sampling

that is,  $Z \sim \mathcal{N}(0, a^2 + b^2)$ . Then,

$$\begin{aligned}
X_{t+1} &= \rho Y_t + \sqrt{1 - \rho^2} \epsilon_{t+1} \\
&= \rho(\rho X_t + \sqrt{1 - \rho^2} \epsilon_t) + \sqrt{1 - \rho^2} \epsilon_{t+1} \\
&= \rho^2 X_t + \sqrt{[\rho^2(1 - \rho^2) + (1 - \rho^2)]} \epsilon \\
&= \rho^2 X_t + \sqrt{[(\rho^2 + 1)(1 - \rho^2)]} \epsilon \\
&= \rho^2 X_t + \sqrt{1 - \rho^4} \epsilon, \quad (\epsilon \sim \mathcal{N}(0, 1))
\end{aligned}$$

□

The Stationary distribution of this chain is  $\mathcal{N}(0, 1)$ .

*Proof.* A normal distribution  $\mathcal{N}(\mu, \tau^2)$  is stationary for the AR(1) chain  $X_{n+1} \sim \mathcal{N}(\theta x_n, \sigma^2)$ , only if

$$\mu = \theta \mu, \quad \tau^2 = \tau^2 \theta^2 + \sigma^2$$

that is, the relationship for stationary is:

$$\begin{aligned}
&\int_B \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{(x - \mu)^2}{2\tau^2}\right) dx \\
&= \int_X dx \int_B dy \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{(x - \mu)^2}{2\tau^2}\right) \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y - \theta x)^2}{2\sigma^2}\right)
\end{aligned}$$

By Fubini's theorem, interchange the integral order:

$$\begin{aligned}
&\int_B \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{(x - \mu)^2}{2\tau^2}\right) dx \\
&= \int_B dy \int_X dx \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{(x - \mu)^2}{2\tau^2}\right) \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y - \theta x)^2}{2\sigma^2}\right)
\end{aligned}$$

and we can change the variable sign of the left side by  $x$  into  $y$ :

$$\begin{aligned}
&\int_B \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{(y - \mu)^2}{2\tau^2}\right) dy \\
&= \int_B dy \int_X dx \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{(x - \mu)^2}{2\tau^2}\right) \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y - \theta x)^2}{2\sigma^2}\right)
\end{aligned}$$

then, drop the integral about  $y$ , it becomes:

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{(y - \mu)^2}{2\tau^2}\right) dy \\
&= \int_X dx \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{(x - \mu)^2}{2\tau^2}\right) \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(y - \theta x)^2}{2\sigma^2}\right)
\end{aligned}$$



delete the same term  $\frac{1}{\sqrt{2\pi}\tau}$ , we have:

$$\begin{aligned}
 & \exp\left(-\frac{(y-\mu)^2}{2\tau^2}\right) dy \\
 &= \int_X dx \exp\left(-\frac{(x-\mu)^2}{2\tau^2}\right) \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\theta x)^2}{2\sigma^2}\right) \\
 &= \int_X dx \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\tau^2} - \frac{(y-\theta x)^2}{2\sigma^2}\right) \\
 &= \int_X dx \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2 - 2x\mu + \mu^2}{2\tau^2} - \frac{y^2 - 2y\theta x + \theta^2 x^2}{2\sigma^2}\right) \\
 &= \int_X dx \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(x^2\left(\frac{1}{\tau^2} + \frac{\theta^2}{\sigma^2}\right) - 2x\left(\frac{\mu}{\tau^2} + \frac{y\theta}{\sigma^2}\right) + \left(\frac{\mu^2}{\tau^2} + \frac{y^2}{\sigma^2}\right)\right)\right) \\
 &= \underbrace{\int_X dx \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\left(\frac{1}{\tau^2} + \frac{\theta^2}{\sigma^2}\right)\left(x - \frac{\frac{\mu}{\tau^2} + \frac{y\theta}{\sigma^2}}{\left(\frac{1}{\tau^2} + \frac{\theta^2}{\sigma^2}\right)}\right)^2\right)\right)}_{:=A} \\
 &\quad \times \exp\left(-\frac{\left(\frac{\mu}{\tau^2} + \frac{y\theta}{\sigma^2}\right)^2}{\frac{1}{\tau^2} + \frac{\theta^2}{\sigma^2}} + \left(\frac{\mu^2}{\tau^2} + \frac{y^2}{\sigma^2}\right)\right)
 \end{aligned}$$

Since  $A = 1$ , we have

$$\left(\frac{1}{\tau^2} + \frac{\theta^2}{\sigma^2}\right) = \frac{1}{\sigma^2}$$

that is

$$\tau^2\sigma^2 + \theta^2 = \tau^2$$

the relationship above becomes:

$$\begin{aligned}
& \exp\left(-\frac{(y-\mu)^2}{2\tau^2}\right) dy \\
&= \exp\left(-\frac{\left(\frac{\mu}{\tau^2} + \frac{y\theta}{\sigma^2}\right)^2}{\frac{1}{\tau^2} + \frac{\theta^2}{\sigma^2}} + \left(\frac{\mu^2}{\tau^2} + \frac{y^2}{\sigma^2}\right)\right) \\
&= \exp\left(-\sigma^2\left(\frac{\mu}{\tau^2} + \frac{y\theta}{\sigma^2}\right)^2 + \left(\frac{\mu^2}{\tau^2} + \frac{y^2}{\sigma^2}\right)\right) \\
&= \exp\left(-\sigma^2\left(\frac{\mu^2}{\tau^4} + \frac{y^2\theta^2}{\sigma^4} + 2\frac{\mu y\theta}{\tau^2\sigma^2}\right) + \frac{\mu^2}{\tau^2} + \frac{y^2}{\sigma^2}\right) \\
&= \exp\left(-\left(\frac{\mu^2\sigma^2}{\tau^4} + \frac{y^2\theta^2}{\sigma^2} + 2\frac{\mu y\theta}{\tau^2}\right) + \frac{\mu^2}{\tau^2} + \frac{y^2}{\sigma^2}\right) \\
&= \exp\left(-y^2\left(\underbrace{-\frac{1}{\sigma^2} + \frac{\theta^2}{\sigma^2}}_{=\frac{1}{\tau^2}}\right) - 2y\frac{\mu\theta}{\tau^2} - \frac{\mu^2\sigma^2}{\tau^4} + \frac{\mu^2}{\tau^2}\right) \\
&= \exp\left(-\frac{y^2}{\tau^2} - 2y\frac{\mu\theta}{\tau^2} - \frac{\mu^2\sigma^2}{\tau^4} + \frac{\mu^2}{\tau^2}\right) \\
&= \exp\left(-\frac{y^2 - 2y\mu\theta}{\tau^2} + C\right) \\
&= \exp\left(-\frac{(y - \mu\theta)^2}{\tau^2} + C'\right)
\end{aligned}$$

therefore, we have

$$\mu\theta = \mu$$

then, the stationary distribution is  $\mathcal{N}(0, \frac{\sigma^2}{1-\theta^2})$ .  $\square$

In our example,  $\theta = \rho^2$ ,  $\sigma^2 = 1 - \rho^4$ , then the stationary distribution is  $\mathcal{N}(0, \frac{1-\rho^4}{1-\rho^4}) = \mathcal{N}(0, 1)$ .

## 9.2 Missing data and latent variables

- with the auxiliary variable  $z$ , the complete-data likelihood or complete -model:

$$L^c(\theta|x, z) = f(x, z|\theta)$$

which corresponds to the observation of the complete data  $(x, z)$ . This is often referred as demarginalization.

- the likelihood can be expressed as

$$g(x|\theta) = \int_{\mathcal{Z}} f(x, z|\theta) dz$$

**Example 1.1. Censored data models.** *Censored data models* are missing data models where densities are not sampled directly. To obtain estimates and make inferences in such models usually requires involved computations and precludes analytical answers.

In a typical simple statistical model, we would observe random variables<sup>3</sup> (rv's)  $Y_1, \dots, Y_n$ , drawn independently from a population with distribution  $f(y|\theta)$ . The distribution of the sample would then be given by the product  $\prod_{i=1}^n f(y_i|\theta)$ . Inference about  $\theta$  would be based on this distribution.

In many studies, particularly in medical statistics, we have to deal with *censored* random variables; that is, rather than observing  $Y_1$ , we may observe  $\min\{Y_1, \bar{u}\}$ , where  $\bar{u}$  is a constant. For example, if  $Y_1$  is the survival time of a patient receiving a particular treatment and  $\bar{u}$  is the length of the study being done (say  $\bar{u} = 5$  years), then if the patient survives longer than 5 years, we do not observe the survival time, but rather the censored value  $\bar{u}$ . This modification leads to a more difficult evaluation of the sample density.

Barring cases where the censoring phenomenon can be ignored, several types of censoring can be categorized by their relation with an underlying (unobserved) model,  $Y_i \sim f(y_i|\theta)$ :

图 2:

- and the conditional distribution of the missing data  $Z$  given the observed data  $x$   $k(z|\theta, x)$  is

$$k(z|\theta, x) = \frac{f(x, z|\theta)}{g(x|\theta)}$$

## 9.3 gibbs sampling on mixture model

## 9 Gibbs sampling

- (i) Given random variables  $Y_i$ , which are, for instance, times of observation or concentrations, the actual observations are  $Y_i^* = \min\{Y_i, \bar{u}\}$ , where  $\bar{u}$  is the maximal observation duration, the smallest measurable concentration rate, or some other truncation point.
- (ii) The original variables  $Y_i$  are kept in the sample with probability  $\rho(y_i)$  and the number of censored variables is either known or unknown.
- (iii) The variables  $Y_i$  are associated with auxiliary variables  $X_i \sim g$  such that  $y_i^* = h(y_i, x_i)$  is the observation. Typically,  $h(y_i, x_i) = \min(y_i, x_i)$ . The fact that truncation occurred, namely the variable  $\mathbb{I}_{Y_i > X_i}$ , may be either known or unknown.

As a particular example, if

$$X \sim \mathcal{N}(\theta, \sigma^2) \quad \text{and} \quad Y \sim \mathcal{N}(\mu, \tau^2),$$

the variable  $Z = X \wedge Y = \min(X, Y)$  is distributed as

$$(1.1) \quad \begin{aligned} & \left[1 - \Phi\left(\frac{z - \theta}{\sigma}\right)\right] \times \tau^{-1} \varphi\left(\frac{z - \mu}{\tau}\right) \\ & + \left[1 - \Phi\left(\frac{z - \mu}{\tau}\right)\right] \sigma^{-1} \varphi\left(\frac{z - \theta}{\sigma}\right), \end{aligned}$$

where  $\varphi$  is the density of the normal  $\mathcal{N}(0, 1)$  distribution and  $\Phi$  is the corresponding cdf, which is not easy to compute.

图 3:

**Example 5.13.** Censored data may come from experiments where some potential observations are replaced with a lower bound because they take too long to observe. Suppose that we observe  $Y_1, \dots, Y_m$ , iid, from  $f(y - \theta)$  and that the  $(n - m)$  remaining  $(Y_{m+1}, \dots, Y_n)$  are censored at the threshold  $a$ . The corresponding likelihood function is then

$$(5.10) \quad L(\theta|\mathbf{y}) = [1 - F(a - \theta)]^{n-m} \prod_{i=1}^m f(y_i - \theta),$$

where  $F$  is the cdf associated with  $f$  and  $\mathbf{y} = (y_1, \dots, y_m)$ . If we had observed the last  $n - m$  values, say  $\mathbf{z} = (z_{m+1}, \dots, z_n)$ , with  $z_i \geq a$  ( $i = m + 1, \dots, n$ ), we could have constructed the (complete data) likelihood

$$L^c(\theta|\mathbf{y}, \mathbf{z}) = \prod_{i=1}^m f(y_i - \theta) \prod_{i=m+1}^n f(z_i - \theta).$$

Note that

$$L(\theta|\mathbf{y}) = \mathbb{E}[L^c(\theta|\mathbf{y}, \mathbf{Z})] = \int_{\mathcal{Z}} L^c(\theta|\mathbf{y}, \mathbf{z}) f(\mathbf{z}|\mathbf{y}, \theta) d\mathbf{z},$$

where  $f(\mathbf{z}|\mathbf{y}, \theta)$  is the density of the missing data conditional on the observed data, namely the product of the  $f(z_i - \theta)/[1 - F(a - \theta)]$ 's; i.e.,  $f(z - \theta)$  restricted to  $(a, +\infty)$ . ◀

图 4:

**Example 7.6.** In Examples 5.13 and 5.14, we treated a censored-data model as a missing-data model. We identify  $g(x|\theta)$  with the likelihood function

$$g(x|\theta) = L(\theta|x) \propto \prod_{i=1}^m e^{-(x_i - \theta)^2/2},$$

and

$$f(x, z|\theta) = L(\theta|x, z) \propto \prod_{i=1}^m e^{-(x_i - \theta)^2/2} \prod_{i=m+1}^n e^{-(z_i - \theta)^2/2}$$

is the complete-data likelihood. Given a prior distribution  $\pi(\theta)$  on  $\theta$ , we can then create a Gibbs sampler that iterates between the conditional distributions

$$\pi(\theta|x, z) \quad \text{and} \quad f(z|x, \theta)$$

and will have stationary distribution  $\pi(\theta, z|x)$ , the posterior distribution of  $(\theta, z)$ .

Taking a flat prior  $\pi(\theta) = 1$ , the conditional distribution of  $\theta|x, z$  is given by

$$\theta|x, z \sim \mathcal{N}\left(\frac{m\bar{x} + (n-m)\bar{z}}{n}, \frac{1}{n}\right),$$

图 5:

while the conditional distribution of  $Z|x, \theta$  is the product of the truncated normals

$$Z_i|x, \theta \sim \varphi(z - \theta) / \{1 - \Phi(a - \theta)\},$$

as each  $Z_i$  must be greater than the truncation point  $a$ . Generating values of  $Z$  can be done via the R function `rtrun` from the package `bayesm` (see Exercises 7.21 and 7.7). The outcome of the Gibbs sampler, whose R core can be written as

```
> for(i in 2:Nsim){
+   zbar[i]=mean(rtrun(mean=rep(that[i-1],n-m),
+   sigma=rep(1,n-m),a=rep(a,n-m),b=rep(Inf,n-m)))
+   that[i]=rnorm(1,(m/n)*xbar+(1-m/n)*zbar[i],sqrt(1/n))
+ }
```

图 6:

**Example 7.7.** Recall the multinomial model of Example 5.16,

$$\mathcal{M}\left(n; \frac{1}{2} + \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right).$$

where we estimated  $\theta$  using either EM or MCEM steps, introducing the latent variable  $Z$  with the demarginalization

$$(z, x_1 - z, x_2, x_3, x_4) \sim \mathcal{M}\left(n; \frac{1}{2}, \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}\right).$$

If we use a uniform prior on  $\theta$ , the full conditionals can be recovered as

$$\theta \sim \mathcal{Be}(z + x_4 + 1, x_2 + x_3 + 1) \text{ and } z \sim \mathcal{Bin}\left(x_1, \frac{\theta}{2 + \theta}\right),$$

图 7:

leading to the Gibbs sampler

```
> x=c(125,18,20,34)           #data
> theta=z=rep(.5,Nsim)        #init chain
> for (j in 2:Nsim){
>   theta[j]=rbeta(1,z[j-1]+x[4]+1,x[2]+x[3]+1)
>   z[j]=rbinom(1,x[1],(theta[j]/(2+theta[j])))
> }
```

whose output is summarized in Figure 7.5.

图 8:

**Example 9.2. Gibbs sampling on mixture posterior.** Consider a *mixture of distributions*

9.1 A General Class of Two-Stage Algorithms 341

$$(9.5) \quad \sum_{j=1}^k p_j f(x|\xi_j),$$

where  $f(\cdot|\xi)$  belongs to an exponential family

$$f(x|\xi) = h(x) \exp\{\xi \cdot x - \psi(\xi)\}$$

and  $\xi$  is distributed from the associated conjugate prior

$$\pi(\xi|\alpha_0, \lambda) \propto \exp\{\lambda(\xi \cdot \alpha_0 - \psi(\xi))\}, \quad \lambda > 0, \quad \alpha_0 \in \mathcal{X},$$

while

$$(p_1, \dots, p_k) \sim \mathcal{D}_k(\gamma_1, \dots, \gamma_k).$$

Given a sample  $(x_1, \dots, x_n)$  from (9.5), we can associate with every observation an indicator variable  $z_i \in \{1, \dots, k\}$  that indicates which component of the mixture is associated with  $x_i$  (see Problems 5.8–5.10). The demarginalization (or *completion*) of model (9.5) is then

$$Z_i \sim \mathcal{M}_k(1; p_1, \dots, p_k), \quad x_i|z_i \sim f(x|\xi_{z_i}).$$

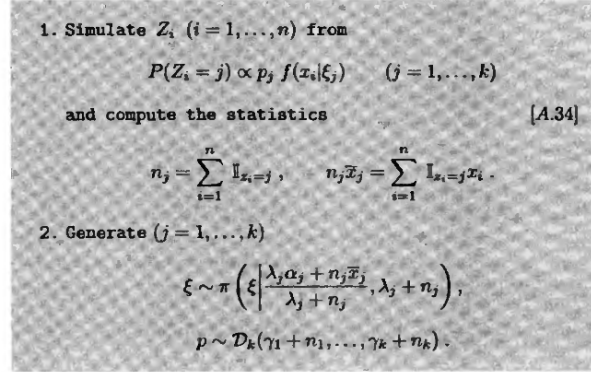
图 9:

Thus, considering  $x_i^* = (x_i, z_i)$  (instead of  $x_i$ ) entirely eliminates the mixture structure since the likelihood of the completed model is

$$\begin{aligned}\ell(p, \xi | x_i^*, \dots, x_i^*) &\propto \prod_{i=1}^n p_{z_i} f(x_i | \xi_{z_i}) \\ &= \prod_{j=1}^k \prod_{i: z_i=j} p_j f(x_i | \xi_j) .\end{aligned}$$

(This latent structure is also exploited in the original implementation of the EM algorithm; see Section 5.3.2.) The two steps of the Gibbs sampler are

**Algorithm A.34 –Mixture Posterior Simulation–**



1. **Simulate**  $Z_i$  ( $i = 1, \dots, n$ ) **from**

$$P(Z_i = j) \propto p_j f(x_i | \xi_j) \quad (j = 1, \dots, k)$$

**and compute the statistics** [A.34]

$$n_j = \sum_{i=1}^n \mathbb{I}_{z_i=j}, \quad n_j \bar{x}_j = \sum_{i=1}^n \mathbb{I}_{z_i=j} x_i .$$

2. **Generate** ( $j = 1, \dots, k$ )

$$\xi \sim \pi \left( \xi \left| \frac{\lambda_j \alpha_j + n_j \bar{x}_j}{\lambda_j + n_j}, \lambda_j + n_j \right. \right),$$

$$p \sim \mathcal{D}_k(\gamma_1 + n_1, \dots, \gamma_k + n_k) .$$

图 10:



## 9.4 generate random number from truncated normal distribution

$$\begin{aligned}
\pi(\xi_j|x, z) &\propto \pi(\xi_j|\alpha_0, \lambda) \prod_{i:z_i=j} f(x_i|\xi_j) \\
&\propto \exp(\lambda(\xi_j \cdot \alpha_0 - \psi(\xi_j))) \prod_{i:z_i=j} h(x_i) \exp(\xi_j \cdot x_i - \psi(\xi_j)) \\
&\propto \exp\left(\xi_j(\lambda \cdot \alpha_0 + \sum_{i:z_i=j} x_i) - (\lambda + \sum_{i:z_i=j} 1)\psi(\xi_j)\right) \\
&\propto \exp\left((\lambda + \sum_{i:z_i=j} 1) \left(\xi_j \cdot \frac{(\lambda \cdot \alpha_0 + \sum_{i:z_i=j} x_i)}{\lambda + \sum_{i:z_i=j} 1} - \psi(\xi_j)\right)\right) \\
&\propto \exp\left((\lambda + n_j) \left(\xi_j \cdot \frac{(\lambda \cdot \alpha_0 + n_j \bar{x}_j)}{\lambda + n_j} - \psi(\xi_j)\right)\right) \\
&\propto \pi(\xi_j | \frac{(\lambda \cdot \alpha_0 + n_j \bar{x}_j)}{\lambda + n_j}, \lambda + n_j)
\end{aligned}$$

$$\begin{aligned}
\pi(p_1, \dots, p_k|x, z) &\propto \mathcal{D}_k(\gamma_1, \dots, \gamma_k) \prod_{j=1}^k \prod_{i:z_i=j} p_j f(x_i|\xi_j) \\
&\propto p_1^{\gamma_1-1} \dots p_k^{\gamma_k-1} p_1^{n_1} \dots p_k^{n_k} \\
&= p_1^{\gamma_1+n_1-1} \dots p_k^{\gamma_k+n_k-1} \\
&\propto \mathcal{D}_k(\gamma_1 + n_1, \dots, \gamma_k + n_k)
\end{aligned}$$

## 9.4 generate random number from truncated normal distribution

### 9.4.1 The Inverse Transform

定义

For a non-decreasing function  $F$  on  $\mathbb{R}$ , the generalized inverse of  $F$ ,  $F^-$ , is the function defined by

$$F^-(u) = \inf\{x : f(x) \geq u\}.$$

As an illustration, consider the same setting as Example 5.19, namely a normal mixture with two components with equal known variance and fixed weights,

$$(9.6) \quad p\mathcal{N}(\mu_1, \sigma^2) + (1-p)\mathcal{N}(\mu_2, \sigma^2).$$

We assume in addition a normal  $\mathcal{N}(0, 10\sigma^2)$  prior distribution on both means  $\mu_1$  and  $\mu_2$ . Generating directly from the posterior associated with a sample  $\mathbf{x} = (x_1, \dots, x_n)$  from (9.6) quickly turns impossible, as discussed for instance in Diebolt and Robert (1994) and Celeux et al. (2000), because of a combinatoric explosion in the number of calculations, which grow as  $\mathcal{O}(2^n)$ .

As for the EM algorithm (Problems 5.8–5.10), a natural completion of  $(\mu_1, \mu_2)$  is to introduce the (unobserved) component indicators  $z_i$  of the observations  $x_i$ , namely,

$$P(Z_i = 1) = 1 - P(Z_i = 2) = p \quad \text{and} \quad X_i | Z_i = k \sim \mathcal{N}(\mu_k, \sigma^2).$$

The completed distribution is thus

$$\begin{aligned} \pi(\mu_1, \mu_2, \mathbf{z} | \mathbf{x}) &\propto \exp\{-(\mu_1^2 + \mu_2^2)/20\sigma^2\} \prod_{z_i=1} p \exp\{-(x_i - \mu_1)^2/2\sigma^2\} \times \\ &\quad \prod_{z_i=2} (1-p) \exp\{-(x_i - \mu_2)^2/2\sigma^2\}. \end{aligned}$$

图 11:

Since  $\mu_1$  and  $\mu_2$  are independent, given  $(\mathbf{z}, \mathbf{x})$ , with distributions  $(j = 1, 2)$ , the conditional distributions are

$$\mathcal{N}\left(\sum_{z_i=j} x_i / (.1 + n_j), \sigma^2 / (.1 + n_j)\right),$$

where  $n_j$  denotes the number of  $z_i$ 's equal to  $j$ . Similarly, the conditional distribution of  $\mathbf{z}$  given  $(\mu_1, \mu_2)$  is a product of binomials, with

$$\begin{aligned} &P(Z_i = 1 | x_i, \mu_1, \mu_2) \\ &= \frac{p \exp\{-(x_i - \mu_1)^2/2\sigma^2\}}{p \exp\{-(x_i - \mu_1)^2/2\sigma^2\} + (1-p) \exp\{-(x_i - \mu_2)^2/2\sigma^2\}}. \end{aligned}$$

图 12:

**引理**

If  $U \sim \mathcal{U}(0, 1)$ , then the random variable  $F^{-}(U)$  has the distribution  $F$ .

*Proof.* For all  $u \in [0, 1]$  and for all  $x \in F^{-}([0, 1])$ , the generalized inverse satisfies:

$$\{u : F^{-} \leq x\} = \{u : F(x) \geq u\}$$

therefore,

$$P(F^{-}(U) \leq x) = P(F(x) \geq U) = F(x)$$

□

**例子 (Exponential variable generation)**

If  $X \sim \mathcal{Exp}(1)$ , so  $F(x) = 1 - e^{-x}$ , then solving for  $x$  in  $u = 1 - e^{-x}$  gives  $x = -\log(1 - u)$ . Therefore, if  $U \sim \mathcal{U}(0, 1)$ , the random variable  $X = -\log U$  has the exponential distribution.

### 9.4.2 generate random number from truncated normal distribution—by the inverse transform

```

for (i in 2:nsim){
  temp=runif(n-m,min=pnorm(a,mean=that[i-1],sd=1),
    max=1)
  zbar[i]=mean(qnorm(temp,mean=that[i-1],sd=1))
  that[i]=rnorm(1,mean=(m/n)*xbar+(1-m/n)*zbar,
    sd=sqrt(1/n))}

```

Since the distribution of the truncated normal distribution is

$$F(x) = \int_a^x \frac{\varphi(x)}{1 - \Phi(a)} dx = \frac{\Phi(x) - \Phi(a)}{1 - \Phi(a)}$$

Then

$$\begin{aligned}
 F(x) &= u \\
 \Leftrightarrow \Phi(x) &= u(1 - \Phi(a)) + \Phi(a) \\
 \Leftrightarrow x &= \Phi^{-1}\left(\underbrace{u(1 - \Phi(a)) + \Phi(a)}_{\text{runif}(n-m, \text{min}=\text{pnorm}(a, \text{mean}=\text{that}[i-1], \text{sd}=1), \text{max}=1)}\right) = F^{-1}(u)
 \end{aligned}$$

## 9.5 The slice sampler

### 9.5.1 The fundamental theorem

This section, we focus on uniform generation on the subgraph of  $f$ :

$$\mathcal{R}(f) = \{(x, u) : 0 \leq u \leq f(x)\}$$

Starting from a point  $(x, u)$  in  $\mathcal{R}(f)$ , the move along the  $u$ -axis will correspond to the conditional distribution

$$U|X = x \sim \mathcal{U}(\{u : u \leq f(x)\})$$

and then the move along the  $x$ -axis to the conditional distribution

$$X|U' \sim \mathcal{U}(\{x : u' \leq f(x)\})$$

**Algorithm A.31 –2D slice sampler–**

**At iteration  $t$ , simulate**

1.  $u^{(t+1)} \sim \mathcal{U}_{[0, f(x^{(t)})]}$ ;
2.  $x^{(t+1)} \sim \mathcal{U}_{A^{(t+1)}}$ , with

[A.31]

$$A^{(t+1)} = \{x : f(x) \geq u^{(t+1)}\}.$$

图 13:

We inaccurately call this method the 2D slice sampler:

**8.1** Referring to Algorithm [A.31] and equations (8.1) and (8.2):

- (a) Show that the stationary distribution of the Markov chain [A.31] is the uniform distribution on the set  $\{(x, u) : 0 < u < f(x)\}$ .
- (b) Show that the conclusion of part (a) remains the same if we use  $f_1$  in [A.31], where  $f(x) = Cf_1(x)$ .

图 14:

if we set  $f(x) = Cf_1(x)$  and use  $f_1$  instead of  $f$ (see problem 8.1), the algorithm remains valid.

Now we want to prove this algorithm forms a Markov chains:

the subgraph of  $f$ : first, if  $x^{(t)} \sim f(x)$  and  $u^{(t+1)} \sim \mathcal{U}_{[0, f_1(x^{(t)})]}$ , then

$$(x^{(t)}, u^{(t+1)}) \sim f(x) \frac{\mathbb{I}_{[0, f_1(x)]}(u)}{f_1(x)} \propto \mathbb{I}_{0 \leq u \leq f_1(x)}.$$

Second, if  $x^{(t+1)} \sim \mathcal{U}_{A^{(t+1)}}$ , then

$$(x^{(t)}, u^{(t+1)}, x^{(t+1)}) \sim f(x^{(t)}) \frac{\mathbb{I}_{[0, f_1(x^{(t)})]}(u^{(t+1)})}{f_1(x^{(t)})} \frac{\mathbb{I}_{A^{(t+1)}}(x^{(t+1)})}{\text{mes}(A^{(t+1)})},$$

where  $\text{mes}(A^{(t+1)})$  denotes the (generally Lebesgue) measure of the set  $A^{(t+1)}$ .  
Thus

$$\begin{aligned} (x^{(t+1)}, u^{(t+1)}) &\sim C \int \mathbb{I}_{0 \leq u \leq f_1(x)} \frac{\mathbb{I}_{f_1(x^{(t+1)}) \geq u}}{\text{mes}(A^{(t+1)})} dx \\ &= C \mathbb{I}_{0 \leq u \leq f_1(x^{(t+1)})} \int \frac{\mathbb{I}_{u \leq f_1(x)}}{\text{mes}(A^{(t+1)})} dx \\ &\propto \mathbb{I}_{0 \leq u \leq f_1(x^{(t+1)})} \end{aligned}$$

and the uniform distribution on  $\mathcal{S}(f)$  is indeed stationary for both steps.

图 15:

#### 例子 (Simple slice sampler)

Consider the density  $f(x) = \frac{1}{2}e^{-\sqrt{x}}$  for  $x > 0$ . we can simulate by the followings:

$$U|x \sim \mathcal{U}(0, \frac{1}{2}e^{-\sqrt{x}}), \quad X|u \sim \mathcal{U}(0, [\log(2u)]^2)$$

Moreover, we can also simulate by problem 8.2:



**8.2** In the setup of Example 8.1, show that the cdf associated with the density  $\exp(-\sqrt{x})$  on  $\mathbb{R}_+$  can be computed in closed form. (*Hint:* Make the change of variable  $z = \sqrt{x}$  and do an integration by parts on  $z \exp(-z)$ .)

图 16:

**例子 (Truncated normal distribution)**

A truncated normal distribution  $N(-3, 1)$  restricted to the interval  $[0, 1]$ :

$$f(x) \propto f_1(x) = \exp\{-(x+3)^2/2\} \mathbb{I}_{[0,1]}(x)$$

**9.6 Back to the Gibbs Sampler**

- The slice sampler can be interpreted as a special case of two-stage Gibbs sampler,
- the slice sampler starts with  $f_X(x)$  and creates a joint density  $f(x, u) = \mathbb{I}(0 < u < f_X(x))$ .
- the associated conditional densities are

$$f_{X|U}(x|u) = \frac{\mathbb{I}(0 < u < f_X(x))}{\int \mathbb{I}(0 < u < f_X(x)) dx}, \quad f_{U|X}(u|x) = \frac{\mathbb{I}(0 < u < f_X(x))}{\int \mathbb{I}(0 < u < f_X(x)) du}$$

- therefore, the  $X$  sequence is also a Markov chain with transition kernel

$$K(x, x') = \int f_{X|U}(x'|u) f_{U|X}(u|x) du$$

and stationary density  $f_X(x)$ .

- Moreover, we can induce a Gibbs sampler for any marginal distribution  $f_X(x)$  by creating a joint distribution that is, formally, arbitrary.
- Starting from  $f_X(x)$ , we can take any conditional density  $g(y|x)$  and create a Gibbs sampler with

$$f_{X|Y}(x|y) = \frac{g(y|x)f_X(x)}{\int g(y|x)f_X(x)dx}, \quad f_{Y|X}(y|x) = \frac{g(y|x)f_X(x)}{\int g(y|x)f_X(x)dy}$$

**9.7 The Hammersley-Clifford Theorem**

- A most surprising feature of the Gibbs sampler is that the conditional distributions contain sufficient information to produce a sample from the joint distribution.

**定理**

The joint distribution associated with the conditional densities  $f_{Y|X}(y|x)$  and  $f_{X|Y}(x|y)$  has the joint density

$$f(x, y) = \frac{f_{Y|X}(y|x)}{\int \frac{f_{Y|X}(y|x)}{f_{X|Y}(x|y)} dy}$$

- By comparison with maximization problems, this approach is similar to maximization an objective function successively in every direction of a given basis.
- It is well known that this optimization method does not necessarily lead to the global maximum, but may end up in a saddlepoint.

*Proof.* Since  $f(x, y) = f_{Y|X}(y|x)f_X(x) = f_{X|Y}(x|y)f_Y(y)$  we have

$$\int \frac{f_{Y|X}(y|x)}{f_{X|Y}(x|y)} dy = \int \frac{f_Y(y)}{f_X(x)} dy = \frac{1}{f_X(x)}$$

and the result follows.  $\square$

## 9.8 Examples

### 例子 (Grouped counting data)

For 360 consecutive time units, consider recording the number of passages of individuals, per unit time, past some sensor.

表 1: Frequencies of passage for 360 consecutive observations

Number of passages	0	1	2	3	4or more
Number of observations	139	128	55	25	13

**Algorithm A.35 –Poisson–Gamma Gibbs Sampler–**

**Given**  $\lambda^{(t-1)}$ ,

- 1. Simulate**  $Y_i^{(t)} \sim \mathcal{P}(\lambda^{(t-1)}) \mathbf{I}_{Y \geq 4}$  ( $i = 1, \dots, 13$ )
- 2. Simulate** [A.35]

$$\lambda^{(t)} \sim \mathcal{Ga} \left( 313 + \sum_{i=1}^{13} y_i^{(t)}, 360 \right).$$

图 17:

- If we assume that every observation is a Poisson  $\mathcal{P}(\lambda)$ , the likelihood of the model corresponding to the table below is

$$l(\lambda|x_1, \dots, x_5) \propto e^{-347\lambda} \lambda^{128+55 \times 2 + 25 \times 3} \left( 1 - e^{-\lambda} \sum_{i=0}^3 \frac{\lambda^i}{i!} \right)$$

- Assume  $\pi(\lambda) = \frac{1}{\lambda}$  and  $\mathbf{y} = (y_1, \dots, y_{13})$ , vector of the 13 units larger in  $\pi(\lambda, y_1, y_2, \dots, y_{13}|x_1, \dots, x_5)$ . The Gibbs sampler is:

- The convergence of the Rao-Blackwellized estimator:

$$\delta_{rb} = \frac{1}{T} \sum_{t=1}^T E[\lambda | x_1, \dots, x_5, y_1^{(t)}, \dots, y_{13}^{(t)}] = \frac{1}{360T} \sum_{t=1}^T \left( 313 + \sum_{i=1}^{13} y_i^{(t)} \right)$$

- Rao-Blackwellized estimator

$$\text{var}[E(\delta(X)|Y)] \leq \text{var}[\delta(X)]$$

**例子 (grouped multinomial data)**

Consider the multinomial model

$$X \sim \mathcal{M}_5(n; a_1\mu + b_1, a_2\mu + b_2, a_3\eta + b_3, a_4\eta + b_4, c(1 - \mu - \eta))$$

where the  $a_i, b_i$  are known.

- This model is equivalent to a sampling from

$$Y \sim \mathcal{M}_9(n; a_1\mu, b_1, a_2\mu, b_2, a_3\eta, b_3, a_4\eta, b_4, c(1 - \mu - \eta))$$

where

$$X_1 = Y_1 + Y_2, \quad X_2 = Y_3 + Y_4, \quad X_3 = Y_5 + Y_6, \quad X_4 = Y_7 + Y_8, \quad X_5 = Y_9$$

- A natural prior distribution on  $(\mu, \eta)$  is the Dirichlet prior  $\mathcal{D}(\alpha_1, \alpha_2, \alpha_3)$ ,

$$\pi(\mu, \eta) \propto \mu^{\alpha_1-1} \eta^{\alpha_2-1} (1 - \eta - \mu)^{\alpha_3-1}$$

where  $\alpha_1 = \alpha_2 = \alpha_3 = 1/2$ .

- We define  $Z = (Z_1, Z_2, Z_3, Z_4) = (Y_1, Y_3, Y_5, Y_7)$ , the completed posterior distribution can be defined as

$$\begin{aligned} \pi(\eta, \mu | x, z) &= \pi(\eta, \mu | y) \\ &\propto \mu^{\alpha_1-1} \eta^{\alpha_2-1} (1 - \eta - \mu)^{\alpha_3-1} \mu^{z_1} \mu^{z_2} \eta^{z_3} \eta^{z_4} (1 - \eta - \mu)^{x_5} \\ &= \mu^{z_1+z_2+\alpha_1-1} \eta^{z_3+z_4+\alpha_2-1} (1 - \eta - \mu)^{x_5+\alpha_3-1} \end{aligned}$$

- thus

$$(\mu, \eta, 1 - \mu - \eta) | x, z \sim \mathcal{D}(z_1 + z_2 + \alpha_1, z_3 + z_4 + \alpha_2, x_5 + \alpha_3)$$

- moreover,

$$\begin{aligned} Z_i | x, \eta, \mu &\sim \mathcal{B}(x_i, \frac{a_i \mu}{a_i \mu + b_i}), \quad (i = 1, 2) \\ Z_i | x, \eta, \mu &\sim \mathcal{B}(x_i, \frac{a_i \eta}{a_i \eta + b_i}), \quad (i = 3, 4) \end{aligned}$$

**Table 7.1.** Observed genotype frequencies on blood type data. The effect of a dominant allele creates a missing-data problem.

Genotype	Probability	Observed	Probability	Frequency
AA	$p_A^2$	A	$p_A^2 + 2p_Ap_O$	$n_A = 186$
AO	$2p_Ap_O$			
BB	$p_B^2$	B	$p_B^2 + 2p_Bp_O$	$n_B = 38$
BO	$2p_Bp_O$			
AB	$2p_Ap_B$	AB	$p_Ap_B$	$n_{AB} = 13$
OO	$p_O^2$	O	$p_O^2$	$n_O = 284$

图 18:

例子 •

$$X \sim \mathcal{M}_4(n; \underbrace{p_A^2 + 2p_Ap_O}_{X_1}, \underbrace{p_B^2 + 2p_Bp_O}_{X_2}, \underbrace{p_Ap_B}_{X_3}, \underbrace{p_O^2}_{X_4})$$

- introduce latent variable:

$$X_1 = Z_1 + Z_2, \quad X_2 = Z_3 + Z_4$$

then, the complete model is

$$(Z_1, X_1 - Z_1, Z_3, X_2 - Z_3, X_3, X_4) \sim \mathcal{M}_6(p_A^2, 2p_Ap_O, p_B^2, 2p_Bp_O, p_Ap_B, p_O^2)$$

- The prior of  $p_A, p_B, p_O$  is defined as following:

$$\begin{aligned} \pi(p_A, p_B, 1 - p_A - p_B) &\sim \mathcal{D}(\alpha_A, \alpha_B, \alpha_O) \\ &\propto p_A^{\alpha_A-1} p_B^{\alpha_B-1} (1 - p_A - p_B)^{\alpha_O-1} \end{aligned}$$

- Then the posterior of  $p_A, p_B, 1 - p_A - p_B$  is

$$\begin{aligned} &\pi(p_A, p_B, 1 - p_A - p_B | X_1, X_2, X_3, X_4, Z_1, Z_3) \\ &\propto p_A^{\alpha_A-1} p_B^{\alpha_B-1} (1 - p_A - p_B)^{\alpha_O-1} \\ &\times (p_A^2)^{Z_1} (2p_Ap_O)^{X_1-Z_1} (p_B^2)^{Z_3} (2p_Bp_O)^{X_2-Z_3} (p_Ap_B)^{X_3} (p_O^2)^{X_4} \\ &\propto (p_A)^{\alpha_A-1+2Z_1+X_1-Z_1+X_3} (p_B)^{\alpha_B-1+2Z_3+X_2-Z_3+X_3} \\ &\times (1 - p_A - p_B)^{\alpha_O-1+X_1-Z_1+X_2-Z_3+2X_4} \\ &\propto (p_A)^{\alpha_A-1+X_1+Z_1+X_3} (p_B)^{\alpha_B-1+X_2+Z_3+X_3} \\ &\times (1 - p_A - p_B)^{\alpha_O-1+X_1-Z_1+X_2-Z_3+2X_4} \end{aligned}$$

## 9 Gibbs sampling

- Since  $X_1 = n_A, X_2 = n_B, X_3 = n_{AB}, X_4 = n_O$

$$\begin{aligned} & \pi(p_A, p_B, 1 - p_A - p_B | X_1, X_2, X_3, X_4, Z_1, Z_3) \\ & \propto (p_A)^{\alpha_A - 1 + n_A + Z_1 + n_{AB}} (p_B)^{\alpha_B - 1 + n_B + Z_3 + n_{AB}} \\ & \times (1 - p_A - p_B)^{\alpha_O - 1 + n_A - Z_1 + n_B - Z_3 + 2n_O} \end{aligned}$$

- define  $Z_1 = Z_A, Z_3 = Z_B$ , the posterio becomes

$$(p_A, p_B, p_O | n_A, n_B, n_{AB}, n_O, Z_A, Z_B) \sim \mathcal{D}(\alpha_A + n_A + Z_A + n_{AB}, \alpha_B + n_B + Z_B + n_{AB}, \alpha_O + n_A - Z_A + n_B - Z_B + 2n_O) \quad (9.1)$$

•

$$Z_A, Z_B | p_A, p_B, p_O, n_A, n_B, n_{AB}, n_O \sim \frac{L^c(p_A, p_B, p_O | n_A, n_B, n_{AB}, n_O, Z_A, Z_B)}{L(p_A, p_B, p_O | n_A, n_B, n_{AB}, n_O)}$$

•

$$\begin{aligned} Z_A & \sim \mathcal{B}(n_A; \frac{p_A^2}{p_A^2 + 2p_A p_O}) \\ Z_B & \sim \mathcal{B}(n_B; \frac{p_B^2}{p_B^2 + 2p_B p_O}) \end{aligned}$$



## 9.9 importance sampling

- Background: Suppose we are interested in  $E(h(\theta)|y)$ , but we can't generate random draws of  $\theta$  from  $p(\theta|y)$ .
- If  $g(\theta)$  is a probability density from which we can generate random draws.
- then we can write

$$E(h(\theta)|y) = \int h(\theta)p(\theta|y)d\theta = \int [h(\theta)\frac{p(\theta|y)}{g(\theta)}]g(\theta)d\theta \quad (9.2)$$

here  $w(\theta) \triangleq \frac{p(\theta|y)}{g(\theta)}$  are called importance ratios.

- Then the algorithm is:

1. sample  $\theta \sim g(\theta)$
2. calculate

$$\frac{1}{S} \sum_{s=1}^S h(\theta^s)w(\theta^s)$$

### Accuracy and efficiency of importance sampling estimates

- aim: to avoid missing some extremely large but rare importance weights.
- method: examine the distribution of sampled importance weights ( the histogram of the logarithms of the largest importance ratios) to discover possible problems:
  - estimates will be poor if the largest ratio are too large relative to the average.
  - In contrast, do not need to worry about the behavior of small importance ratios, because they have little influence on equation 9.2.
- If the variance of the weights is finite, the effective sample size can be estimated using an approximation:

$$S_{eff} = \frac{1}{\sum_{s=1}^S (w(\theta^s))^2}$$

### Importance resampling/sampling-importance resampling, SIR

Once  $S$  draws,  $\{\theta^1, \dots, \theta^S\} \sim g(\theta)$ , a sample  $k < S$  draws can be simulated as follows:

- sample a value  $\theta$  from the set  $\{\theta^1, \dots, \theta^S\}$ , with the probability  $w(\theta^s)$ ;
- sample a second value using the same procedure, but excluding the already sampled value from the set;

## 9 Gibbs sampling

- repeatedly sample without replacement  $k-2$  times.

Why sample without replacement?

- If the importance weights are moderate, sampling with and without replacement gives similar results.
- consider a bad case, with a few large weights and many small weights. Sampling with replacement will pick the same few values of  $w_i$  repeatedly; in contrast, sampling without replacement yields a more desirable intermediate approximation somewhere between the starting and target densities.

# 10 chapter 11 Basics of Markov chain simulation

## the background of MC

- The key to the method's success, however, is not the Markov property but rather that the approximate distributions are improved at each step in the simulation, in the sense of converging to the target distribution.
- As we shall see in Section 11.2, the Markov property is helpful in proving this convergence.
- Markov chain simulation is used when
  1. it is not possible
  2. not computationally efficientto sample directly from  $p(\theta|y)$ ;
- instead we sample iteratively in such a way that at each step of the process the distribution becomes closer to  $p(\theta|y)$ .
- after we have obtain the overall simulations, we should also
  1. check the convergence of the simulated sequences
  2. construct an expression for the effective number of simulation draws for a correlated sample.

## 10.1 11.1 Gibbs sampler/ alternating conditional sampling

Suppose the parameter vector  $\theta = (\theta_1, \dots, \theta_d)$ , the algorithm is:

1. sample  $\theta_j^t \sim p(\theta_j | \theta_{-j}^{t-1}, y)$

repeat this step for  $j = 1, \dots, d$ .

here denote  $\theta_j^t$  as the  $j$ th component of  $\theta$  at iteration  $t$ . and  $\theta_{-j}^{t-1} = (\theta_1^{t-1}, \dots, \theta_{j-1}^{t-1}, \theta_{j+1}^{t-1}, \dots, \theta_d^{t-1})$ , which represents all the components of  $\theta$ , except for  $\theta_j$ , at their current values.

**例子 (Bivariate normal distribution)**

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \bigg| y \sim \mathcal{N} \left( \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

## 10.2 11.2 Metropolis and M-H 算法

### 10.2.1 Metropolis 算法

Metropolis 算法是一种带有 A/C 规则的随机游走算法。

step 1. 选一个初始的  $\theta^0 \sim p(\theta^0)$ , 使得  $p(\theta^0 | y) > 0$ ;

step 2. 对于  $t = 1, 2, \dots$

(a) 提出建议  $\theta^* \sim J_t(\theta^* | \theta^{t-1})$ , 对称性要求  $J_t(\theta_a | \theta_b) = J_t(\theta_b | \theta_a)$ ,  $\forall \theta_a, \theta_b, t$ ;

(b) 计算  $r = \frac{p(\theta^* | y)}{p(\theta^{t-1} | y)}$ ;

(c) 设

$$\theta^t = \begin{cases} \theta^* & \text{以概率 } \min(r, 1) \\ \theta^{t-1} & \text{otherwise} \end{cases}$$

**转移分布**

$$T_t(\theta^t | \theta^{t-1}) = \delta_{\theta^{t-1}}(\theta^t) + J_t(\theta^t | \theta^{t-1}) \min(r, 1)$$

是一个  $\theta^t = \theta^{t-1}$  的单点估计量和跳分布  $J_t(\theta^t | \theta^{t-1})$  的加权, 权重为接受概率。

**算法要求具有一下能力**

(a)  $r$  可以计算

(b) 可以从  $J_t(\theta^* | \theta)$  采样  $\theta$

(c) step c 需要产生一个均匀分布随机数。

例子 (二元正态分布) •  $p(\theta|y) = \mathcal{N}(\theta|0, I)$

- $J_t(\theta^*|\theta^{t-1}) = \mathcal{N}(\theta^*|\theta^{t-1}, 0.2^2 I)$

- $r = \frac{\mathcal{N}(\theta^*|0, I)}{\mathcal{N}(\theta^{t-1}|0, I)}$

与优化的关系:

- 若这次跳使得后验分布增加, 设  $\theta^t = \theta^*$ ;
- 若这次跳使得后验分布减小, 以概率  $r$  设  $\theta^t = \theta^*$ , 否则设  $\theta^t = \theta^{t-1}$

可以视为分段找 mode 的随机版本: 若增大密度, 接受这一步; 若降低密度, 有时也接受。

收敛性

### 10.2.2 Metropolis-Hastings 算法

相比 M 算法

(1)  $J_t$  不再对称

(2)  $r$  变成

$$r = \frac{p(\theta^*|y)/J_t(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y)/J_t(\theta^{t-1}|\theta^*)}$$

收敛性

remark

- 理想的 M-H 跳是  $J_t(\theta^*|\theta) = p(\theta^*|y)$ , 就是目标分布,  $\forall \theta^*$ , 这时  $r \equiv 1$ , 但这是没有意义的;
- 好的 M-H 的性质
  1.  $\forall \theta$ ,  $J_t(\theta^*|\theta)$  易于采样
  2.  $r$  容易计算
  3. 每次跳在参数空间距离合适
  4. 不要拒绝太频繁

# 11 Monitoring Convergence to the Stationary Distribution

## 11.1 推断、评价收敛

从迭代的 simulation 中推断带来的问题

- 如果迭代不足, simulation 不足以代表 target 分布。即使 simulation 已经足够接近收敛, 早期的迭代仍然影响总体的近似效果;
- 序列内的相关性: 收敛后, 序列内的相关性影响有效样本数。

用三种方法来处理

- (1) 设计 simulation 使得可以进行有效的 convergence monitor, 例如把初始  $\theta$  在参数空间选得分散开
- (2) 通过大致比较序列内和序列间的方差 (between and within simulated sequence)
- (3) 若有效性仍旧不可接受, 换算法。

具体来说

1. 去掉前一半的 simulation, warm-up
2. 对序列内的相关性, 隔几个留一个样本
3. 从分散开的初始点多跑几条链
4. monitor 标量变化
5. monitoring convergence 的挑战: mixing and stationarity

a) 如图 11.3(a), 每条链单独看似乎均已稳定, 但两条链并列说明还没收敛到共同的分布

b) 如图 11.3(b), 两条似乎存在一个共同的分布, 但均不平稳。

这两个例子说明了当评价收敛时应考虑 between-sequence 和 within-sequence 的信息。

6. 把每条保存下来的链分成两半, 如图1

- 把每条链分成两半, 检查最终全部的半链均已 mixed
- 假设 warm-up 部分已经被删去

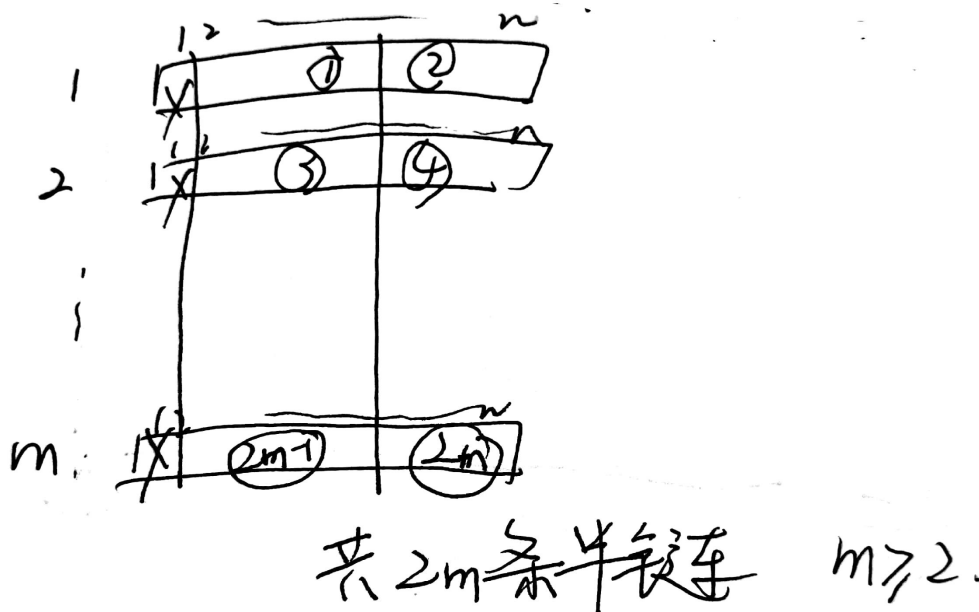


图 1: 共有  $m(m \geq 2)$  条链, 每条链有  $n$  个样本, 删去 warm-up 后, 每条链分为两半。共有  $2m$  条半链

7. 用序列内和序列间的方差 (Between-sequence, within-sequence) 评估 mixing:  $\theta_{ij}$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, m$

$$\text{Between-sequence } B = \frac{n}{m-1} \sum_{j=1}^m (\bar{\theta}_{\cdot,j} - \bar{\theta}_{\cdot\cdot})^2$$

$$\text{Within-sequence } W = \frac{1}{m} \sum_{j=1}^m s_j^2, \quad s_j^2 = \frac{1}{n-1} \sum_{i=1}^n (\theta_{ij} - \bar{\theta}_{\cdot,j})^2$$

$$\bar{\theta}_{\cdot j} = \frac{1}{n} \sum_{i=1}^n \theta_{ij}, \quad \bar{\theta}_{..} = \frac{1}{m} \sum_{j=1}^m \bar{\theta}_{\cdot j}$$

总方差定义为：

$$\text{var}(\theta|y) = \hat{\text{var}}^+(\theta|y) = \frac{n-1}{n}W + \frac{1}{n}B$$

这个统计量再平稳或者极限 ( $n \rightarrow \infty$ ) 的意义下是无偏的。

一般用伸缩变换后的方差：

$$\hat{R} = \sqrt{\frac{\hat{\text{var}}^+(\theta|y)}{W}}$$

当  $n \rightarrow \infty$  时，降到 1。

若  $\hat{R}$  较大，有理由相信，更多的迭代将改进推断。

### 例子 (二元正态分布)

以二元正态跳核

#### 11.1.1 simulation draws 的有效数目

•

$$\lim_{n \rightarrow \infty} mn \text{var}(\bar{\theta}_{..}) = (1 + 2 \sum_{t=1}^{\infty} \rho_t) \text{var}(\theta|y)$$

$\rho_t$  是  $\theta$  在 lag  $t$  处的自相关系数。

•

$$n_{eff} = \frac{mn}{1 + 2 \sum_{t=1}^{\infty} \rho_t}$$

• 下面估计  $\rho_t$

定义

$$V_t = \frac{1}{m(n-t)} \sum_{j=1}^m \sum_{i=t+1}^n (\theta_{ij} - \theta_{i-t,j})^2$$

由  $E(\theta_i - \theta_{i-t})^2 = 2(1 - \rho_t) \text{var}(\theta)$  所以有

$$\hat{\rho}_t = 1 - \frac{V_t}{2\hat{\text{var}}(\theta|y)}$$

但  $\rho_t$  不能取到  $\infty$ ，部分和一般从 lag 0 开始，直到连续两个 lag  $\hat{\rho}_{2t} + \hat{\rho}_{2t+1} < 0$

$$n_{eff} = \frac{mn}{1 + 2 \sum_{t=1}^T \rho_t}$$

何时终止 simulation 计算  $\hat{R}$  和  $n_{eff}$

• 如果  $\hat{R}$  不是近似为 1，继续迭代；一般以 1.1 为阈值

•  $\hat{n}_{eff}$  给出精度的意义：一般要保证  $\hat{n}_{eff} \geq 5m$ ，(例如，如果共有两条链，每条链至少有 10 个有效样本。)



## 11.2 Graphical diagnoses

### 例子 (Logit model using Monte Carlo EM and Metropolis-Hastings)

A simple random effect logit model processed in Booth and Hobert(1999) represents observations  $y_{i,j}$  ( $i = 1, \dots, n, j = 1, \dots, m$ ) as distributed conditionally on one covariate  $x_{i,j}$  as a logit model:

$$P(y_{ij} = 1 | x_{ij}, u_i, \beta) = \frac{\exp(\beta x_{ij} + u_i)}{1 + \exp(\beta x_{ij} + u_i)}$$

where  $u_i \sim \mathcal{N}(0, \sigma^2)$  is an unobserved random effect. The vector of random effects  $(U_1, \dots, U_n)$  therefore corresponds to the missing data  $Z$ . When considering the function  $Q(\theta' | \theta, x, y)$ ,

$$\begin{aligned} Q(\theta' | \theta, x, y) &= \sum_{i,j} y_{ij} \mathbb{E}[\beta' x_{ij} + U_i | \beta, \sigma, x, y] \\ &\quad - \sum_{i,j} \mathbb{E}[\log[1 + \exp\{\beta' x_{ij} + U_i\}] | \beta, \sigma, x, y] \\ &\quad - \sum_i \mathbb{E}[U_i^2 | \beta, \sigma, x, y] / 2\sigma'^2 - n \log \sigma' \end{aligned}$$

with  $\theta = (\beta, \sigma)$ . It is impossible to compute the expectations in  $U_i$ . The M-step would then almost straightforward for maximizing  $Q(\theta' | \theta, x, y)$  in  $\sigma'$  leads to

$$\sigma'^2 = \frac{1}{n} \sum_i \mathbb{E}[U_i^2 | \beta, \sigma, x, y]$$

while maximizing  $Q(\theta' | \theta, x, y)$  in  $\beta'$  produces the fixed-point equation

$$\sum_{i,j} y_{ij} x_{ij} = \sum_{i,j} \mathbb{E} \left[ \frac{\exp\{\beta' x_{ij} + U_i\}}{1 + \exp\{\beta' x_{ij} + U_i\}} | \beta, \sigma, x, y \right] x_{ij}$$

The alternative to EM is to simulate the  $U_i$ 's conditional on  $\beta, \sigma, x, y$  in order to replace the expectations above with Monte Carlo approximations.

$$\pi(u_i | \beta, \sigma, x, y) \propto \frac{\exp\{\sum_j y_{ij} u_i - u_i^2 / 2\sigma^2\}}{\prod_j [1 + \exp\{\beta x_{ij} + u_i\}]}$$

Opting for a standard random walk Metropolis-Hastings algorithm, we simulate both  $u_i, \beta$  from Normal distributions centered at the previous values of those parameters:

$$u_i^{(t)} \sim \mathcal{N}(u_i^{(t-1)}, \sigma^2), \quad \beta^{(t)} \sim \mathcal{N}(\beta^{(t-1)}, \tau)$$

The scale parameter  $\sigma$  can be simulated directly from an inverse gamma distribution:

$$\sigma \sim I\Gamma(1, 4 * \sum_i u_i^2 / n)$$

## 11.3 Nonparametric tests of stationarity

Kolmogorov-Smirnov statistic

$$K = \frac{1}{M} \sup \left| \sum_{g=1}^M \mathbb{I}_{(0,\eta)}(x_1^{(gG)}) - \sum_{g=1}^M \mathbb{I}_{(0,\eta)}(x_2^{(gG)}) \right|$$

## 11.4 A missing Mass

To assess how much of the support of the target distribution has been explored by the chain via an evaluation of

$$\int_{\mathcal{A}} f(x) dx$$

If  $\mathcal{A}$  denotes the support of the distribution of the chain. This is not necessarily easy, especially in large dimensions, but we can use the Riemann approximation method to assess it. When  $f$  is a one-dimensional density, the quantity

$$\sum_{t=1}^T [\theta^{(t+1)} - \theta^{(t)}] f(\theta^{(t)})$$

converges to 1, even when the  $\theta^{(t)}$  are not generated from the density  $f$ .

### 例子 (Bimodal target)

Consider the density

$$f(x) = \frac{\exp(-x^2/2)}{\sqrt{2\pi}} \frac{4(x - 0.3)^2 + 0.01}{4(1 + (0.3)^2) + 0.01}$$

## 11.5 Geweke.diag function

Geweke.diag utilize Spectral analysis. Geweke takes the first  $T_A$  and the last  $T_B$  observations from a sequence of length  $T$  to derive

$$\delta_A = \frac{1}{T_A} \sum_{t=1}^{T_A} h(x^{(t)}), \quad \delta_B = \frac{1}{T_B} \sum_{t=T-T_B+1}^T h(x^{(t)}),$$

and the estimates  $\sigma_A$  and  $\sigma_B$  based on both subsamples, respectively. The test statistic is then the asymptotically normal so-called Z-score

$$\sqrt{T}(\delta_A - \delta_B) / \sqrt{\frac{\sigma_A^2}{\tau_A} + \frac{\sigma_B^2}{\tau_B}},$$

with  $T_A = \tau_A T$ ,  $T_B = \tau_B T$ , and  $\tau_A + \tau_B < 1$ . This is a t test which assess the equality of the means of the first and last parts of the Markov chain.

## 11.6 Kolmoforov-Smirnov statistic

The empirical distribution function  $F_n$  for  $n$  iid observations  $X_i$  is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{[-\infty, x]}(X_i)$$

The Kolmoforiv-Smirnov statistic for a given cumulative distribution function  $F(x)$  is

$$D_n = \sup_x |F_n(x) - F(x)|$$

- By the Glivenko-Cantelli theorem,  $D_n$  converges to 0 almost surely.
- Kolmogorov strengthened this result, by effectively providing the rate of this Convergence

$$\lim_n P(\sqrt{n}D_n \leq x) = 1 - 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2 x^2}$$

- the goodness-of-fit test or the Kolmogorov-Smirnov test is constructed by using the critical values of the Kolmogorov distribution. The null hypothesis is rejected at level  $\alpha$  if

$$\sqrt{n}D_n > K_\alpha$$

where  $K_\alpha$  is found from

$$Pr(K \leq K_\alpha) = 1 - \alpha$$

$\alpha$	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.22	1.36	1.48	1.63	1.73	1.95

## 11.7 Two-sample Kolmoforov-Smirnov test

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$$

where  $F_{1,n}$  and  $F_{2,m}$  are the empirical distribution functions of the first and the second sample respectively. The null hypothesis is rejected at level  $\alpha$  if

$$D_{n,m} > c(\alpha) \sqrt{\frac{n+m}{nm}}$$

and in general by

$$c(\alpha) = \sqrt{-\frac{1}{2} \ln\left(\frac{\alpha}{2}\right)}$$

## 11.8 summary–convergence diagnosing

- randogibs.R:

```
library(coda)
plot(mcmc(cbind(beta, sigma)))
browser()
cumuplot(mcmc(cbind(beta, sigma)))
list(beta = beta, sigma = sigma)
```

- kscheck.R

```
#old ks
ks=NULL
M=10

for (t in seq(T/10,T,le=100)){
  beta1=beta[1:(t/2)]
  beta2=beta[(t/2)+(1:(t/2))]
  beta1=beta1[seq(1,t/2,by=M)]
  beta2=beta2[seq(1,t/2,by=M)]
  ks=c(ks, ks.test(beta1, beta2)$p)
}
```

$t = \text{seq}(1000, 10000, \text{by} = 90)$

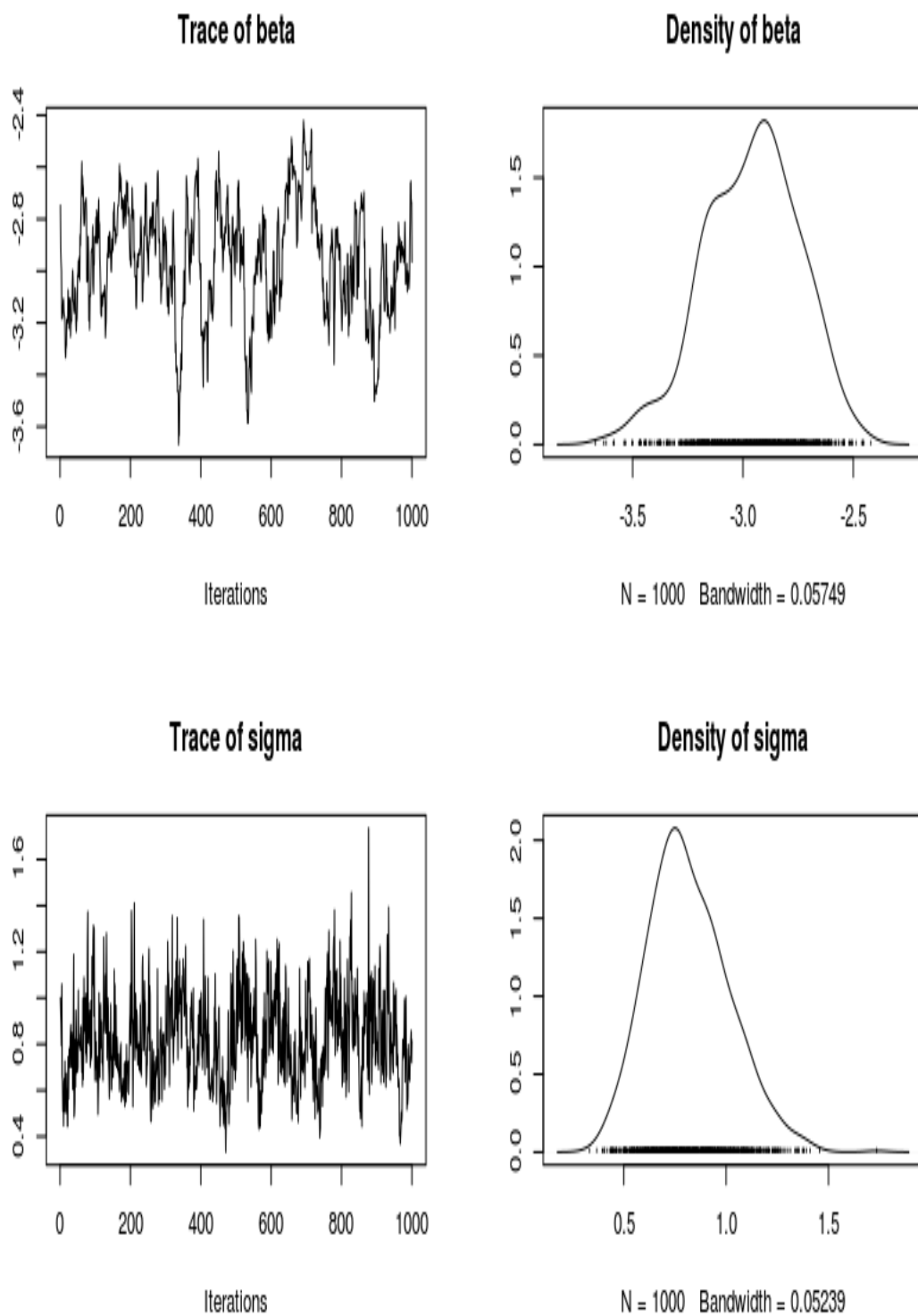
for example, if  $t = 1000$ ,

$\beta_1 = \beta(1 : 500), \beta_2 = \beta(501 : 1000),$

$\beta_1 = \beta_1(\text{seq}(1, 500, \text{by} = 10)), \beta_2 = \beta_2(\text{seq}(1, 500, \text{by} = 10))$

```
oldbeta=beta[seq(1,T,by=M)]
olks=ks

#dual chain KS:
#new ks
beta=beta[seq(1,T,by=M)]
ks=NULL
for (t in seq((T/(10*M)), (T/M), le=100))
  ks=c(ks, ks.test(beta[1:t], oldbeta[1:t])$p)
```



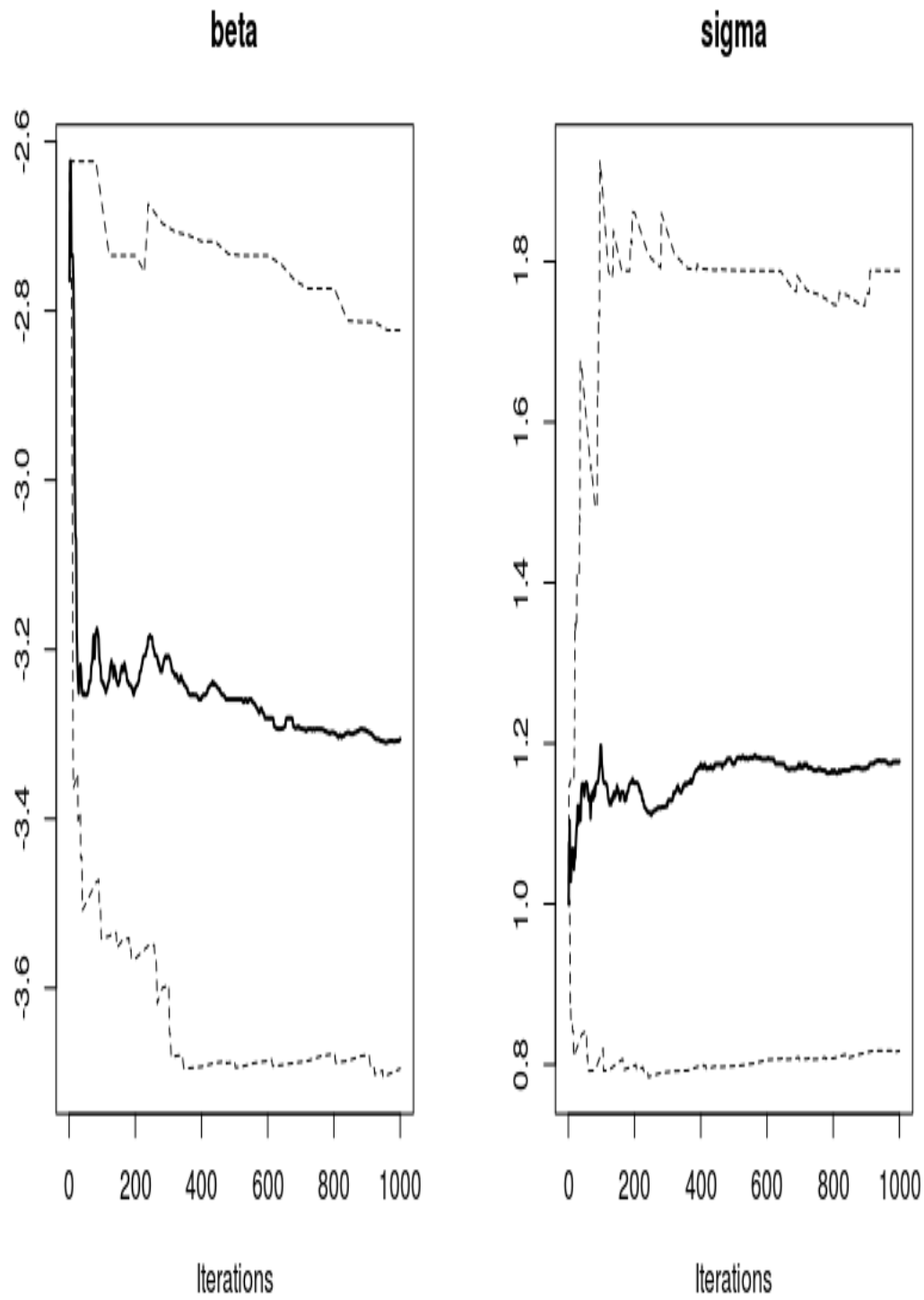


图 3: cumuplot

## 11 Monitoring Convergence to the Stationary Distribution

```
 $\beta = \beta(\text{seq}(1, 10000, \text{by} = 10))$   
 $t = \text{seq}(100, 1000, \text{by} = 9)$   
if  $t = 100, \beta(1 : 100), \text{old}\beta(1 : 100)$   
if  $t = 1000, \beta(1 : 1000), \text{old}\beta(1 : 1000)$ 
```

```
#figure  
par(mar=c(4,4,1,1), mfrow=c(2,1))  
plot(seq(1,T, le=100), olks, pch=19, cex=.7,  
      xlab="Iterations", ylab="p-value")  
plot(seq(1,T, le=100), ks, pch=19, cex=.7,  
      xlab="Iterations", ylab="p-value")
```

从图中可以看出，第一种 ks 计算方法，看不出是否平稳；第二幅图是比较了两条不同的链。图中说明了需要更长时间达到平稳。迭代 4000 次左右两个样本有类似的经验 cdf，之后又探索了不同的空间。

```
heidel.diag(mcmc(beta))  
geweke.diag(mcmc(beta))
```

在 sqar 中讲解。

- sqar.R

```
thn=jitter(xmc[seq(1,T,by=M)])  
kst=NULL  
for (m in seq(T/(10*M), T/M, le=100))  
  kst=c(kst, ks.test(thn[1:(m/2)],  
                    thn[(m/2)+(1:(m/2))])$p)
```

kscheck 中第一种方法。

```
hist(xmc, pro=T, col="grey85", nclass=150, main="",  
      ylab="", xlab="")  
ordin=apply(as.matrix(seq(min(xmc), max(xmc), le=200)),  
            1, eef)  
lines(seq(min(xmc), max(xmc), le=200),  
      ordin*max(density(xmc)$y)/max(ordin), lwd=2,  
      col="gold4")  
plot(seq(1,T, le=100), kst, pch=19, cex=.5,
```



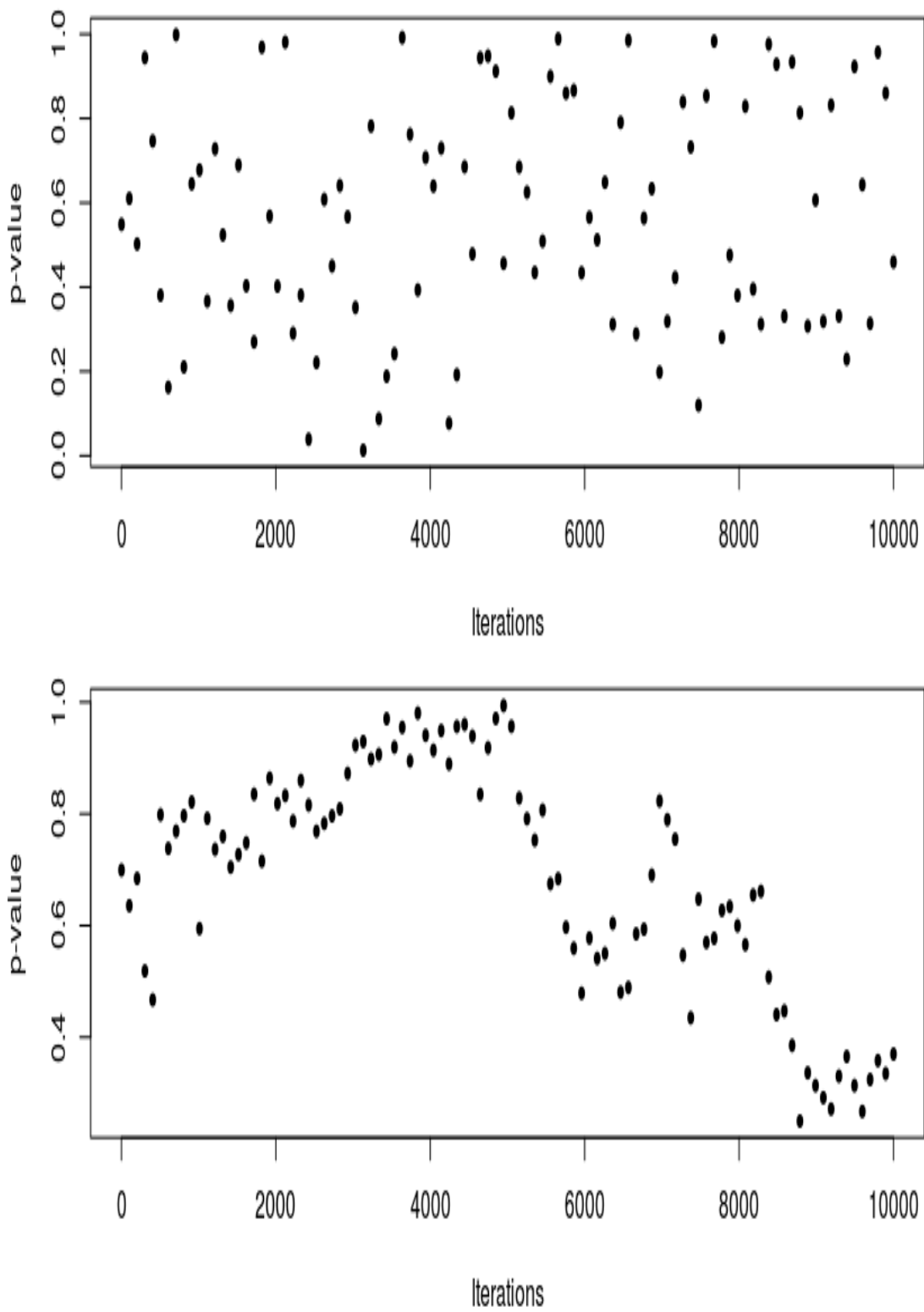


图 4: kscheck

```
xlab="Iterations",ylab="p-value")
```

画出直方图和分布, ks 统计量。

```
print(geweke.diag(mcmc(xmc)))
```

谱分析。统计量服从  $\mathcal{N}(0,1)$ 。返回值是 p-value。执行双边检验的假设检验。计算  $\Phi(|X| > p - value)$  与 significance level 进行比较。如果返回负值, 利用 geweke.diag 查看。

```
print(heidel.diag(mcmc(xmc)))
```

- sqar.R,multiple chains:

```
par(mfrow=c(3,2),mar=c(4,2,1,1))
for (t in 2:5)
  plot(smpl[,t],type="l",ylim=range(smpl),
       xlab="Iterations",ylab="",col=heat.colors(10)[6-t])

plot(smpl[,1],type="l",ylim=range(smpl),
     xlab="Iterations",ylab="",col=heat.colors(10)[5])
for (t in 2:5) lines(smpl[,t],col=heat.colors(10)[6-t])
```

```
par(mfrow=c(3,2),mar=c(4,2,1,1))
for (t in 1:5)
  plot(density(smpl[,t],n=1024),main="",ylab="",
       xlab=paste("Bandwith",format(density(smpl)$b,
       dig=3),sep=" "),lwd=2)
  plot(density(smpl,n=1024),main="",ylab="",
       xlab=paste("Bandwith",format(density(smpl)$b,
       dig=3),sep=" "),lwd=2)
```

```
par(mfrow=c(1,2),mar=c(4,2,1,1))
plot(smpl[,1],type="l",ylim=range(smpl),xlab="Iterations",
     ylab="",col=heat.colors(10)[5])
for (t in 2:5) lines(smpl[,t],col=heat.colors(10)[6-t])
plot(density(smpl,n=1024),main="",ylab="",
     xlab=paste("Bandwith",format(density(smpl)$b,dig=3),sep=" "),lwd=2)
```

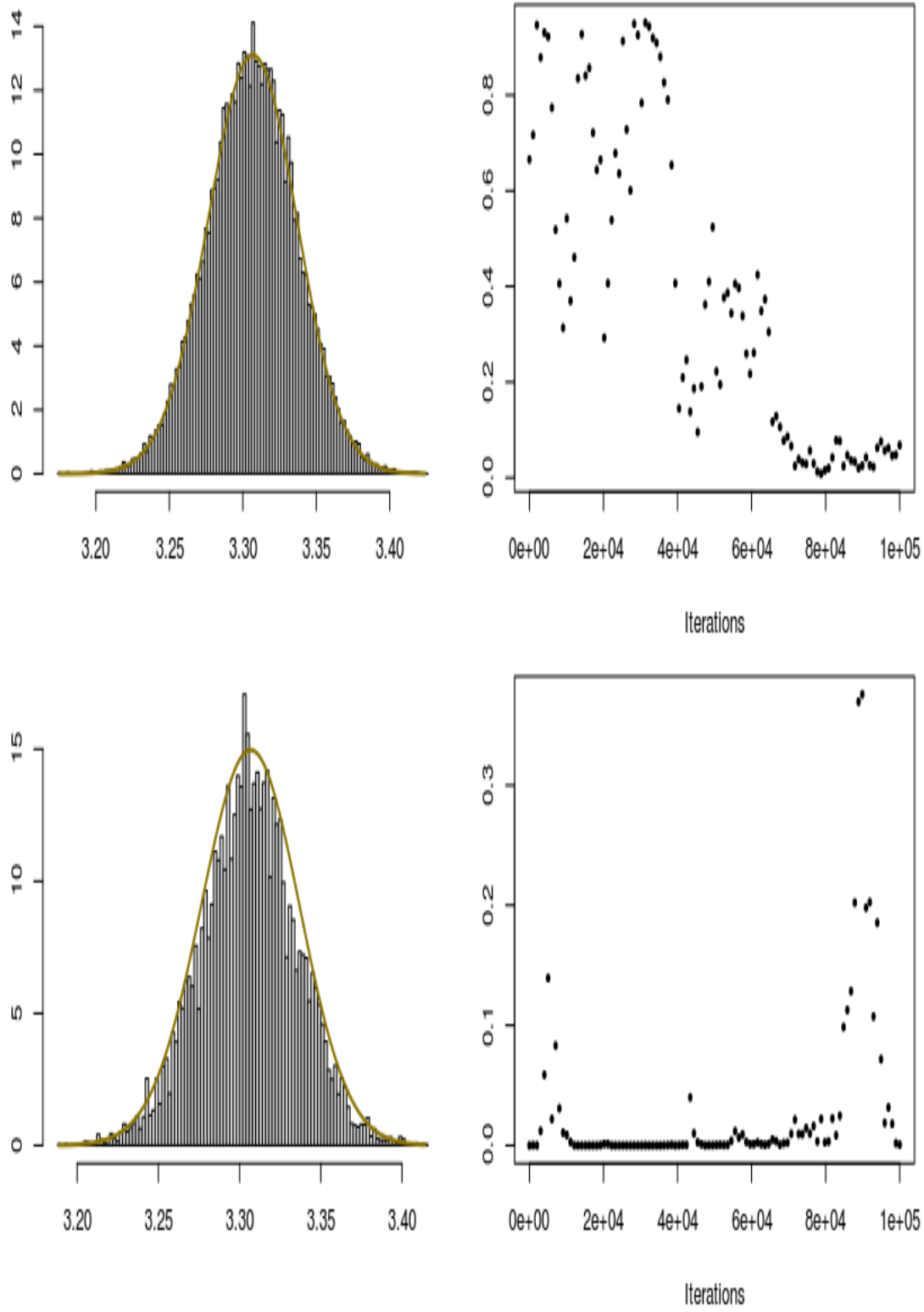


图 5: single chains

```
Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

var1
0.9636

Stationarity start      p-value
test      iteration
var1 passed      1      0.934

Halfwidth Mean Halfwidth
test
var1 passed      3.31 0.001
```

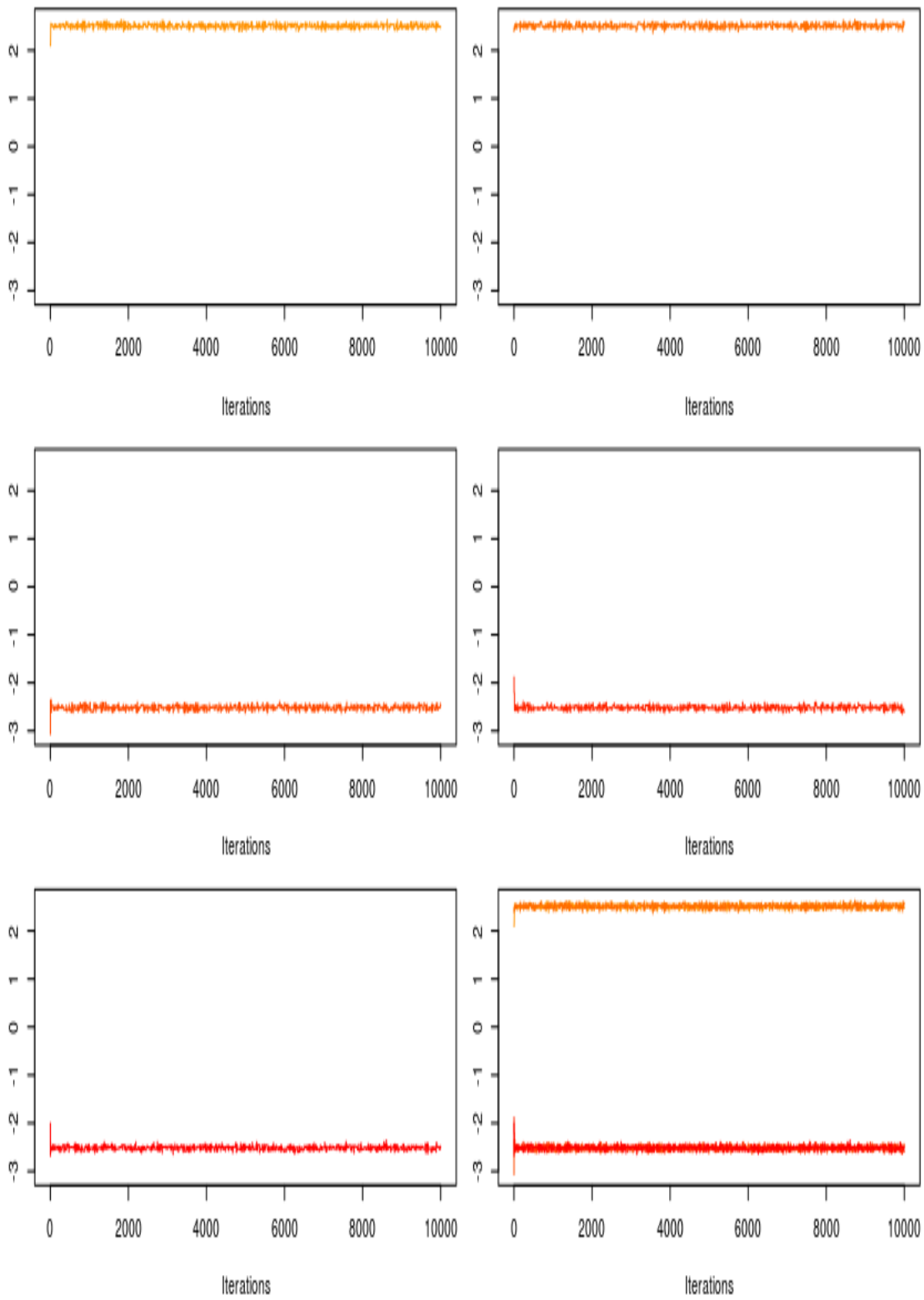
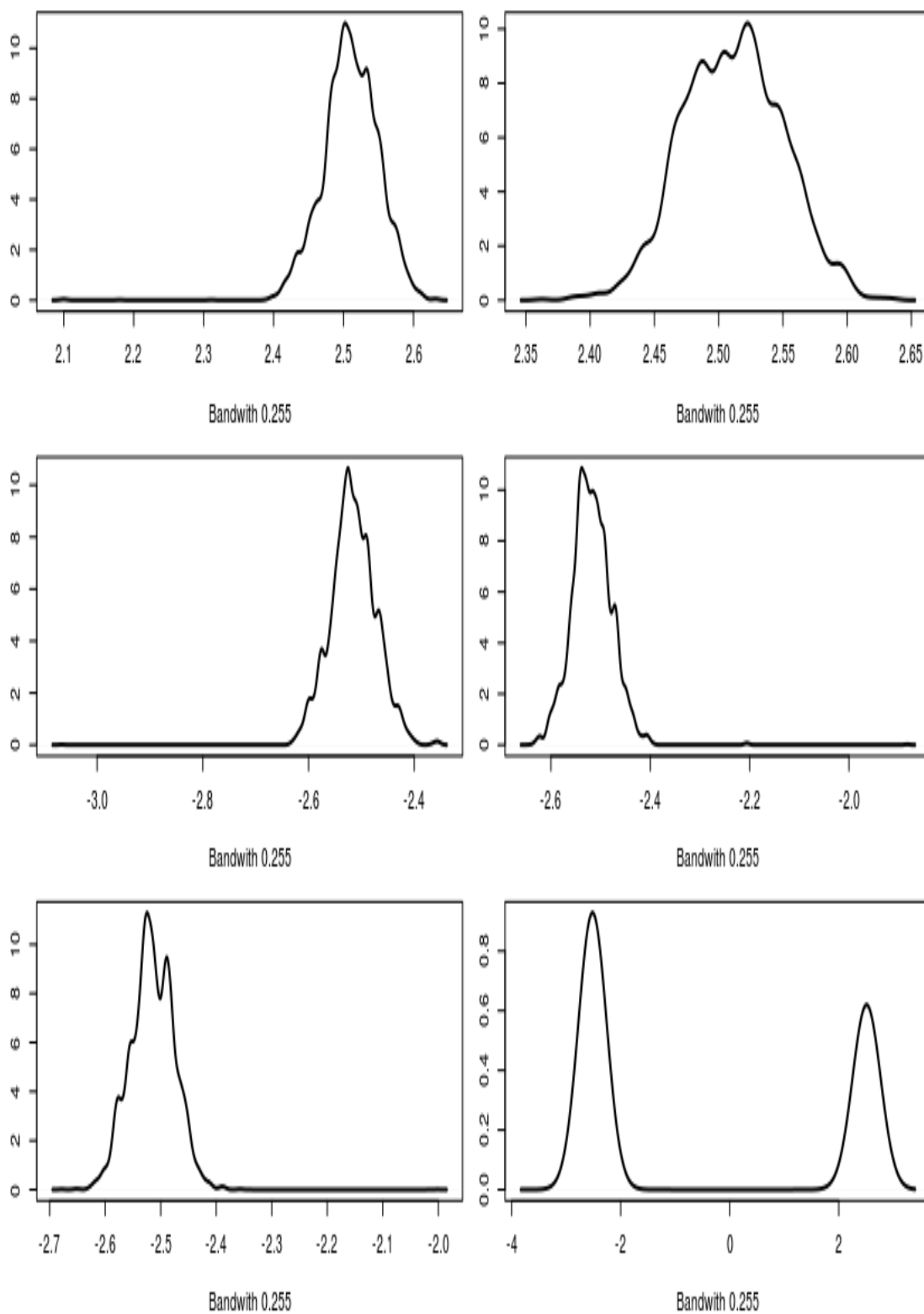


图 7: multiple chains



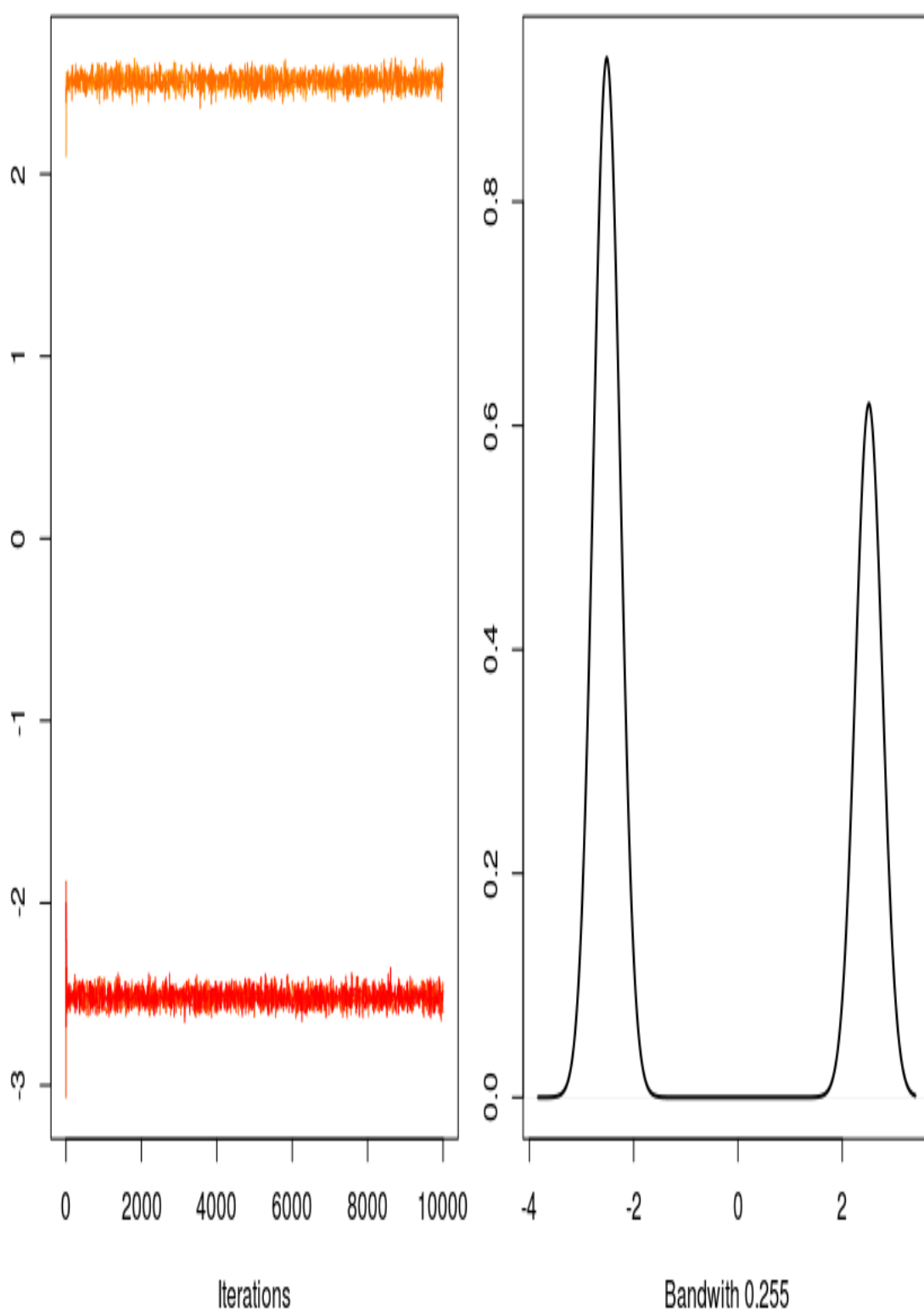


图 9: multiple chains

```
plot(mcmc.list(mcmc(smpl[,1]),mcmc(smpl[,2]),
mcmc(smpl[,3]),mcmc(smpl[,4]),mcmc(smpl[,5]))))
```

metropolis-hastings:

- hist:

```
ks.test(jitter(X),rbeta(5000,a,b))

par(mfrow=c(1,2),mar=c(2,2,1,1))
hist(X,nclass=150,col="grey",main="Metropolis-Hastings",freq=FALSE)
curve(dbeta(x,a,b),col="sienna",lwd=2,add=TRUE)
hist(rbeta(5000,a,b),nclass=150,col="grey",main="Direct Generation",freq=FALSE)
curve(dbeta(x,a,b),col="sienna",lwd=2,add=TRUE)
```

- acf

```
par(mfrow=c(2,2),mar=c(4,4,2,2))
hist(X1,col="grey",nclas=125,freq=FALSE,xlab="",main="Accept-Reject",xlim=c(0,1))
curve(dgamma(x,a,rate=1),lwd=2,add=TRUE)
hist(X2[2500:nsim],nclas=125,col="grey",freq=FALSE,xlab="",main="Metropolis-Hastings",xlim=c(0,1))
curve(dgamma(x,a,rate=1),lwd=2,add=TRUE)
acf(X1,lag.max=50,lwd=2,col="red") #Accept-Reject
acf(X2[2500:nsim],lag.max=50,lwd=2,col="blue") #Metropolis-Hastings
```

- ```
plot(cumsum(X<3)/(1:nsim),lwd=2,ty="l",ylim=c(.85,1),xlab="iterations",xlim=c(0,10000))
lines(cumsum(Z<3)/(1:nsim),lwd=2,col="sienna")
```

- mass

```
#Corresponding sequence of masses
mass=0*(1:2000)
that=thet[1]
for (i in 2:2000){
  that=sort(c(that,thet[i]))
  mass[i]=sum((that[2:i]-that[1:(i-1)])*f(that[1:(i-1)]))
}
```



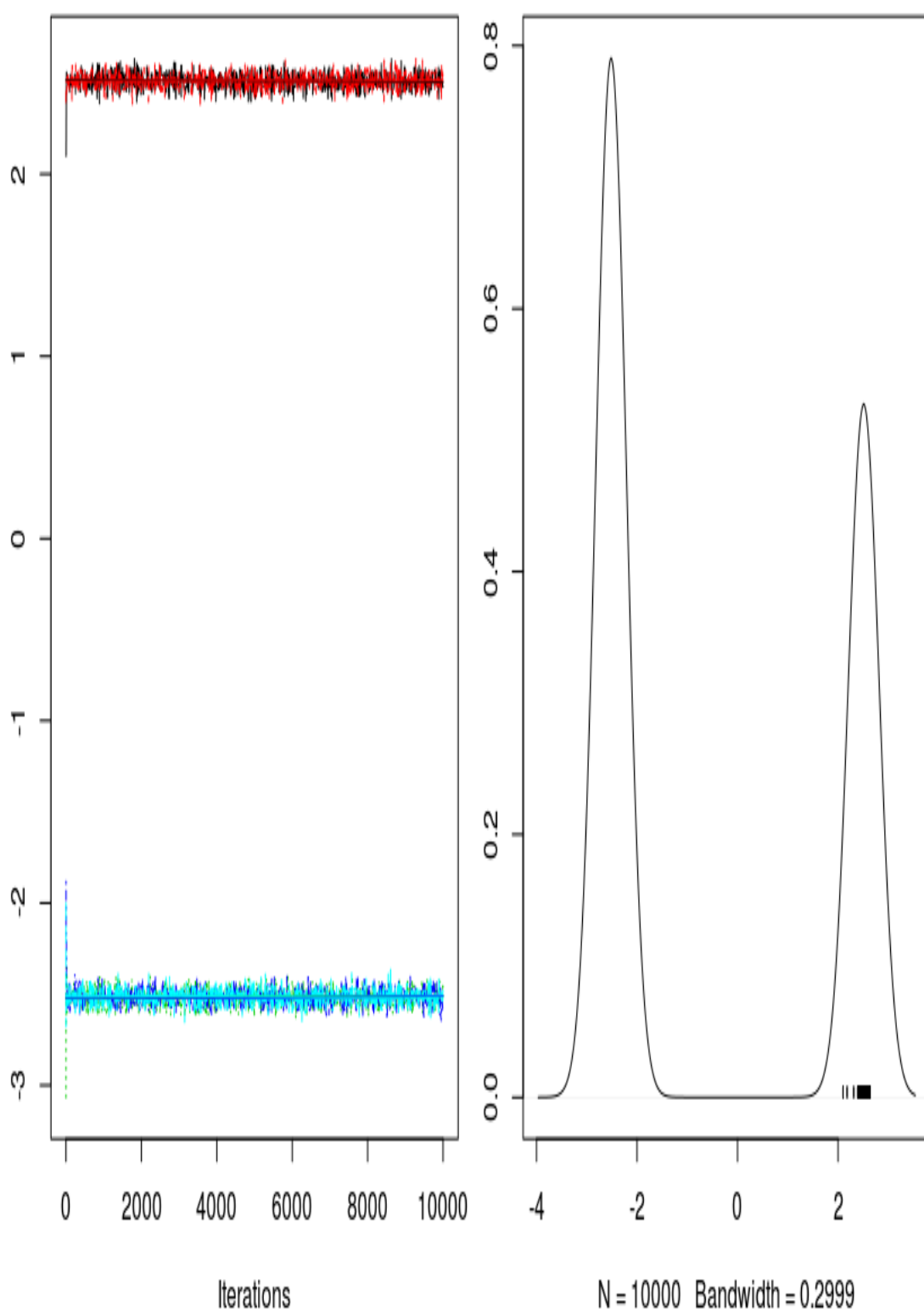
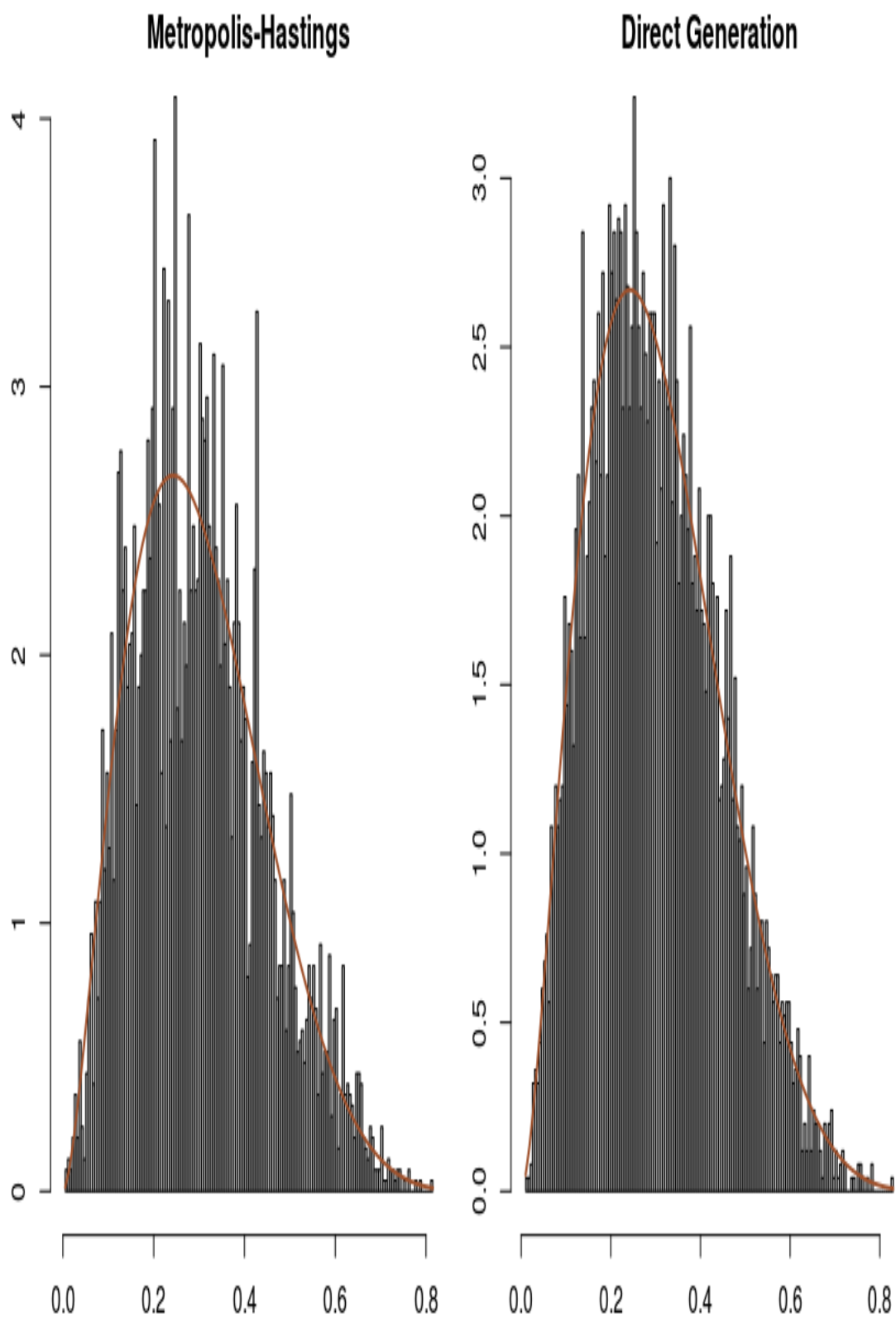


图 10: multiple chains



```
> ks.test(jitter(X),rbeta(5000,a,b))
```

Two-sample Kolmogorov-Smirnov test

data: jitter(X) and rbeta(5000, a, b)

D = 0.03, p-value = 0.02222

alternative hypothesis: two-sided

图 12: ks.test

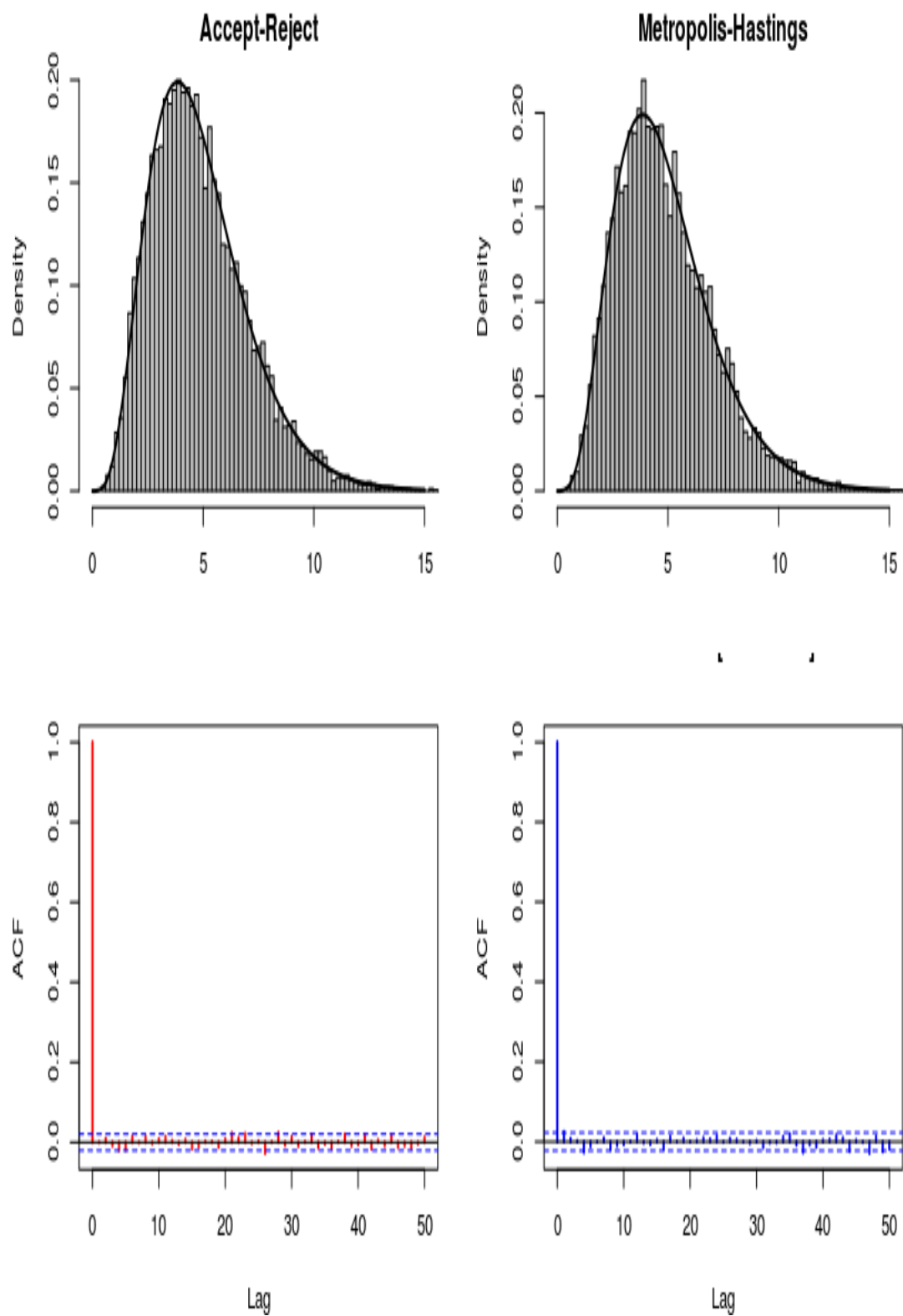


图 13: acf

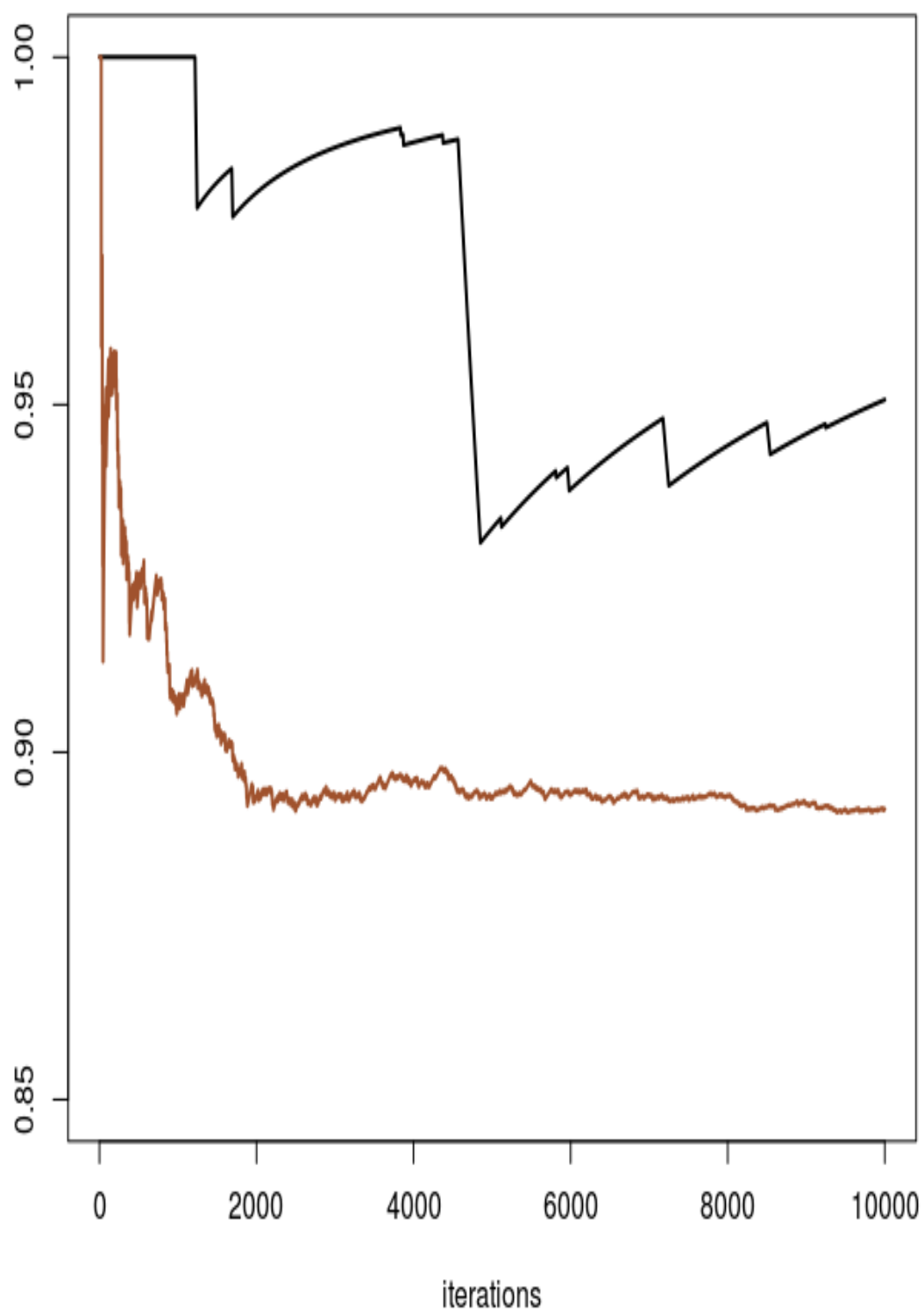


图 14: cumsum

## 11 Monitoring Convergence to the Stationary Distribution

```
# Plots  
plot(mass,type="l",ylab="mass",col="sienna4",lwd=2)  
par(new=T);plot(thet,pch=5,axes=F,cex=.3,col="steelblue2",ylab="")
```

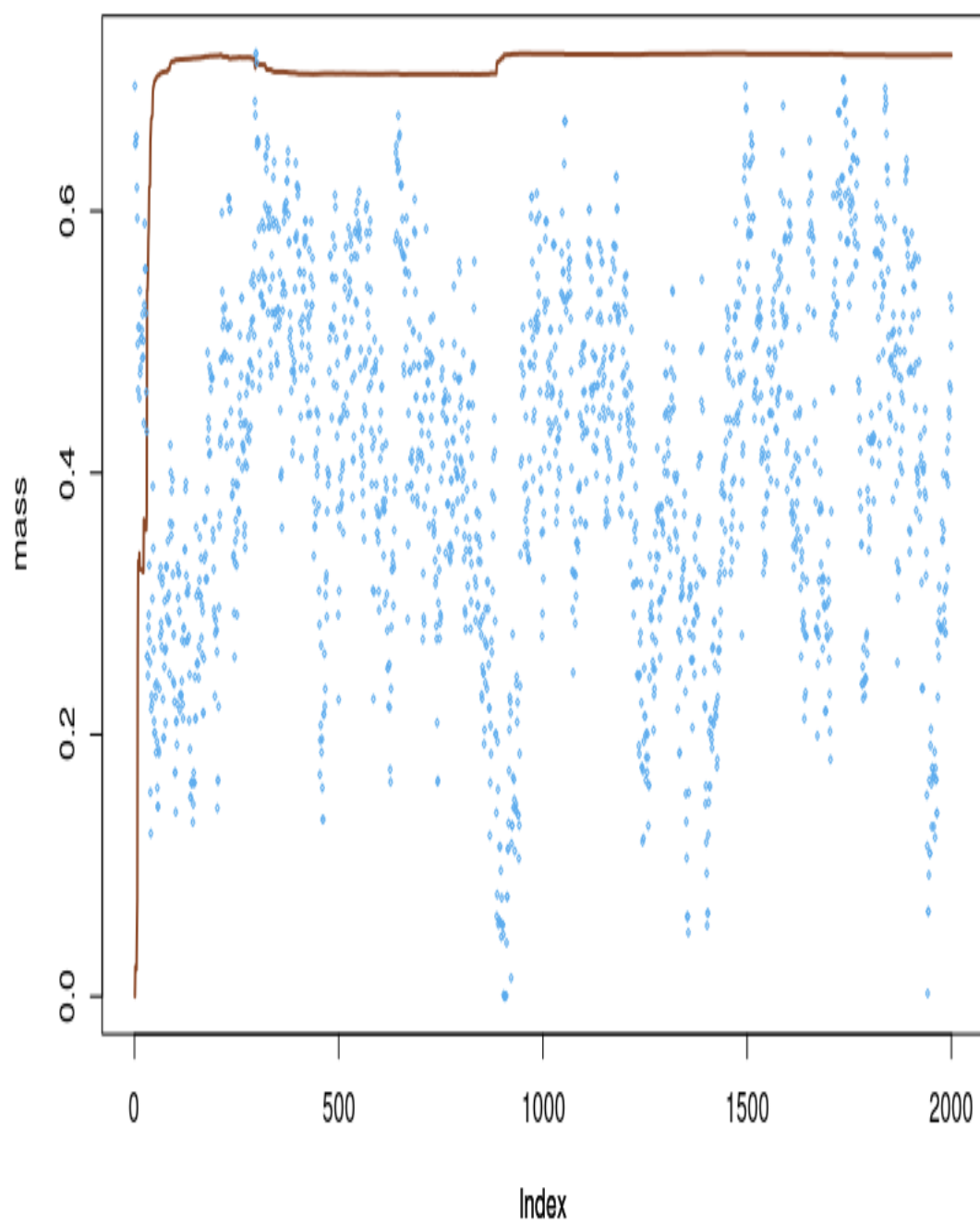


图 15: mass