第6章
贝叶斯统计计算方法

贝叶斯统计的基本理论和方法是简单易懂的，但是，当把贝叶斯统计运用到实际问题中时，由于后验分布的复杂性，往往无法得到它的解析表达式，从而得用随机模拟方法去估计有关的参数或做其它统计推断，而要做到这些，需要有一整套的理论和方法。其实，近三十年来，由于计算机的高速发展以及优良算法的发明，随机模拟方法不但在统计也在众多其它学科广泛运用，解决了许多经典方法难以解决的问题，从而愈来愈受到重视，是二十世纪发明的开创性方法，任何对统计感兴趣的人都应该对它有所了解，更何况我们统计专业的学生。本章就是对现在常用的随机模拟方法 MCMC 做一个实用性的概要介绍。

§6.1　什么是 MCMC 方法

6.1.1　蒙特卡罗法

蒙特卡罗原本只是欧洲历史上著名的赌城的城市名。当时蒙特卡罗的赌徒们为了赢得赌博，肯钻研爱学习，遇到不易解决的赌博上的概率统计问题，往往会请教数学家和概率统计学家，例如大名鼎鼎的拉普拉斯，从而客观上促进了概率统计的研究和发展，而蒙特卡罗也和概率统计沾上了关系。到了第二次世界大战时期，美国为了与德国竞赛研制原子弹，提出了著名的曼哈顿计划（Manhattan Project）。在研制原子弹过程中，遇到了复杂的有关核反应的计算问题，波兰裔美国数学家斯塔尼斯拉夫·乌拉姆（Stanislaw Ulam）创造性地提出了随机模拟方法，用来计算遇到的问题，并由另一位大名鼎鼎的科学家冯·纽曼（von Neumann）在计算机上实现。为了保密，就将该方法称为蒙特卡罗（Monte Carlo）法，因此蒙特卡罗就变成了随机模拟的意思。另一方面，一种革命性的尖端工具---电子计算机那时也在美国发明并制造出来。 这样，把蒙特卡罗方法与电子计算机的快速运算能力结合起来就迸发出巨大的能力，大大加快了原子弹的成功研制（也有人认为制造氢弹时才用上蒙特卡罗方法）。1945 年 8 月 6 日和 9 日，美国向日本的广岛和长崎各投下一颗原子弹，造成了日本巨大的人员伤亡和财产损失，加速迫使日本天皇宣布日本无条件投降。

蒙特卡罗方法的基本思想是模拟从总体抽取样本，然后，利用抽取的模拟样本进行估计、假设检验等统计推断。模拟的实施主要在计算机上来进行，所以蒙特卡罗方法的兴起是和计算机的发展密切相关的，现代计算机的高速计算能力和优良的抽样算法使得大多数模拟抽样可以轻而易举的实现，从而使蒙特卡罗方法发展成为现代及其重要且应用广泛的统计方法。

Monte Carlo, Monaco

Casino in Monte Carlo City

例 6.1 分别模拟容量 200 的服从二项分布 *Bin*(10, 0.6) 的样本和服从泊松分布 *Poisson*(0.6) 的样本，然后分别算出样本均值并与总体均值进行比较。

解：（1）二项分布 *Bin*(10, 0.6) 表示做 10 次的独立试验而且每次的成功概率是 0.6。产生模拟样本所用 R 命令如下

```
rb<-rbinom(n=200,size=10,prob=0.6)
  mean(rb)
[1] 6.005
```

这里 R 命令中参变量的意义一望可知，不需解释。样本均值为 6.005 与总体均值$10\times0.6=6$非常接近。

（2）泊松分布 *Poisson*(0.6) 表示其均值和方差是 0.6。产生模拟样本所用 R 命令如下

```
rp<-rpois(n=200,lambda=0.6)
mean(rp); var(rp)
[1] 0.595
[1] 0.603995
```

这里样本均值为 0.595 与总体均值 0.6；样本方差与总体方差同样非常接近。

例 6.2 分别模拟容量 500 的服从均匀分布$U(0,1)$、正态分布$N(1.2,25)$和贝塔分布$Beta(1.5,2)$的样本，然后分别算出样本均值并与总体均值进行比较。

解：（1）产生容量 500 的服从均匀分布$U(0,1)$的模拟样本所用 R 命令如下

```
ru<-runif(n=500,min=0,max=1)
mean(ru)
[1] 0.4979093
```

这里样本均值为 0.4979 与总体均值 0.5 非常接近。

（2）产生容量 500 的服从正态分布$N(1.2,25)$的模拟样本所用 R 命令如下

```
rn<-rnorm(n=500,mean=1.2, sd=5)
mean(rn)
[1] 0.9525769
```

这里样本均值为 0.9526 与总体均值 1.2 有较大差距。

（3）产生容量 500 的服从贝塔分布 $Beta(1.5,2)$ 的样本所用 R 命令如下

```
rbet<-rbeta(n=500, shape1=1.5, shape2=2)
  mean(rbet)
[1] 0.4389338
```

这里样本均值为 0.4389 与总体均值 1.5/(1.5+2)=0.4286 非常接近。

注：在（2）中，由容量 500 的服从正态分布 $N(1.2,25)$ 的模拟样本计算所得的样本均值 0.9526 与总体均值 1.2 有较大差距，这并不奇怪，主要原因是样本是随机的且容量不够大。事实上，根据强大数定律（见定理 6.1），当样本容量趋近于无穷大时，样本均值几乎必然收敛于总体均值。现在我们让样本容量增大到 10000，再看看结果如何。

```
  rn1<-rnorm(n=10000,mean=1.2, sd=5); rn2<-rnorm(n=10000,mean=1.2, sd=5)
  mean(rn1); mean(rn2)
[1] 1.174578
[1] 1.136559
```

我们看到这时样本均值与总体均值就比较接近了。读者还可以让样本容量增大到 100000 来看看结果如何。对于计算机来说，这个计算量是小菜一碟。

虽然这里我们用模拟样本的均值只是估计了总体均值，但在连续情形，总体均值是由定积分来计算的，例如，贝塔分布 $Beta(1.5,2)$ 的均值

$$E(\theta)=\frac{\Gamma(3.5)}{\Gamma(1.5)\Gamma(2)}\int_0^1\theta\theta^{1.5-1}(1-\theta)^{2-1}d\theta$$

因此，我们也把一个定积分估计出来了，即

$$\frac{\Gamma(3.5)}{\Gamma(1.5)\Gamma(2)}\int_0^1\theta^{1.5}(1-\theta)^{2-1}d\theta=E(\theta)\approx\bar{\theta}=0.4389$$

这其实就是蒙特卡罗方法的基本思想，其理论基础之一是强大数定律。

定理 6.1（强大数定律）设 $X_1,X_2,\cdots$ 是独立同分布的随机序列，存在绝对值期望 $\mathrm{E}|X_t|<\infty$，又设 $\bar{X}_T=(1/T)\sum_{t=1}^{T}X_t$，$\mathrm{E}(X_t)=\mu$，那么

$$\lim_{T\to\infty}\bar{X}_T=\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}X_t=E(X_t)=\mu\ \ a.s.$$

其中 a.s.=almost surely 表示几乎必然成立，即以概率 1 成立。

一般地，设 $\pi(x)$ 是随机变量 X 的概率密度函数，$h(x)$ 是任意但我们感兴趣的可积函数。考虑期望（定积分）

$$E^{\pi}[h(X)]=\int_{\chi}h(x)\pi(x)dx$$

的估计问题。如果我们能够从概率密度函数 $\pi(x)$ 抽取独立同分布的样本 $(X_1,X_2,\cdots,X_T)$，那么由强大数定律，均值

$$\bar{h}_T=\frac{1}{T}\sum_{i=1}^{T}h(X_i)$$

几乎必然收敛于 $E^{\pi}[h(X)]$。换句话说，只要样本容量足够大，就可以用 $\bar{h}_T$ 估计期望（定积分）$E^{\pi}[h(X)]$。另外，如果 $h(X)$ 的方差 $Var[h(X)]$ 存在，就可以用

$$s_T^2=\frac{1}{T-1}\sum_{i=1}^{T}[h(x_i)-\bar{h}_T]^2$$

来估计之，而且根据中心极限定理，$\sqrt{T}\left[\bar{h}_T-E^{\pi}[h(X)]\right]\Big/s_T$ 渐近服从标准正态分布 $N(0,1)$，从而可以构造出相应的置信区间并对估计量 $\bar{h}_T$ 做检验，同时估计量 $\bar{h}_T$ 的标准误可以由下式估计

$$se_{\bar{h}_T}=\sqrt{\frac{1}{T^2}\sum_{i=1}^{T}Var[h(X_i)]}=\sqrt{\frac{1}{T}Var[h(X)]}=\sqrt{\frac{1}{T(T-1)}\sum_{i=1}^{T}[h(x_i)-\bar{h}_T]^2}$$

以上所述就是经典蒙特卡罗方法的原理，估计量 $\bar{h}_T$ 也称为蒙特卡罗估计量。

显然，蒙特卡罗方法关键的一步是能够从概率密度函数 $\pi(x)$ 抽取独立同分布的样本，但不幸的是在许多情形下无法容易地从 $\pi(x)$ 抽取独立同分布的样本。这时一个变通的方法是寻找一个容易抽样的密度函数 $g(x)$，它满足当 $h(x)\pi(x)\neq 0$ 时 $g(x)>0$，于是

$$E^{\pi}[h(X)]=\int_{\chi}\frac{h(x)\pi(x)}{g(x)}g(x)dx=E^{g}[\frac{h(X)\pi(X)}{g(X)}]$$

从而就可以估计出 $E^{\pi}[h(X)]$ 为

$$E^{\pi}[h(X)]=E^{g}[\frac{h(X)\pi(X)}{g(X)}]\approx\frac{1}{T}\sum_{i=1}^{T}\frac{h(x_i^g)\pi(x_i^g)}{g(x_i^g)}$$

其中 $(x_1^g,x_2^g,\cdots,x_T^g)$ 是抽取自密度函数 $g(x)$ 的独立同分布样本。我们称密度函数 $g(x)$ 为重要性函数，这种方法为重要性抽样法。

应用经典蒙特卡罗法和重要性抽样法虽然可以解决不少期望（定积分）的估计问题，但仍然还有许多定积分的估计问题它们解决不了，因为它们要求待抽样的密度函数具有完全已知的解析表达式，但是，待抽样的密度函数（称为目标分布（密度）），以贝叶斯统计的后验分布 $\pi(\theta|\mathbf{x})$ 为例，往往没有完全的解析表达式，这样就不能用以上所讨论的抽样方法去直接抽样进而估计后验期望和进行别的统计推断，而必须去寻找新的方法。在 1953 年，物理学家梅切波利斯（Metropolis，1953）等学者从粒子物理的计算问题得到启发，发明了一套算法，可以得到具有马尔可夫性而且近似于来自待抽样的密度函数的样本，从而使得大量极其复杂的定积分的估计问题得到解决。为了理解这套算法，我们首先要知道什么是马尔可夫性和马尔可夫链。

6.1.2 马尔可夫链

马尔可夫链（Markov chain，简称为马氏链）是一种具有马尔可夫性（简称为马氏性）的特别随机过程，它的取值空间称为状态空间 S，为了简单和易于理解起见，我们假设状态空间 S 中的元素是可数的。马氏链的正式定义如下：

定义 6.1 一列有序随机变量 $\{X_t\}$ 称为马氏链，如果已知现在 X_t，过去 $\{X_i;\ 0 \le i \le t-1\}$ 与将来 X_{t+1} 相互独立。这个性质称为马氏性，用公式表示为

$$\begin{aligned} &P(X_{t+1}=s_{t+1}, X_{t-1}=s_{t-1}, \cdots, X_0=s_0 \mid X_t=s_t) \\ &=P(X_{t+1}=s_{t+1} \mid X_t=s_t)P(X_{t-1}=s_{t-1}, \cdots, X_0=s_0 \mid X_t=s_t) \end{aligned}$$

注：以上公式等价于

$$P(X_{t+1}=s_{t+1} \mid X_t=s_t, X_{t-1}=s_{t-1}, \cdots, X_0=s_0)=P(X_{t+1}=s_{t+1} \mid X_t=s_t)$$

即将来 X_{t+1} 只依赖于现在 X_t，而不依赖于过去 $\{X_i;\ 0 \le i \le t-1\}$。

定义 6.2 马氏链$\{X_t\}$称为时间齐次的（简称为时齐的），如果对任何时间点t，任何两个状态$i, j \in S$，有

$$P(X_{t+1} = j \mid X_t = i) = P(X_1 = j \mid X_0 = i) = p_{ij}$$

并且称 p_{ij} 为状态i到状态 j 的一步转移概率（转移核，马氏核），$\mathbf{P} = (p_{ij})$为转移概率矩阵。另外，$p_{ij}(t) = P(X_t = j \mid X_0 = i)$称为$t$步转移概率。

时齐马氏链$\{X_t\}$具有如下性质，其任意有限维分布可由转移概率和初始（值）分布表示：

$$\begin{aligned}
&P(X_t = s_t, X_{t-1} = s_{t-1}, \cdots, X_0 = s_0) \\
&= P(X_t = s_t \mid X_{t-1} = s_{t-1}, \cdots, X_0 = s_0) P(X_{t-1} = s_{t-1}, \cdots, X_0 = s_0) \\
&= p_{s_{t-1}s_t} P(X_{t-1} = s_{t-1}, \cdots, X_0 = s_0) \\
&= p_{s_{t-1}s_t} \cdots p_{s_0 s_1} P(X_0 = s_0)
\end{aligned}$$

这就是说，只要给定转移概率（矩阵）和初始分布，那么时齐马氏链的统计规律也就确定了。以下总是设马氏链$\{X_t\}$是时齐马氏链。

例 6.3 随机序列$\{X_t\}$的状态空间为整数集，而且对任时间t满足

$$P(X_{t+1}=i-1\,|\,X_t=i)=p,\ P(X_{t+1}=i+1\,|\,X_t=i)=q$$

其中$0<p<1$，$q=1-p$，则它是一个时齐马氏链。

解： 易知对任时间t有

$$P(X_{t+1}=j\,|\,X_t=i)=p_{ij}=\begin{cases}p, j=i-1\\ q, j=i+1\\ 0, \text{其余情形}\end{cases}$$

因上式右端与时间t无关，故$\{X_t\}$是时齐马氏链，同时，它的一步转移概率为p_{ij}。

例 6.4 给定初始值θ_1，通过转移核$p(\theta\,|\,\theta_t)$产生的随机序列$\{\theta_t\}$就是一个时齐马氏链。例如，设$\theta\,|\,\theta_t\sim N(0.6\theta_t,4)$，初始值$\theta_1=20$，那么就产生了一个时齐马氏链（模拟样本），再让$\theta_1=-20$就又产生了另一个时齐马氏链（模拟样本）。从图 6.1（称为马氏链的轨迹图）可以看出，虽然初始值（出发点）完全不同，但经过几次迭代，它们就没有什么区别了，并且趋于平稳。这里所用的 R 程序如下：

```
theta=c();theta[1]=20
for(t in 2:1000){theta[t]=rnorm(1, mean=0.6*theta[t-1],sd=2)}    # for 是迭代语句
plot(ts(theta),xlab="迭代次数", ylim=c(-20,20))       #参变量 ylim 表示 y 轴的取值范围
theta[1]=-20
for(t in 2:1000){theta[t]=rnorm(1, mean=0.6*theta[t-1],sd=2)}
lines(1:1000, theta, col="blue")   #命令 lines 是把第二个链的图形附加到第一个链的图形中
```

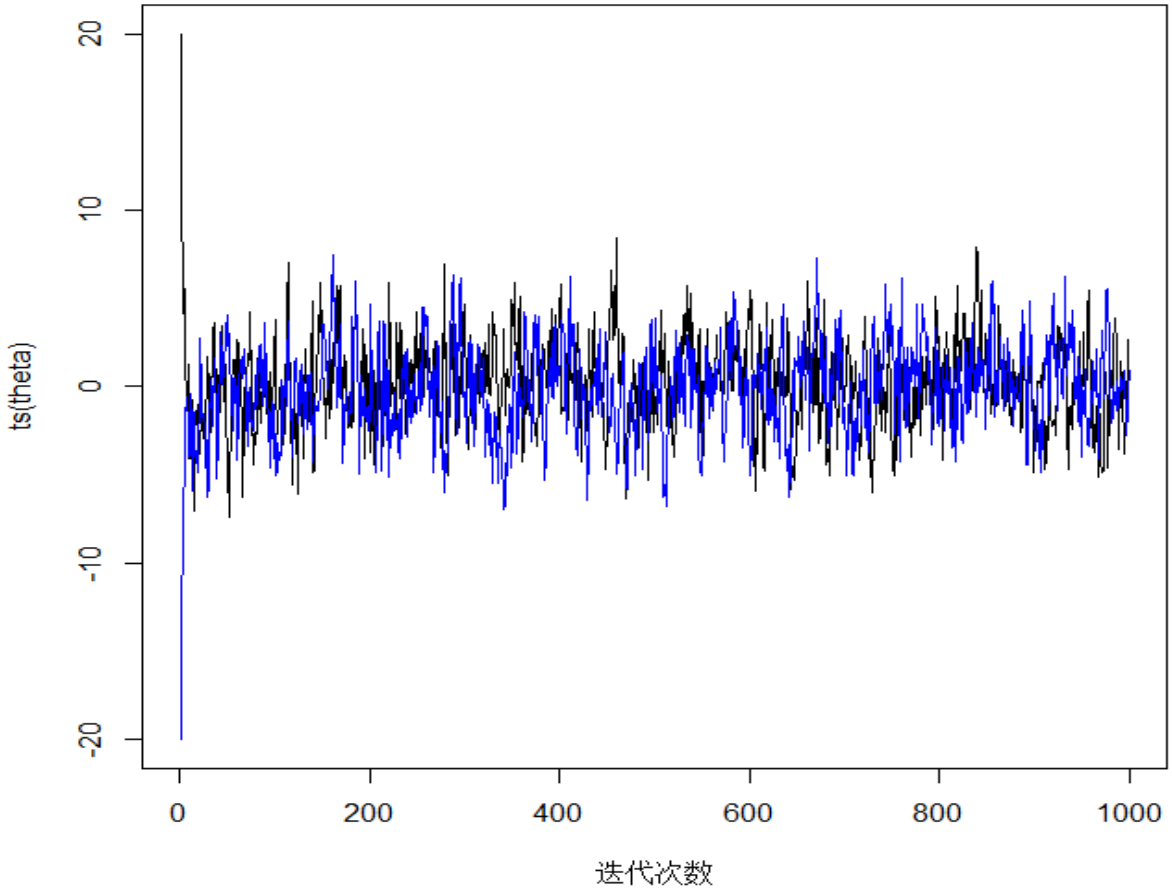

图 6.1 时齐马氏链图

定义 6.3（不可约性）马氏链$\{X_t\}$称为不可约的，如果对任何两个状态$i, j \in S$存在时点$t > 0$使得t步转移概率$p_{ij}(t) > 0$。换句话说，不可约性就是从任意一个状态出发总可以到达任意的另一个状态。

定义 6.4（非周期性）称状态i是非周期的，如果

$$\gcd\{t; P(X_t = i \mid X_0 = i) > 0\} = 1$$

其中 gcd 表示最大公约数。称马氏链$\{X_t\}$为非周期的，如果它的所有状态都是非周期的。

定义 6.5（正常返性）设$X_0 = i, i \in S$，$T_i = \min\{t \geq 1; X_n = i\}$。如果概率$P(T_i < \infty) = 1$，称状态$i$是常返的。如果$E(T_i) < \infty$，称状态$i$是正常返的。称马氏链$\{X_t\}$为正常返的，如果它的所有状态都是正常返的。

注： $T_i = \min\{t \geq 1; X_t = i\}$表示首次返回状态$i$的时间。

定义 6.6（遍历性）称马氏链的状态i是遍历的，如果它是非周期且正常返的。称马氏链$\{X_t\}$是遍历的，如果它是不可约的而且所有状态是遍历的。

定义 6.7 状态空间S上的分布$\boldsymbol{\pi} = \{\pi_j\}$称为马氏链$\{X_t\}$的平稳分布（不变分布），如果由它是初始值$X_0$的分布可以推出对任意时点$t$，它也是$X_t$的分布。

注：具有平稳分布的马氏链本身显然是（严）平稳的。

以下几个定理是 MCMC 的理论基础。如果读者有兴趣，它们的证明可参考有关马氏链的专著。而有了这些准备，我们就可以讨论 MCMC 本身了。

定理 6.2 如果马氏链$\{X_t\}$是不可约、非周期和正常返的（即遍历的），那么它具有唯一的平稳分布$\boldsymbol{\pi}=\{\pi_j\}$满足对于任意状态$i, j \in S$

$$\lim_{t\to\infty} p_{ij}(t) = \lim_{t\to\infty} P(X_t = j \mid X_0 = i) = \pi_j$$

并且是方程$\boldsymbol{\pi}\mathbf{P}=\boldsymbol{\pi}, \sum_{j=0}^{\infty}\pi_j = 1$唯一的非负解。

定义 6.8 如果对于任意状态$i, j \in S$和时间t，存在常数$\rho<1$和$C_{ij}<\infty$使得$\left|p_{ij}(t)-\pi_j\right| \le C_{ij}\rho^t$，则称马氏链$\{X_t\}$是几何遍历的，如果还有$\sup\{C_{ij}\}<\infty$，则称马氏链$\{X_t\}$是一致几何遍历的。

定理 6.3（马氏链强大数定律）如果马氏链$\{X_t\}$是不可约的且具有唯一的平稳分布$\boldsymbol{\pi}$，随机变量$X \sim \boldsymbol{\pi}$，函数$h(x)$满足$E^{\pi}|h(X)|<\infty$，那么

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} h(X_t) = E^{\pi}[h(X)] = \int_{\chi} h(x)\pi(\mathrm{x})\mathrm{dx} \quad a.s.$$

定理 6.4（马氏链中心极限定理）如果马氏链$\{X_t\}$是一致几何遍历的，则

$$\sqrt{T}\left[\frac{\bar{h}-E^{\pi}[h(X)]}{\sqrt{Var[h(X)]}}\right] \xrightarrow{d} \mathrm{N}(0,1) \quad （依分布收敛）$$

其中，$\bar{h}=\sum_{i=1}^{T} h(X_i)/T$，而$Var[h(X)]$可用$\bar{h}$的方差$\sigma_{\bar{h}}^2 = \frac{\sigma_h}{T}[1+2\sum_{i=1}^{\infty}\rho(h)]$代替。

6.1.3 马氏链蒙特卡罗法

在前面谈到，为了寻找新的估计复杂定积分方法，在 1953 年，物理学家梅切波利斯（Metropolis，1953）等学者从粒子物理的优化计算问题得到启发，发明了一套算法。现在从定理 6.2，我们看到只要抽取的样本是满足一定条件的马尔可夫链（不要求独立!），那么就可以用它来估计期望（定积分），从而使得大量极其复杂的定积分的估计问题得到解决。这套算法开始时叫做梅切波利斯算法，后来，哈斯廷斯（Hastings，1970）意识到这个算法的重大意义并将它进行了一般化。现在人们普遍把它称为梅切波利斯-哈斯廷斯算法（简称为 MH 算法），并把它誉为二十世纪最重要的十大算法之一。

在 1984 年，吉曼（Geman，1984）兄弟在研究数字图像恢复问题时提出了吉布斯抽样法（Gibbs sampling）并给予该名称，虽然它的来源不同，但可以看成 MH 算法的一个特例，同时它也有独立存在的意义，因为它实现的形式是不同的，而且特别适用于处理失缺数据问题、高维定积分和潜变量模型等等。类似的方法， 除了 MH 算法和吉布斯抽样法之外，还有模拟退火法（simulated annealing），这是由应用数学家提出来的。现在人们将这三种方法和各种各样的推广统称为马氏链蒙特卡罗（MCMC）法。

比较蹊跷的一件事是国际统计学界在长达三十多年的时间里，对于马氏链蒙特卡罗法几乎无动于衷，没有意识到它对于统计学的开创性的重要性，直到1990年，盖尔芳德和史密斯（Gelfand and Smith，1990）才使统计学界开始认识到马氏链蒙特卡罗法在统计计算中的威力，并引发了大量的理论研究和实际应用研究，取得了许多重要成果，为统计学界挽回了一点面子。

§6.2 吉布斯抽样

上一节扼要讨论了马氏链蒙特卡罗法的发展历史以及马氏链的基本概念和性质。本节讨论吉布斯抽样（Gibbs sampling）的具体方法。一般而言，一元分布的抽样当然比多元分布的抽样容易，而吉布斯抽样的特点是可以通过来自目标分布的一元（或较低维）分布的抽样来获得多元分布本身的样本，因此很适合用于高维问题的场合。另外，与MH算法相比，吉布斯抽样的结果不会被拒绝（参看6.3节）。

6.2.1 二阶段吉布斯抽样

为了初学者更好地理解吉布斯抽样，我们从二维的情形讨论起。设随机向量 $\mathbf{X}=(X_1,X_2)\sim\pi(x_1,x_2)$（分布密度或概率函数），那么我们有两个边际分布密度（按照习惯积分区域略去不写）

$$\pi_1(x_1)=\int\pi(x_1,x_2)dx_2\,,\quad \pi_2(x_2)=\int\pi(x_1,x_2)dx_1$$

和两个条件分布密度（称为满条件分布密度）

$$\pi(x_1\mid x_2)=\frac{\pi(x_1,x_2)}{\pi_2(x_2)}\,,\quad \pi(x_2\mid x_1)=\frac{\pi(x_1,x_2)}{\pi_1(x_1)}$$

我们称如下抽样为二阶段吉布斯抽样：

1. 给定初始值 $\mathbf{x}^{(0)}=(x_1^{(0)},x_2^{(0)})$（其实只要给定 $x_1^{(0)}$ 或 $x_2^{(0)}$）
2. 对于 $t=1,2,\cdots,T$，产生样本 $\mathbf{x}^{(t)}=(x_1^{(t)},x_2^{(t)})$，做法是

（1）从条件密度 $\pi(x_1\mid x_2^{(t-1)})$ 抽取 $x_1^{(t)}$；

（2）从条件密度 $\pi(x_2\mid x_1^{(t)})$ 抽取 $x_2^{(t)}$

3. 对于 $t+1$，回到第2步。

这样依次抽取就得到一个样本 $\{\mathbf{x}^{(t)}=(x_1^{(t)},x_2^{(t)});1\le t\le T\}$，从抽样的过程可以看出，如果已知现在 $\mathbf{x}^{(t)}$，则将来 $\mathbf{x}^{(t+1)}$ 与过去 $\{\mathbf{x}^{(i)};i<t\}$ 无关，因此它是马氏链，而且在一定的条件下，其平稳分布就是 $\pi(x_1,x_2)$。不仅如此，还可以证明两个分量样本序列 $\{x_1^{(t)};1\le t\le T\}$ 和 $\{x_2^{(t)};1\le t\le T\}$ 是分别具有平稳分布

$$\pi_1(x_1)=\int\pi(x_1,x_2)dx_2,\quad \pi_2(x_2)=\int\pi(x_1,x_2)dx_1$$

的马氏链。另外，使人惊奇的是两个一维的满条件分布 $\pi(x_1\mid x_2)$ 和 $\pi(x_2\mid x_1)$ 合在一起就包含了联合发布 $\pi(x_1,x_2)$ 的全部信息，即后者可由前两者表示出来

$$\pi(x_1,x_2)=\frac{\pi(x_2\mid x_1)}{\int[\pi(x_2\mid x_1)/\pi(x_1\mid x_2)]dx_2}$$

事实上

$$\int[\pi(x_2\mid x_1)/\pi(x_1\mid x_2)]dx_2=\int\frac{\pi(x_1,x_2)}{\pi_1(x_1)}\times\frac{\pi_2(x_2)}{\pi(x_1,x_2)}dx_2=\int\frac{\pi_2(x_2)}{\pi_1(x_1)}dx_2=\frac{1}{\pi_1(x_1)}$$

例 6.5 利用吉布斯抽样法产生来自二元正态分布 $N(\mu_1,\sigma_1^2;\mu_2,\sigma_2^2;\rho)$ 的马氏链样本 $X^{(t)}=(X_1^{(t)},X_2^{(t)})$，要求参数为 $(1.1,3^2;1.8,4^2;0.6)$，样本量为 5000。然后，考察样本服从的分布是什么。最后，利用样本计算五个参数 $(\mu_1,\sigma_1^2;\mu_2,\sigma_2^2;\rho)$ 的估计值并与参数真值进行比较。

解： 二元正态分布密度函数为

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\times e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]}$$

不难证明其两个边际分布也是正态分布，两个满条件分布密度分别是

$$\pi(x_1\mid x_2)\sim N(\mu_1+\rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2),(1-\rho^2)\sigma_1^2)\ \pi(x_2\mid x_1)\sim N(\mu_2+\rho\frac{\sigma_2}{\sigma_1}(x_1-\mu_1),(1-\rho^2)\sigma_2^2)$$

因此，利用吉布斯抽样法就能将二元抽样化为一元抽样。本例的吉布斯抽样以及相关的 R 命令如下(各参变量的含义一望可知)

```
library(BayesianStat)
X<-Normsig12Gibbs(n=5000, mu1=1.1,sigma1=3,mu2=1.8,sigma2=4,rho=0.6)     #抽样并赋予 X
plot(X,xlab=bquote(X[1]),ylab=bquote(X[2]))     #画出马氏链两分量的散点图
library(mvnormtest)   #此包要先下载安装
mshapiro.test(t(X))   #多元正态性 Shapiro-Wilk 检验，t(X)是转置样本矩阵 X
colMeans(X)     #计算均值（期望）向量
cov(X)              #计算协方差矩阵
cor(X)              #计算相关系数
```

马氏链两分量的散点图（图 6.2）显示出二元正态分布密度等高线所具有的椭圆特征以及相关系数为 0.6 的正相关特征，初步判断马氏链样本可以看成来自二元正态分布。为了进一步明确这个断言，我们对样本进行多元正态性 Shapiro-Wilk 检验，从 P-值可以确认样本服从二元正态分布。此外，从以上 R 命令的计算结果可以看出参数的样本估计值为 $(1.11, 9.29; 1.83, 16.18; 0.59)$ 与参数真值 $(1.1, 3^2; 1.8, 4^2; 0.6)$ 非常接近，估计精度相当高。综合以上二个结论可以进一步断言给定的二元正态分布是模拟马氏链的平稳分布。

注：对马氏链的第一个分量序列进行 Shapiro-Wilk 正态性检验

```
shapiro.test(X[,1])
  p-value = 0.3551
```

从 P-值看出第一个分量服从正态分布，同时其样本均值为 1.11 方差为 9.29，即样本可以说来自正态分布 $N(1.11, 9.29)$，另一方面，边际分布为 $N(1.1, 3^2)$，因此两个正态分布可以说是相同的，即第一个分量序列的平稳分布是边际分布 $\pi_1(x_1) = N(1.1, 3^2)$。同样可知第二个分量的平稳分布是 $\pi_2(x_2) = N(1.8, 4^2)$。

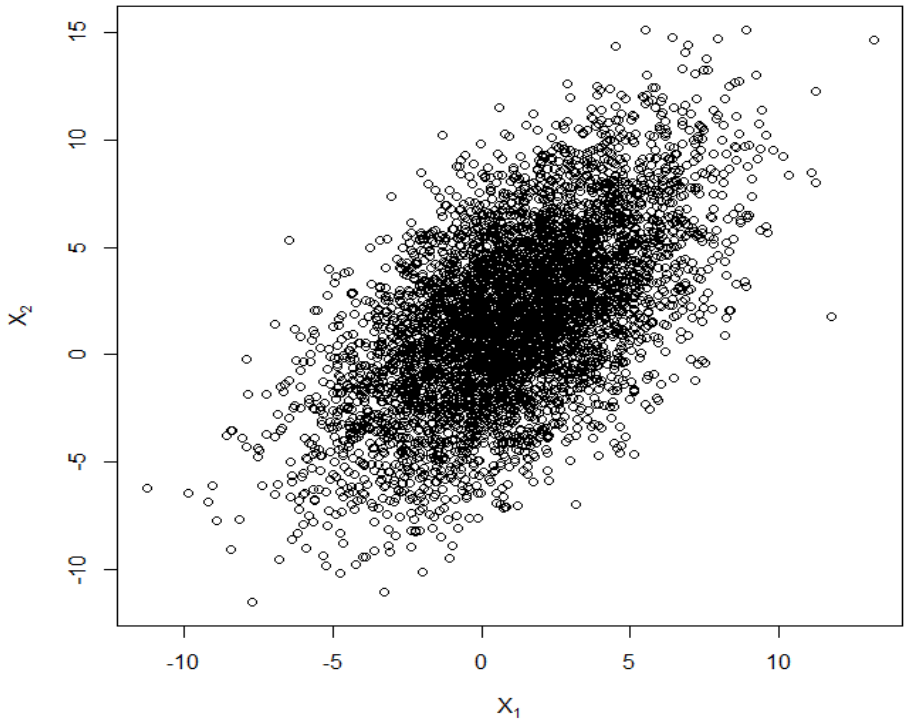

图 6.2 吉布斯抽样产生的马氏链两个分量的散点图

例 6.6 在 4.4.3 节中，我们得知当正态分布 $N(\mu,\sigma^2)$ 的两参数取无信息先验（杰弗里斯先验） $\pi(\mu,\sigma^2)=1/\sigma^2$ 时，后验分布

$$\pi(\mu,\sigma^2\,|\mathbf{x})\propto p(\mathbf{x}\,|\mu,\sigma^2)\pi(\mu,\sigma^2)\propto(\sigma^2)^{-(\frac{n}{2}+1)}\exp\left(-\frac{s^2+n(\mu-\bar{x})^2}{2\sigma^2}\right)$$

其中， $\mathbf{x}=(x_1,\cdots,x_n)$ 为样本， $s^2=\sum_{i=1}^{n}(x_i-\bar{x})^2$ 。作为二元后验分布，我们无法断定其核是什么分布的核，因此，直接抽样不可能做到。但是，当 σ^2 给定时，我们已知 μ 的条件后验分布为

$$\pi(\mu|\sigma^2,\mathbf{x})=N(\bar{x},\sigma^2/n)$$

另一方面，当 μ 给定时， σ^2 的条件后验分布为

$$\pi(\sigma^2|\mu,\mathbf{x})\propto(\sigma^2)^{-(\frac{n}{2}+1)}\exp\left(-\frac{s^2+n(\mu-\bar{x})^2}{2\sigma^2}\right)$$

令

$$y=[s^2+n(\mu-\bar{x})^2]/\sigma^2 \text{ 或 } \sigma^2=[s^2+n(\mu-\bar{x})^2]/y$$

则

$$\pi(\sigma^2|\mu,\mathbf{x})\propto(y)^{\frac{n+4}{2}-1}\exp(-\frac{y}{2})$$

上式右边是自由度为 $n+4$ 的卡方分布 $\chi^2(n+4)$ 的核，这表明

$$Y\,|\,(\mu,\mathbf{x})=[s^2+n(\mu-\bar{x})^2]/\sigma^2\,|\,(\mu,\mathbf{x})\sim\chi^2(n+4)$$

从而能抽取出Y的样本，进而就能得到σ^2的样本。综上知可用吉布斯抽样来获取样本，其算法如下：

1. 给定初始值$\sigma^{2(0)}$

2. 对于$t=1,2,\cdots,T$，产生样本$(\mu^{(t)},\sigma^{2(t)})$，做法是

（1）从$\mu|(\sigma^{2(t-1)},\mathbf{x})\sim N(\bar{x},\sigma^{2(t-1)}/n)$抽取$\mu^{(t)}$；

（2）从$Y|(\mu^{(t)},\mathbf{x})\sim\chi^2(n+4)$抽取$y$并$\sigma^{2(t)}=[s^2+n(\mu^{(t)}-\bar{x})^2]/y$

3. 对于$t+1$，回到第 2 步。

现在就用这个吉布斯抽样，对案例 4.3 中正态分布的两个参数进行 MCMC 估计，其 R 命令如下：

```
library(BayesianStat)
data(marathontime)
attach(marathontime) #此命令让我们可用数据的抬头 mtime 作为数据（对象）名
musigma<-NormmusigGibbs(n=5000, x=mtime,sig0= 6)  #吉布斯抽样，其中 sig0 为初始值
ts.plot(musigma[,1], xlab="迭代次数",ylab="mu")  #第一个参数马氏链轨迹图，平稳特征明显
ts.plot(musigma[,2], xlab="迭代次数",ylab="sigma^2")  #第二个参数马氏链轨迹图，结论同上
mean(musigma[,1]); mean(musigma[,2])
```

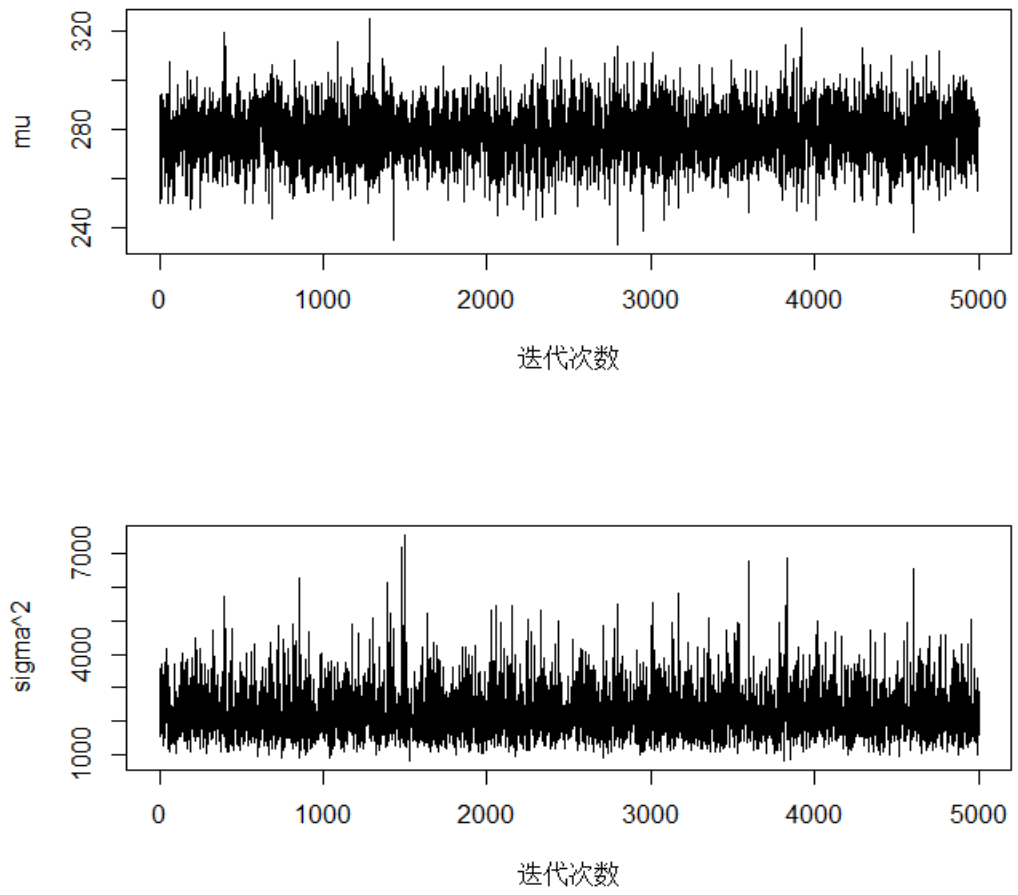

图 6.3 吉布斯抽样产生的马氏链

从图 6.3 我们看到，模拟的马氏链样本应该是平稳的，当然这里的证据还很弱，在 6.4 节将进一步讨论马氏链的收敛性。两个参数(μ,σ^2)的估计也容易地算出来了。

注：案例 4.3 中两个参数的估计值到底与参数真值有多接近无从判断。但是，可以做一个随机模拟来说明 MCMC 估计的优良性。令正态分布为$N(200,10^2)$，模拟容量为 1000 的样本 Z，然后假定两个参数未知，用上面的吉布斯抽样来估计两个参数，所用 R 命令如下

```
Z<-rnorm(1000,200,10)
X<-NormmusigGibbs(n=2000,x=Z,sig0=20)
mean(X[,1]); mean(X[,2])
[1] 200.3305
[1] 100.2064
```

我们看到 MCMC 方法估计出来的参数值与参数真值$(200,10^2)$非常近似！

6.2.2 多阶段吉布斯抽样

现在设随机向量 $\mathbf{X}=(X_1,\cdots,X_n)\sim\pi(\mathbf{x})$，其中 $\mathbf{x}=(x_1,\cdots,x_n)$。称条件概率密度函数 $\pi_i(x_i\mid x_1,\cdots,x_{i-1},x_{i+1},\cdots,x_n)=\pi_i(x_i\mid x_{-i})$，$i=1,\cdots,n$ 为满（或全）条件概率密度，它们是一元密度（也可以是低维的多元密度），而吉布斯抽样只要用到它们就可以了，这通常是一大优势，使得抽样可以容易进行。另外，关于满条件概率密度函数还有一个不易想到的性质，即联合密度 $\pi(\mathbf{x})$ 可以反过来通过其满条件密度函数表示。这点我们在二元情形已经看到了，现在考虑多元情形，为此需要先引入一个定义。

定义6.9 设 $\pi(\mathbf{x})$ 的边际密度为 $\pi_i(x_i),i=1,\cdots,n$，如果由 $\prod_{i=1}^{n}\pi_i(x_i)>0$ 可以推出 $\pi(\mathbf{x})>0$，则称联合密度 $\pi(\mathbf{x})$ 满足**正性条件**。

定理6.3 （Hammersley and Clifford,1970; Besag, 1974）如果联合密度 $\pi(\mathbf{x})$ 满足正性条件，那么对任意的点 $\mathbf{x}'=(x_1',\cdots,x_n')\in\chi$，有

$$\pi(\mathbf{x})=\pi(x_1,\cdots,x_n)\propto\prod_{i=1}^{n}\frac{\pi_i(x_i\mid x_1,\cdots,x_{i-1},x_{i+1}',\cdots,x_n')}{\pi_i(x_i'\mid x_1,\cdots,x_{i-1},x_{i+1}',\cdots,x_n')}$$

现在转入讨论多阶段吉布斯抽样，它是二阶段情形的自然推广，其具体抽样步骤如下（这里每个分量本身可以是向量）：

1. 给定初始值 $\mathbf{x}^{(0)}=(x_1^{(0)},\cdots,x_n^{(0)})$

2. 对于 $t=1,2,\cdots,T$，产生样本 $\mathbf{x}^{(t)}=(x_1^{(t)},\cdots,x_n^{(t)})$，做法是

（1）从条件密度 $\pi_1(x_1\mid x_2^{(t-1)},\cdots,x_n^{(t-1)})$ 抽取 $x_1^{(t)}$；

（2）从条件密度 $\pi_2(x_2\mid x_1^{(t)},x_3^{(t-1)},\cdots,x_n^{(t-1)})$ 抽取 $x_2^{(t)}$；

$\vdots$

（n）从条件密度 $\pi_n(x_n\mid x_1^{(t)},x_2^{(t)},\cdots,x_{n-1}^{(t)})$ 抽取 $x_n^{(t)}$

3. 对于 $t+1$，回到第2步。

这样依次抽取就得到了一个样本 $\{\mathbf{x}^{(t)}=(x_1^{(t)},\cdots,x_n^{(t)});1\le t\le T\}$，从抽样的做法可知这个样本形成了一个马氏链，而且只要它是不可约的，它的平稳分布就是 $\pi(\mathbf{x})$（目标分布），从而由定理6.2，对可积函数 $h(\mathbf{x})$ 有

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}h(\mathbf{x}^{(t)})=E^{\pi}[h(\mathbf{X})]=\int h(\mathbf{x})\pi(\mathbf{x})d\mathbf{x}\quad a.s.$$

例6.7 已知二元正态分布 $N(\mu_1,\sigma_1^2;\mu_2,\sigma_2^2;\rho)$ 的密度为

$$f(x_1,x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\times e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2}-2\rho\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]}$$

（1）当参数为 $(1.1,3^2;1.8,4^2;0.6)$ 时，模拟出容量1000的二元正态样本。

（2）现在假定参数真值遗失了，但得知均值向量和方差协方差阵参数 $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ 的杰弗里斯无信息先验为 $\pi(\boldsymbol{\mu},\boldsymbol{\Sigma})\propto|\boldsymbol{\Sigma}|^{-3/2}$。试利用（1）得到的样本和 MCMC 抽样计算参数的贝叶斯估计并与参数真值进行比较。

解： 先将二元正态密度的矩阵形式写出，令 $\mathbf{x}=(x_1,x_2)'$，那么矩阵形式为

$$f(\mathbf{x}\,|\,\boldsymbol{\mu},\boldsymbol{\Sigma})=(2\pi)^{-1}\,|\boldsymbol{\Sigma}|^{-1/2}\exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})]$$

由已知条件，均值向量和方差协方差矩阵是

$$\boldsymbol{\mu}=\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}=\begin{pmatrix}1.1\\ 1.8\end{pmatrix},\quad \boldsymbol{\Sigma}=\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}=\begin{pmatrix}9 & 7.2\\ 7.2 & 16\end{pmatrix}$$

（1）模拟容量 1000 的二元正态样本所用 R 程序如下

```
library(MASS)
#此包自动随基本包 base 一起下载，下面的二元正态抽样函数在此包中
mu<-c(1.1,1.8); Sigma<-matrix(c(9,7.2,7.2,16),2,2)
X<-mvrnorm(n = 1000, mu, Sigma)     #抽取二元正态样本
```

（2）给定样本 $\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n)'$，那么后验分布

$$\pi(\boldsymbol{\mu},\boldsymbol{\Sigma}\,|\,\mathbf{X})\propto|\,\boldsymbol{\Sigma}\,|^{-(n+3)/2}\exp[-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})]$$

这个分布无法直接抽样（因为不知是何分布）。但是，当方差协方差矩阵 $\boldsymbol{\Sigma}$ 已知时，与一元情形类似，可推出 $\boldsymbol{\mu}\,|\,(\boldsymbol{\Sigma},\mathbf{X})\sim N(\bar{\mathbf{x}},\boldsymbol{\Sigma}/n)$；当均值向量 $\boldsymbol{\mu}$ 已知时，可推出 $\boldsymbol{\Sigma}\,|\,(\boldsymbol{\mu},\mathbf{X})\sim IWishart(n;(\mathbf{X}-\boldsymbol{\mu})'(\mathbf{X}-\boldsymbol{\mu}))$（自由度为 n 的逆维希特分布，维希特分布本身是卡方分布的多维推广）。既然 $\boldsymbol{\mu}$ 和 $\boldsymbol{\Sigma}$ 的条件分布都是有名有姓可直接抽样的分布，所以可应用吉布斯抽样法，其 R 程序如下

```
library(BayesianStat)   #以下吉布斯抽样函数在此包中
PP<-MultiNormGibbs(n=6000,D=X)
#二元正态参数的吉布斯抽样，初值内部给定了
colMeans(PP)     #计算模拟马氏链样本均值
```

§6.3 梅切波利斯-哈斯廷斯算法

梅切波利斯-哈斯廷斯（Metropolis-Hastings，MH）**算法**是马氏链蒙特卡罗（MCMC）方法中的核心抽样法并具有一般性，被誉为二十世纪最重要的十大算法之一，由此可知它的重要性。MH算法的出发点是我们有一个待抽样的目标分布 $\pi(x)$，但它难以直接抽样。为了利用MH算法，要挑选一个适当的条件分布 $q(y|x)$，它在抽样中作为一个工具来使用，因此被称为工具分布（也称为建（提）议分布），是比较容易抽样的。由此，可以给出MH算法的一般步骤如下：

任意给定 $x^{(1)}$，对 $t=1,2,\cdots,T$

1. 分别抽取 $y_t \sim q(y|x^{(t)})$ 和 $u \sim U(0,1)$（均匀分布）；
2. 如果 $u \le \alpha(x^{(t)}, y_t)$，则取 $x^{(t+1)} = y_t$，否则，取 $x^{(t+1)} = x^{(t)}$，其中

$$\alpha(x,y) = \min\{1, \pi(y)q(x|y)/\pi(x)q(y|x)\}$$

3. 对于给定的 $x^{(t+1)}$，回到第1步。

反复进行这些步骤，我们就得到一个马氏链（样本）$\{x^{(t)}\}$。另外，概率 $\alpha(x,y)$ 称为**接受概率**。可以证明如下定理6.7，这一定理表明MH算法广泛的适用性。

定理6.7 只要工具分布的支撑 $\{y; q(y|x)>0\}$ 包含目标分布的支撑 $\{y;\pi(y)>0\}$，则目标分布 $\pi(x)$ 就是马氏链 $\{x^{(t)}\}$ 的平稳分布。

现在我们来看看工具分布 $q(y|x)$ 的一些特殊情形：

（1）如果工具分布是对称的，即 $q(y|x)=q(x|y)$，那么接受概率 $\alpha(x,y)=\min\{1,\pi(y)/\pi(x)\}$。这时抽样就直接称为梅切波利斯（Metropolis）抽样，是梅切波利斯等人早在1953年提出的。

（2）如果工具分布 $q(y|x)$ 独立于 x，即 $q(y|x)=q(y)$，则称抽样为独立MH抽样，这时接受概率简化为

$$\alpha(x,y)=\min\{1,\pi(y)q(x)/\pi(x)q(y)\}$$

（3）如果工具随机变量 $Y\sim q(y|x)$ 按方式 $Y_t=X^{(t)}+\varepsilon_t$ 产生，其中 ε_t 具有独立于 $X^{(t)}$ 的分布，那么称抽样为随机游动MH抽样。这时的工具分布具有形式 $q(y|x)=q(y-x)$（这是因为 $Y_t-X^{(t)}=\varepsilon_t$），例如

当 $\varepsilon_t\sim U(0,1)$ 时，有 $Y_t-X^{(t)}\sim U(0,1)$，即 $Y_t\sim U(x^{(t)},x^{(t)}+1)$

当 $\varepsilon_t\sim N(0,\sigma^2)$ 时，有 $Y_t-X^{(t)}\sim N(0,\sigma^2)$，即 $Y_t\sim N(x^{(t)},\sigma^2)$

如果 $q(z)$ 还是原点对称函数，即满足 $q(-z)=q(z)$，那么，接受概率就简化为 $\alpha(x,y)=\min\{1,\pi(y)/\pi(x)\}$。

例6.8 已知目标分布 $\pi(x)$ 为贝塔分布 $Beta(2.4,5.1)$，取工具分布为均匀分布 $U(0,1)$。利用MH算法抽取马氏链样本 $\{X^{(t)}\}$，然后，（1）回答：马氏链 $\{X^{(t)}\}$ 的平稳分布是贝塔分布 $Beta(2.4,5.1)$ 吗？（2）利用样本计算样本均值和方差并与总体均值和方差进行比较。

解：本例 MH 算法的抽样以及其它相关 R 命令如下

```
library(BayesianStat)   #以下抽样函数在此包中
X<-BetaMH(n=500,a=2.4, b=5.1)
#a, b是贝塔分布的两个参数，此命令抽取马氏链样本
ts.plot(X, xlab="迭代次数",col="blue")   #画马氏链轨迹图（图6.4）
ks.test(X,"pbeta",2.4,5.1)     #对马氏链样本进行柯尔莫哥诺夫-斯米尔诺夫检验
        One-sample Kolmogorov-Smirnov test
data:   X
D = 0.046, p-value = 0.2396
mean(X); var(X)
[1] 0.3210151，[1] 0.02958076
2.4/(2.4+5.1); 2.4*5.1/((2.4+5.1)^2*(2.4+5.1+1)) #计算贝塔分布的均值和方差
[1] 0.32，[1] 0.0256
```

从MH算法模拟得到的马氏链轨迹图（图6.4），我们看到从工具分布抽取的候选点有的被拒绝了，所以样本轨迹有时是水平前行的。

（1）对马氏链进行柯尔莫哥诺夫-斯米尔诺夫检验(Kolmogorov-Smirnov test)。根据P值，我们可以接受马氏链样本来自于贝塔分布 $Beta(2.4,5.1)$，即马氏链的平稳分布是贝塔分布 $Beta(2.4,5.1)$。

（2）利用样本计算所得样本均值为0.3210方差为0.0296与理论均值0.32方差0.0256非常近似

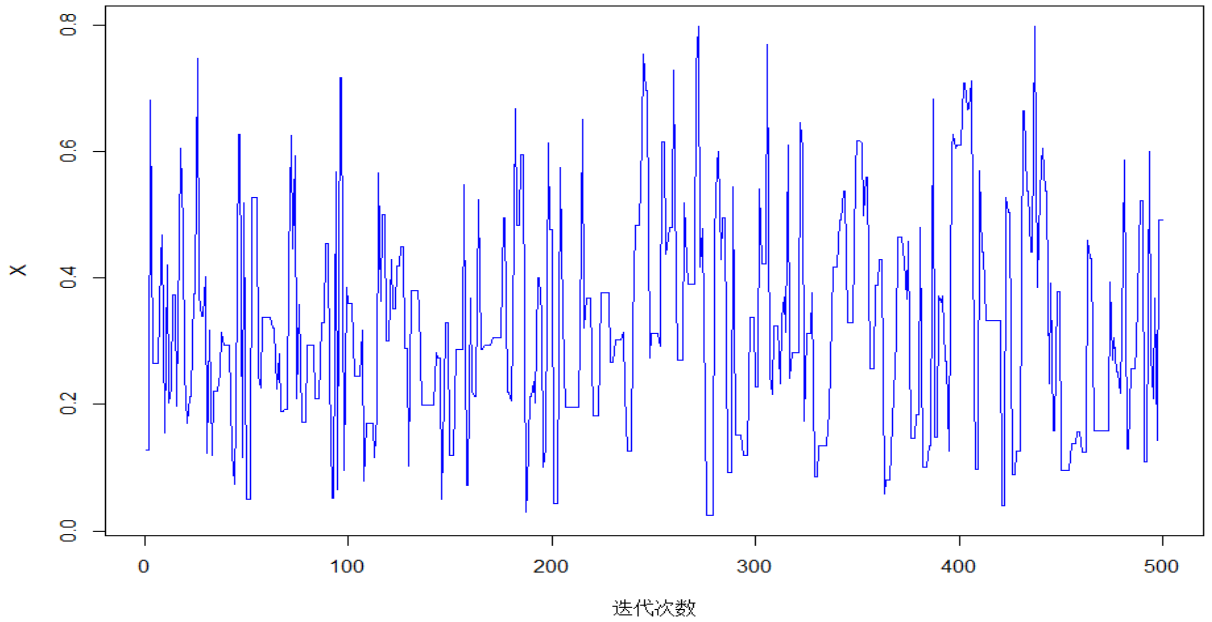

图6.4　MH算法模拟得到的马氏链

例6.9 已知二元正态分布 $N(0,1;0,1;\rho)$ 的密度为

$$f(x_1,x_2\mid\rho)=\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left[-\frac{x_1^2-2\rho x_1x_2+x_2^2}{2(1-\rho^2)}\right]$$

其容量为1000的样本在R包BayesianStat中，文件名为mydata。现在知道相关系数 ρ 的杰弗里斯无信息先验为 $\pi(\rho)\propto(1-\rho^2)^{-3/2}$。试利用贝叶斯和MCMC方法估计参数 ρ。

解： 设给定样本为 $\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n)'$，其中，$\mathbf{x}_i=(x_{1i},x_{2i})'$，那么样本的联合分布为 $g(\mathbf{X}\mid\rho)=\prod_{i=1}^{n}f(\mathbf{x}_i\mid\rho)$，相关系数 ρ 的后验分布为

$$\pi(\rho\mid\mathbf{X})\propto g(\mathbf{X}\mid\rho)\pi(\rho)\propto(1-\rho^2)^{-\frac{n+3}{2}}\exp\left[-\frac{\sum x_{1i}^2-2\rho\sum x_{1i}x_{2i}+\sum x_{2i}^2}{2(1-\rho^2)}\right]$$

这里看不出后验分布的核属于什么分布，因此无法直接抽样，也无法用吉布斯抽样法。现在我们利用独立 MH 算法，取贝塔分布 $Beta(a,b)$ 为工具（提议）分布，那么其核为 $x^{a-1}(1-x)^{b-1}$，因此接受概率

$$\alpha(x,y)=\min\{1,\pi(y\mid\mathbf{X})x^{a-1}(1-x)^{b-1}/\pi(x\mid\mathbf{X})\,y^{a-1}(1-y)^{b-1}\}$$

在本例中，取 $a=2,b=2.5$，初值 rho0=0.9，那么相应的 R 命令如下

```
library(BayesianStat) #以下抽样函数在此包中
data(mydata)
X<-mydata      #样本
Rho<-MNormrhoInM(n=6000,X,rho0=0.9,a=2,b=2.5) #独立MH抽样，rho0为初值
ts.plot(Rho,xlab="迭代次数")      #画出马氏链轨迹图
Rhoo<-MNormrhoInM(n=6000,X,rho0=0.1,a=2,b=2.5) #改初值为rho0=0.1再抽样
lines(1:6000,Rhoo,col="blue")      #将新的马氏链轨迹画在前面的图中
RRho<-Rho[1001: 6000]      #把最前面1000个样本烧掉不要，以消除初值影响
mean(RRho)
[1] 0.591974
```

我们得到相关系数 ρ 的后验均值估计为 0.5920。这里，我们以差距较大的两个初始值为起点抽取了两条马氏链，从图 6.5，我们看到这两条马氏链很快收敛（平稳或混合）在一起。但是最初各自都受到初始值的影响，为消除这种影响，在估计参数之前，把最前面一段链切除，然后再应用剩下的部分估计参数。

§6.4　MCMC 的收敛性问题

本章前面三节讨论了 MCMC 的思想和实现问题，我们知道只有当模拟的马氏链收敛于平稳分布（目标分布）时，才有理由利用该马氏链来估计有关参数或进行其它统计推断，因此诊断 MCMC（算法）的收敛性（即 MCMC 产生的马氏链的收敛性）在 MCMC 的应用中是至关重要的问题。在前面三节的各个例子中，有的从某个侧面验证了产生的马氏链收敛，有的没有验证，严格而言这当然是不行的，因而我们将弥补上这点。本节将对模拟的马氏链的收敛性从实用的角度加以扼要介绍，但必须知道的是诊断马氏链是否收敛是一个复杂的问题，最好用多种方法从不同侧面进行验证。如果多种方法一致认定马氏链收敛，那么结论的说服力自然就强。总的来看，MCMC 产生的马氏链的收敛性受三个因素的影响：（1）初始值；（2）后验密度（目标分布密度）的形状；（3）在 MH 算法中的工具分布。

（1）**初始值** 在用计算机进行模拟的时候总是要有一个开头，这就是初始值。我们要利用一切可得的信息尽量使初始值离参数真值较近。参数真值当然是未知的，不过可能也有信息可以利用，例如，某个参数是大于零的，那么初始值最好取正数。如果对后验分布密度有所了解，那么初始值就应当在后验高密度也就是中心区域。另外，为了消除初始值的影响，应该把开始一段的马氏链弃之不用。

（2）**后验分布密度（目标分布密度）** 如果后验密度的形状是单峰的，问题相对简单。如果后验密度的形状是多峰的，要小心伪收敛现象，即陷入非最大值的其它极值点这个现象。

（3）**工具分布** 它对模拟马氏链的收敛性的影响是不言而喻的，好的工具分布的中心区域应与后验密度（目标分布）的中心区域有较多重合，尾部较厚，支撑包含后验密度的支撑。

MCMC 算法的收敛性的检验（诊断）可以从图形和数量两方面来考虑，有时数量也可以通过图形来表示，以增加直观性。在 R 包 coda 中有许多诊断马氏链收敛性的函数，把它下载安装后，可多加应用。下面介绍一些常用的方法。

（1）马氏链轨迹图检验法。收敛的马氏链应该没有趋势和周期，而是在一水平线上下小幅波动。图 6.1、图 6.3、图 6.4 和图 6.5 显示的马氏链都可以初步看成是收敛的马氏链。下面我们来画例 6.7 中五个参数的马氏链轨迹图，图 6.6 从上到下分别是均值 μ_2, μ_1 和相关系数 ρ 的马氏链轨迹图；图 6.7 从上到下分别是方差 σ_1^2, σ_2^2 的马氏链轨迹图，从各个马氏链轨迹图，我们可以初步认为这些马氏链是收敛的。当然，仅仅观察马氏链轨迹图就断定马氏链收敛的说服力是不够强的，要继续用其它方法进一步诊断这些马氏链的收敛性。

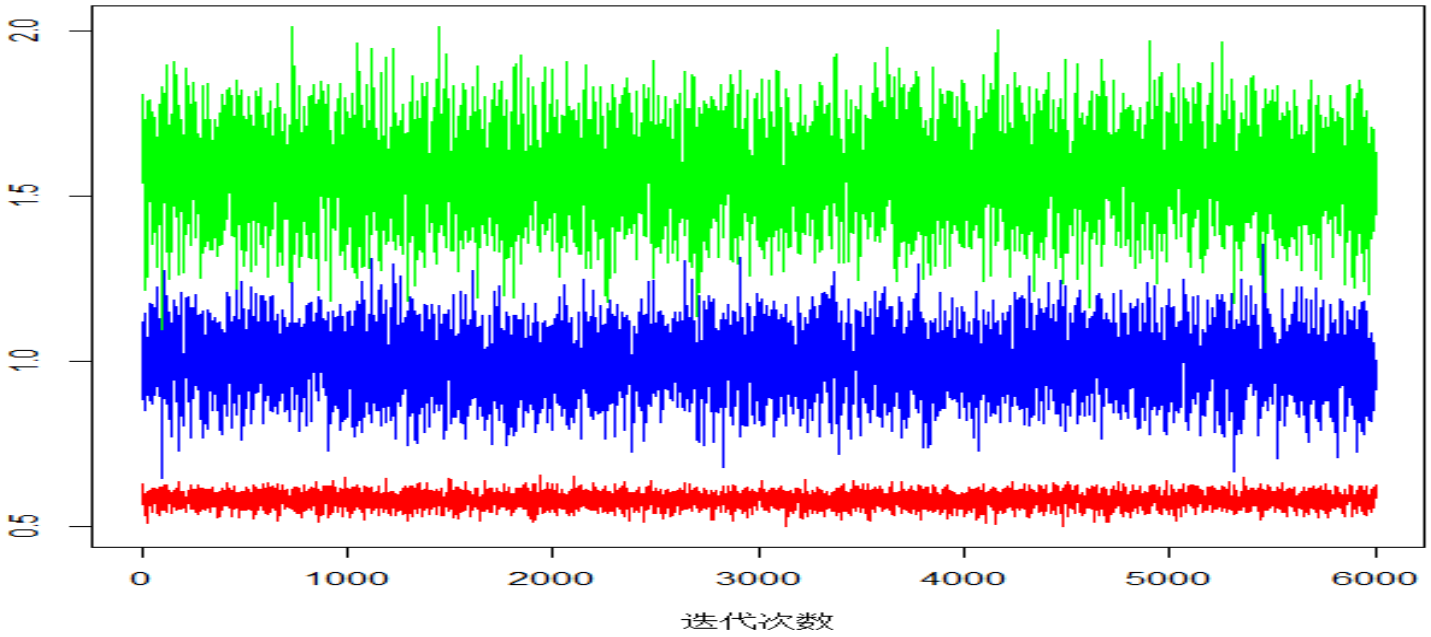

图6.6　例6.7中两个均值参数和相关系数的马氏链轨迹

（2）Geweke检验法。该法由Geweke (1992)提出， 它实际上由数量诊断和图形诊断两部分组成，图形诊断是由Steve Brooks建议的。它利用如果马氏链收敛了，那么它来自同一平稳分布，因而它的前一部分与后一部分的均值应该相等这一事实。在马氏链前一部分与后一部分渐进独立的假设条件下，Geweke构造的检验统计量渐进服从标准正态分布（因此是Z检验）。需要注意的是，由于假定马氏链前一部分与后一部分渐进独立，在应用Geweke检验法时马氏链前一部分与后一部分不可以重叠。实际应用时，Geweke检验对每个分量马氏链算出检验统计量的值，当值的绝对值小于2时，马氏链被认为是收敛的。但是，Geweke检验法的数量诊断每次仅算出一个统计量值，不如其图形诊断来的合理。图形诊断的做法是用与数量诊断同样的方法算出一个统计量值，然后，把最前面的一小段马氏链切除掉，再次用同样的方法算出第二个统计量值，以此类推，一共算出若干（默认20）个统计量值，最后在坐标系上画出它们的位置，同时，用虚线画出置信度0.95的置信区间形成的置信带。如果绝大多数统计量值在可信带内，则可以认为马氏链是收敛的。以下是例6.7中五个参数的马氏链的Geweke检验的R命令和相关结果，我们看到无论是数量诊断还是图形诊断都可以断定五个参数的马氏链收敛了（第4个参数的马氏链从数量诊断上不能断定其收敛，但从图形诊断可以断定其收敛）（图6. 8）。最后，要注意Geweke检验中的样本容量要小于十万。

```
library(coda)
geweke.diag(as.mcmc(PP))
geweke.plot(as.mcmc(PP))   #画诊断图
```

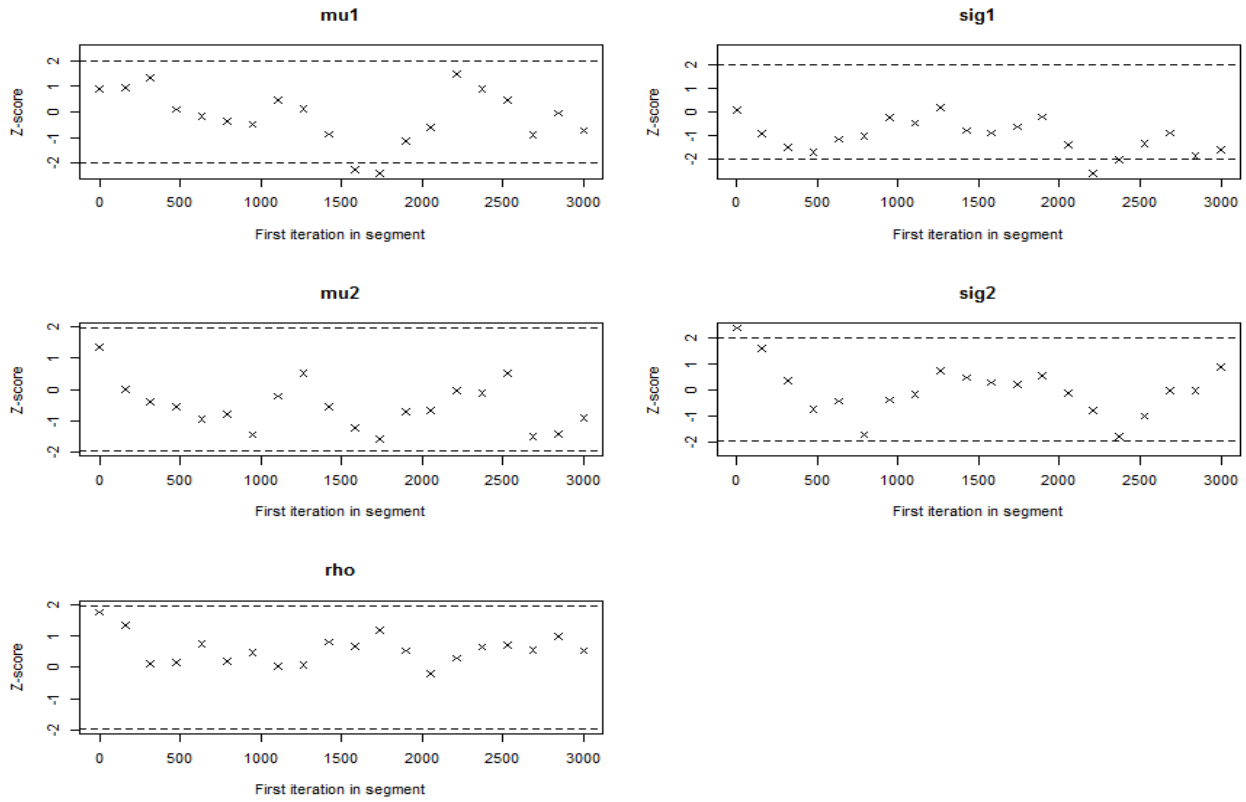

图6.8　例6.7中五个参数的马氏链收敛性的Geweke检验

（3）分位数检验法。该法滚动利用经验（累积）分布函数计算 0.025,0.5,0.975 分位数，如果马氏链收敛，则三个分位数应该逐步形成水平直线。以下是例 6.7 中五个参数的马氏链的分位数检验及其 R 命令，我们看到每个参数的马氏链对应的三个分位数线都逐渐形成水平直线，因此五个参数的马氏链都可以看成是收敛的（图 6.9）。

```
cumuplot(as.mcmc(PP))
```

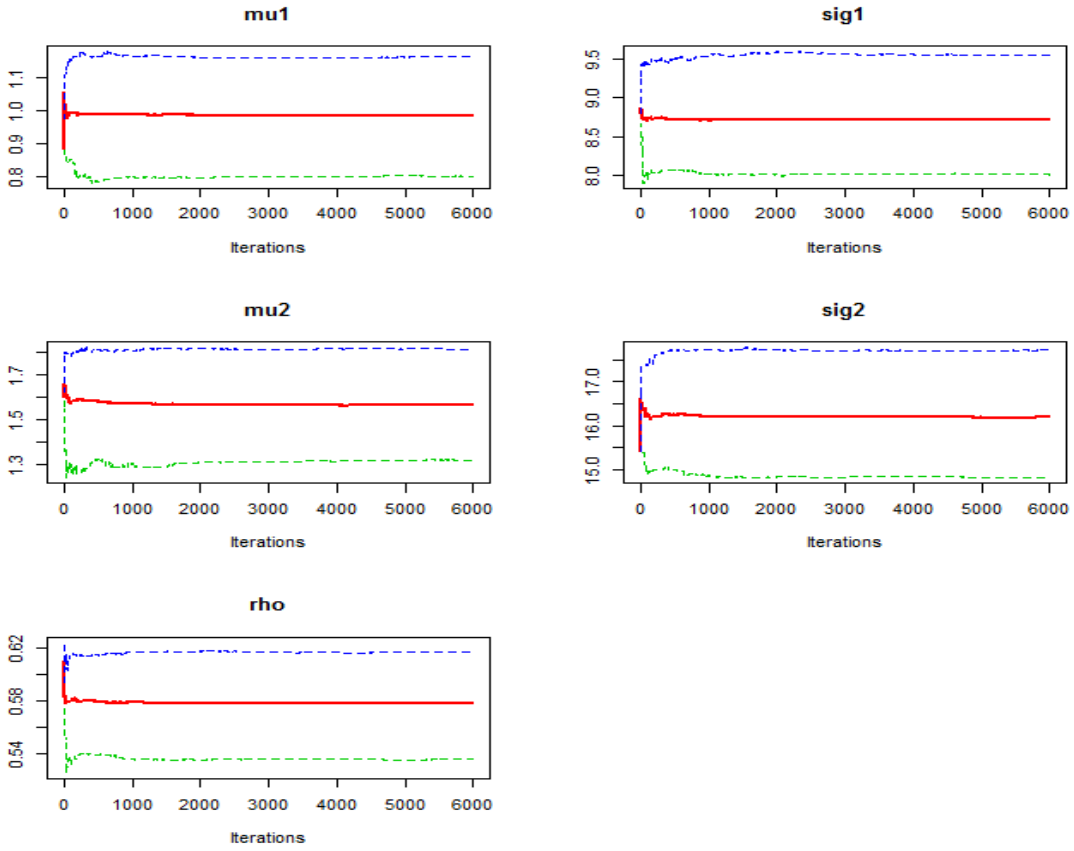

图6.9　例6.7中五个参数的马氏链收敛性的分位数检验

（4）Heidelberger检验法。此检验由两部分组成：第一部分为马氏链平稳性检验，是利用Cramer–von Mises 统计量检验马氏链是否来自同一个平稳分布。如果不能通过检验则会指示要产生更长的马氏链；如果通过了检验则会报告收敛从何时开始。因此，这个检验方法可以用来控制产生马氏链的迭代的次数。第二部分被称为半宽检验（Halfwidth test），它先是利用谱分析方法把渐进方差估计出来，而后用方差和已通过收敛检验的那段马氏链去估计置信度0.95的均值的置信区间。所谓半宽（Halfwidth）就是指这个置信区间的一半宽度。最后，将半宽与均值的估计值进行对比，如果比值的绝对值小于0.1，则检验获得通过，即认为马氏链的长度已能使均值的估计足够精确；否则，检验不能通过，即还要增加马氏链的长度以获得足够精确的均值估计，换句话说，半宽检验只是考察均值估计的精度是否足够，意义较小。

注：半宽与均值的估值的比值的绝对值小于0.1意味着置信区间很窄，从而估值的精度高。

以下是例6.7中五个参数的马氏链的Heidelberger检验。从第一部分看到五个参数的马氏链都通过了平稳性检验，而且从第一次迭代开始就收敛了（其实已经抛弃了受初始值影响的最初一段）。从第二部分看到五个参数的马氏链都通过了半宽检验，而且不难看出半宽与均值的估值的比的绝对值都远远小于0.1。

```
heidel.diag(as.mcmc(PP))
```

（5）Gelman 检验法。Gelman 和 Rubin (1992)发现仅用一条马氏链样本去诊断马氏链的收敛性有时会出现误判，即把未真正收敛的马氏链诊断为收敛了。为了解决这一问题，他们提出同时考察多条初值尽可能分散的马氏链样本，如果这些马氏链都是平稳（收敛）的，那么它们的统计特征就应该是一样的，比如，样本均值和样本方差应该相等，进而利用与方差分析中的做法类似，构造出一个统计量用以诊断马氏链的收敛性。现在设随机变量φ的分布是目标分布，具有均值μ和方差σ^2。再设通过 MCMC 方法产生了长度都为n的m条马氏链样本$\{\varphi_{jt}\}$，其中，$j=1,2,\cdots,m$表示第j马氏链，$t=1,2,\cdots,n$表示第t次迭代，那么，与方差分析中的做法类似，令

$$\bar{\varphi}_{j.}=\frac{1}{n}\sum_{t=1}^{n}\varphi_{jt}\ ,\quad \bar{\varphi}_{..}=\frac{1}{nm}\sum_{j=1}^{m}\sum_{t=1}^{n}\varphi_{jt}$$

则链间方差为

$$A=\frac{1}{m-1}\sum_{j=1}^{m}(\bar{\varphi}_{j.}-\bar{\varphi}_{..})^2$$

链内方差为

$$W=\frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{t=1}^{n}(\varphi_{jt}-\bar{\varphi}_{j.})^2$$

令$B=nA$，用W和B的加权平均来估计方差σ^2得

$$\hat{\sigma}_{+}^{2}=\frac{n-1}{n}W+\frac{B}{n}$$

如果初始值就来自目标分布（此时所有马氏链是收敛的），那么，估计量 $\hat{\sigma}_+^2$ 是 σ^2 的无偏估计量。如果初始值相对于目标分布而言是过度分散的，那么 $\hat{\sigma}_+^2$ 将高估 σ^2（因为波动的比目标分布厉害）。

注：从方差分析，我们知道总变差

$$\sum_{j=1}^{m}\sum_{t=1}^{n}(\varphi_{jt}-\bar{\varphi}_{..})^2=\sum_{j=1}^{m}\sum_{t=1}^{n}(\varphi_{jt}-\bar{\varphi}_{j.})^2+n\sum_{j=1}^{m}(\bar{\varphi}_{j.}-\bar{\varphi}_{..})^2$$

因此方差 σ^2 可用下式估计

$$\frac{1}{mn}\sum_{j=1}^{m}\sum_{t=1}^{n}(\varphi_{jt}-\bar{\varphi}_{..})^2=\frac{1}{mn}\sum_{j=1}^{m}\sum_{t=1}^{n}(\varphi_{jt}-\bar{\varphi}_{j.})^2+\frac{1}{m}\sum_{j=1}^{m}(\bar{\varphi}_{j.}-\bar{\varphi}_{..})^2$$

$$=\frac{n-1}{n}W+\frac{(m-1)}{m}\frac{B}{n}=\tilde{\sigma}^2<\hat{\sigma}_+^2$$

考虑到抽样的波动性，将 $\hat{\sigma}_+^2$ 再放大一点点，令

$$\hat{V}=\frac{n-1}{n}W+\frac{m+1}{m}\frac{B}{n}$$

有 $\tilde{\sigma}^2<\hat{\sigma}_+^2<\hat{V}$，因此，估计量 $\hat{V}$ 是从上方趋近于方差 σ^2。考虑两个估计量 $\hat{V}$ 和 W 的比值的正开方

$$R=\sqrt{\hat{V}/W}$$

称 R 为潜尺度缩减因子（Potential Scale Reduction Factor, PSRF），当样本量（迭代次数）增大时（从而马氏链趋向收敛），R 从上方趋近于 1。因此，可以用潜尺度缩减因子来诊断马氏链的收敛性，一般而言，$R \le 1.1$可以作为收敛的标准。后来，Brooks 和 Gelman (1998)对潜尺度缩减因子进行了微小的改进，得到修正的 R 如下（仍然称为潜尺度缩减因子）

$$R_c = R\sqrt{(d+3)/(d+1)} = \sqrt{\hat{V}(d+3)/[W(d+1)]}$$

其中，d 是目标分布的估计即 t 分布的自由度，可用矩估计为 $d = 2\hat{V}^2/Var(\hat{V})$。

Gelman 检验法在 R 包 coda 上的实施由两部分组成，第一个命令是 gelman.diag，它计算出潜尺度缩减因子的点估计和一个可信度 97.5%的可信上限，据以诊断马氏链的收敛性。但是，只看一个数值有时会产生误诊，因为在未收敛时，偶尔潜尺度缩减因子也会很接近于 1。因此，在包 coda 中用另一个命令 gelman.plot 来滚动算出一系列潜尺度缩减因子的可信度 97.5%的可信上限并做出图来直观显示可信上限的演变，从而用以诊断马氏链的收敛性，这时可信上限被简称为缩减因子（shrink factor）。下面来看一个例子。

例 6.9 续 对例 6.9 中的问题产生四条初始值分别为（0.9，0.7，0.4，0.01）的马氏链，然后用 Gelman 检验法诊断其收敛性。

解：产生四条初始值不同的马氏链和做收敛性诊断的命令如下

```
RR=matrix(0, 2000,4)   #产生一个 2000*4 的矩阵
RR[,1]<-MNormrhoInM(n=2000,X,rho0=0.9,a=2,b=2.5)
RR[,2]<-MNormrhoInM(n=2000,X,rho0=0.7,a=2,b=2.5)
RR[,3]<-MNormrhoInM(n=2000,X,rho0=0.4,a=2,b=2.5)
RR[,4]<-MNormrhoInM(n=2000,X,rho0=0.01,a=2,b=2.5)
ZZ<-mcmc.list(mcmc(RR[,1]),mcmc(RR[,2]),mcmc(RR[,3]),mcmc(RR[,4]))
gelman.diag(ZZ)   #数量诊断
Potential scale reduction factors:
     Point est.      Upper C.I.
[1,]          1.01           1.02
gelman.plot(ZZ)     #图形诊断
```

从命令 gelman.diag 的结果，我们看到潜尺度缩减因子的点估计为 1.01 ，可信度 97.5%的可信上限为 1.02，因此似乎可以断定马氏链收敛了。但要注意的问题是如果我们只产生 450 左右长度的马氏链，那么潜尺度缩减因子及其可信度 97.5%的可信上限同样很接近于 1，这时断定马氏链收敛那就是误判了（图 6.10）。因此，为了诊断正确，我们还要用命令 gelman.plot 进行图形诊断。从图 6.10 看到可信度 97.5%的可信上限形成的折线（或称为缩减因子线）在 1500 这个时点之前是有比较大波动的（偶尔很接近于 1），只有在 1500 这个时点之后，这条折线才明显收敛于 1，换句话说，要 1500 这个时点之后的马氏链才是真正收敛的（图 6.10 中的“中位线”仅是参考，可以不管它）。

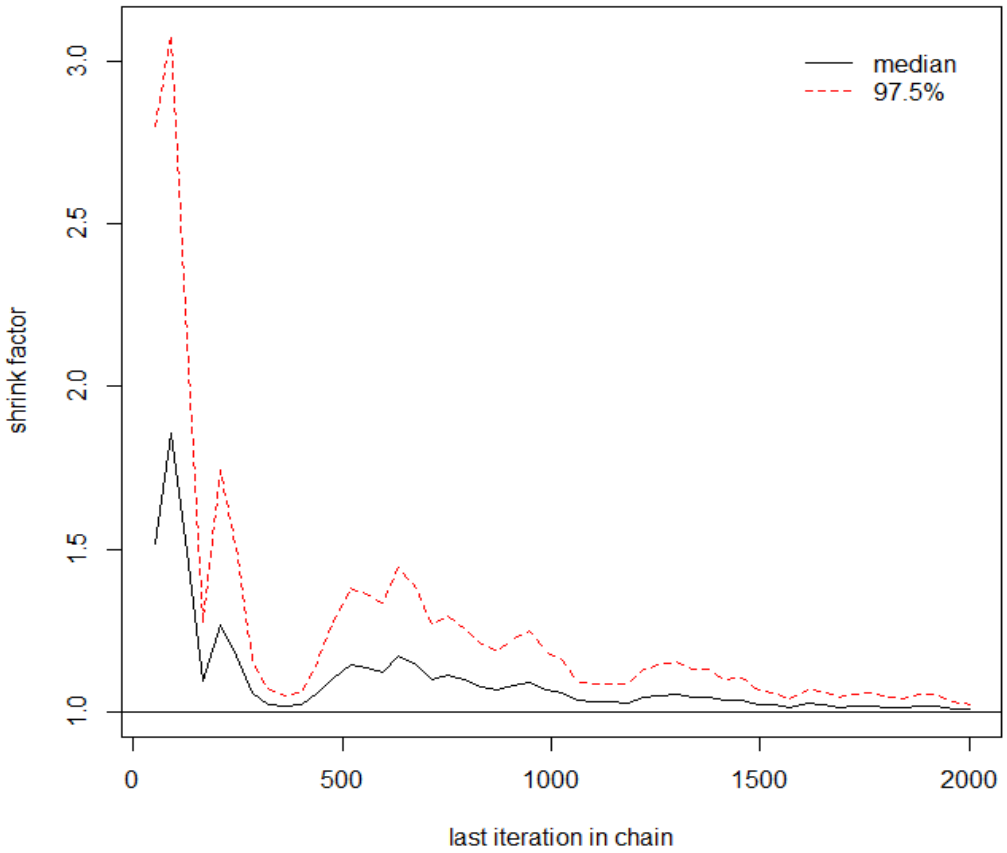

图 6.10 缩减因子演化图

作业：PP137-138，1，2，3，4，5，7