Metropolis-Hastings 算法, HMC 算法与 SMC 算法

王璐

有些 Bayesian 模型不存在 conjugate priors, 甚至参数的 full conditional distributions 也不是常见的分布或很难进行抽样。这种情况下，无法使用 Gibbs sampler 估计参数的后验分布。本章介绍一种更通用的 MCMC 方法 — Metropolis-Hastings 算法，它几乎适用估计任何 priors 下的 Bayesian 模型。我们首先以一个 Bayesian Poisson 回归模型为例引入该方法。

1 A Bayesian Poisson Regression Model

一项针对麻雀的研究记录了 52 只雌雀在一个夏天的繁殖数据，图1用 boxplot 展示了这些雌雀繁殖的后代数与它们年龄的关系。从图1可以看到，2 岁的雌雀繁殖后代数的中位数 (median) 最高。

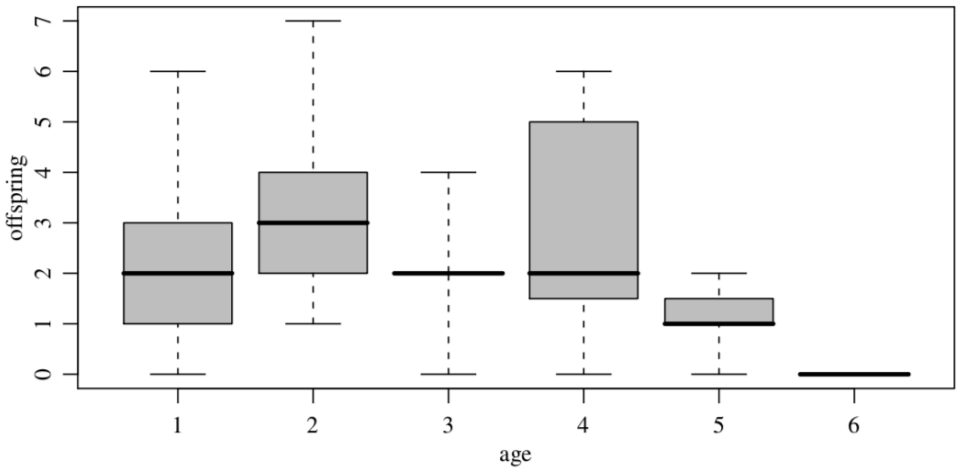

Figure 1: 雌雀繁殖的后代数与年龄的 boxplot. Picture sourse: Hoff (2009)

我们希望用一个概率模型拟合该数据以预测雌雀各年龄繁殖后代数的期望。响应变量 (response) y_i 对应雌雀 i 繁殖的后代数，是一个非负整数 $y_i \in \{0, 1, 2, \dots\}$；解释变量 x_i 为雌雀 i

的年龄。考虑采用如下的 Poisson 模型:

$$y_i \mid x_i \sim Po(\theta(x_i)), \quad i = 1, \ldots, n. \tag{1}$$

根据图1，可以假设 y_i 的期望 $\theta(x_i)$ 是 x_i 的二次函数，即 $\theta(x_i) = \beta_1 + \beta_2 x_i + \beta_3 x_i^2$，但是该模型有一个问题：估计的系数 $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ 可能使 $\theta(x_i) < 0$. 因此假设 $\log \theta(x_i)$ 是 x_i 的二次函数：

$$\log \theta(x_i) = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 \tag{2}$$

此时 y_i 的期望 $\theta(x_i) = \exp(\beta_1 + \beta_2 x_i + \beta_3 x_i^2) > 0, \forall \boldsymbol{\beta}$.

为了更好地描述估计的系数 $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ 的不确定性 (如方差、置信区间等), 考虑采用 Bayesian 分析。为上述 Poisson 回归模型(1)-(2)的参数 $\boldsymbol{\beta}$ 设定如下的多元正态 prior:

$$\boldsymbol{\beta} \sim N_3(\mathbf{0}, 100 I_3) \tag{3}$$

在 prior (3)下，$\boldsymbol{\beta}$ 的后验分布不是多元正态分布，其各分量的 full conditional distribution 也不是常见的容易抽样的分布，此时无法使用 Gibbs sampler, 但 Metropolis 方法仍然能通过构建 Markov chain 获得 $\boldsymbol{\beta}$ 后验分布的样本。

2 Metropolis 算法

对一般的 Bayesian 模型，用 $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ 表示观察值的似然函数，参数 $\boldsymbol{\theta} \in \mathbb{R}^p$ 的 prior 为 $p(\boldsymbol{\theta})$. $\boldsymbol{\theta}$ 的后验分布为

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$

但该分布的 normalizing constant $\int p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$ 通常很难计算，也很难直接从该后验分布抽样。

Metropolis 算法通过持续地在参数空间随机 “游走” 寻找目标后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 概率密度较高的区域。假设在当前时刻 Markov chain 得到的样本为 $\boldsymbol{\theta}^{(t)}$, 在 $\boldsymbol{\theta}^{(t)}$ 附近随机产生一点，如果该点对应的 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 的值高于 $p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})$, 则让 Markov chain 沿该点的方向继续移动；反之以一定概率决定是否沿较低概率密度的方向移动，这点对于有效地探索多峰值的后验分布很重要。

运行 Metropolis 算法需要先选取一个对称的 proposal 分布 $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$, $g(\cdot \mid \cdot)$ 满足 $g(\boldsymbol{\theta}_a \mid \boldsymbol{\theta}_b) = g(\boldsymbol{\theta}_b \mid \boldsymbol{\theta}_a)$. 比如常用的 proposal 分布为:

- $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) \sim \boldsymbol{U}\left(\boldsymbol{\theta}^{(t)} - \boldsymbol{\delta}, \boldsymbol{\theta}^{(t)} + \boldsymbol{\delta}\right)$
- $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)}) \sim N_p(\boldsymbol{\theta}^{(t)}, \mathrm{diag}(\boldsymbol{\delta}))$

后面我们会讨论如何选取 $\boldsymbol{\delta}$ 使算法运行更有效率。

选定 proposal 分布 $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ 后，Metropolis 算法如下产生样本 $\boldsymbol{\theta}^{(t+1)}$:

1. 抽取 $\boldsymbol{\theta}^\star \sim g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$

2. 计算接受比率 (acceptance ratio)

$$r_t = \frac{p(\boldsymbol{\theta}^\star \mid \boldsymbol{y})}{p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})} = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}^\star)p(\boldsymbol{\theta}^\star)}{p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(t)})p(\boldsymbol{\theta}^{(t)})}.$$

3. 令

$$\boldsymbol{\theta}^{(t+1)} = \begin{cases} \boldsymbol{\theta}^\star & \text{概率 } \min(r_t, 1) \\ \boldsymbol{\theta}^{(t)} & \text{概率 } 1 - \min(r_t, 1) \end{cases}$$

2.1 Metropolis 算法的收敛性

根据 Markov chain 理论，一条遍历的 Markov chain 会收敛到唯一的 stationary distribution. 此时对于 transition pdf $p(\boldsymbol{x} \mid \boldsymbol{x}')$, 如果能找到分布 $\pi(\boldsymbol{x})$ 满足

$$\pi(\boldsymbol{x}) = \int p(\boldsymbol{x} \mid \boldsymbol{x}')\pi(\boldsymbol{x}')d\boldsymbol{x}', \quad \forall \boldsymbol{x}, \boldsymbol{x}' \tag{4}$$

则 $\pi(\boldsymbol{x})$ 就是 Markov chain 收敛到的 stationary distribution.

(4)成立的一个充分条件如下：

$$p(\boldsymbol{x}' \mid \boldsymbol{x})\pi(\boldsymbol{x}) = p(\boldsymbol{x} \mid \boldsymbol{x}')\pi(\boldsymbol{x}') \quad \forall \boldsymbol{x}, \boldsymbol{x}' \tag{5}$$

对(5)两边关于 $\boldsymbol{x}'$ 积分可得(4). 称(5)为 **detail balance** 条件。因此对一条遍历的 Markov chain, 找到满足 detail balance (5)的分布 $\pi(\boldsymbol{x})$ 就找到了 Markov chain 的 stationary distribution.

接下来我们证明 Metropolis 算法产生的 Markov chain 的 stationary distribution 为参数的后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$.

- 如果 $\boldsymbol{\theta}^{(t+1)} \neq \boldsymbol{\theta}^{(t)}$, 说明接受了候选样本，对应的 transition density 为

$$\begin{aligned} p(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) &= g(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)})P(\boldsymbol{\theta}^{(t+1)} \text{ is accepted}) \\ &= g(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \min\left\{\frac{p(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{y})}{p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})}, 1\right\} \end{aligned}$$

  此时

$$p(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)})p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y}) = g(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \min\{p(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{y}), p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})\} \tag{6}$$

  由于 proposal density $g(\cdot \mid \cdot)$ 是对称的，因此(6)的右端关于 $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)})$ 对称，所以

$$p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t+1)})p(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{y}) = p(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)})p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})$$

  即 detail balance 条件对后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 成立。

- 如果 $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$，不论 transition density 具有何种形式，detail balance 条件对后验分布总是成立的，因为 $p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)})p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y}) = p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)})p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})$.

因此如果 Metropolis 方法生成的 Markov chain 是遍历的，由于 detail balance 条件对后验分布成立，该 Markov chain 会收敛到后验分布。

Remarks

1. 如果参数 $\boldsymbol{\theta}$ 是一个 p 维连续向量，将 proposal distribution $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ 选为 $N_p(\boldsymbol{\theta}^{(t)}, \text{diag}(\boldsymbol{\delta}))$ 可以得到遍历的 Markov chain.

2. 在 Metropolis 算法中，接受比率并不是越高越好。如果 proposal distribution $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ 中选取的 $\|\boldsymbol{\delta}\|$ 很小，每一步迭代产生的候选样本 $\boldsymbol{\theta}^\star$ 会很接近当前样本 $\boldsymbol{\theta}^{(t)}$，这导致 $p(\boldsymbol{\theta}^\star \mid \boldsymbol{y}) \approx p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})$，因此候选样本很容易被接受。但会导致 Markov chain 在参数空间中移动地非常缓慢，需要很长时间才能收敛到 stationary distribution, 如图2的左图所示。此时 Markov chain 上样本之间的相关性较高，这使得样本均值对后验期望的近似精度下降（回顾 effective sample size 的定义）。在 Gibbs sampling 中，我们无法直接控制 Markov chain 的相关性，但在 Metropolis 算法中，可以通过调节 $\|\boldsymbol{\delta}\|$ 来调整 Markov chain 的相关性。

3. 如果在 proposal distribution $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ 中令 $\|\boldsymbol{\delta}\|$ 很大，每步产生的候选样本 $\boldsymbol{\theta}^\star$ 虽然能更自由地探索参数空间，其对应的 $p(\boldsymbol{\theta}^\star \mid \boldsymbol{y})$ 很可能非常小，因而很容易被拒绝。这会造成 Markov chain 在一段时间 “困在” 局部某点处不动，导致样本间相关性较高，如图2的右图所示。

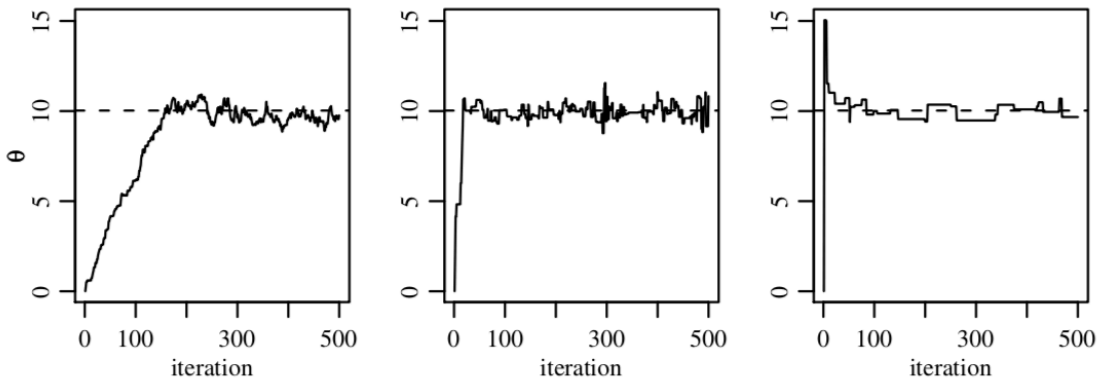

Figure 2: 在 proposal distribution $N(\theta^{(t)}, \delta^2)$ 中选取不同的 δ 得到的 Markov chains. 从左到右分别对应 $\delta^2 = 1/32$, 2, 64. Markov chain 的总体接受比率从左到右分别为 87%, 35%, 5%. Picture sourse: Hoff (2009)

4. 实践中经常采用以下做法为 proposal distribution 选取合适的 $\boldsymbol{\delta}$: 在不同 $\boldsymbol{\delta}$ 取值下，先运行 Metropolis 算法产生一些较短的 Markov chains, 选取 $\boldsymbol{\delta}$ 使得 Markov chain 的接受比率大致在 20% 到 50% 之间 (Hoff, 2009)；确定一个合理的 $\boldsymbol{\delta}$ 后，再运行 Metropolis 算法产生较长的 Markov chain 做统计推断。

3 Bayesian Poisson 回归模型的 Metropolis 算法

本节介绍如何使用 Metropolis 算法估计 Section 1中的 Bayesian Poisson 回归模型(1)-(3). 在每步迭代中，我们整体对参数 $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)$ 进行抽样，因为 block sampling 可以减少样本间的相关性。从 proposal distribution $g(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(t)})$ 产生一个候选样本 $\boldsymbol{\beta}^\star$，令 $\boldsymbol{x}_i = (1, x_i, x_i^2)^\top$，矩阵 $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, 此时 Metropolis 算法的接受比率为:

$$
\begin{aligned}
r &= \frac{p(\boldsymbol{\beta}^\star \mid X, \boldsymbol{y})}{p(\boldsymbol{\beta}^{(t)} \mid X, \boldsymbol{y})} \\
&= \frac{N_3(\boldsymbol{\beta}^\star \mid \mathbf{0}, 100 I_3) \prod_{i=1}^n \mathrm{Po}\left(y_i \mid \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}^\star)\right)}{N_3(\boldsymbol{\beta}^{(t)} \mid \mathbf{0}, 100 I_3) \prod_{i=1}^n \mathrm{Po}\left(y_i \mid \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}^{(t)})\right)}.
\end{aligned}
$$

在本例中，我们选取的 proposal distribution 为 $N(\boldsymbol{\beta}^{(t)}, \hat{\sigma}^2 (X^\top X)^{-1})$, 其中 $\hat{\sigma}^2$ 是 $\{\log(y_1 + 1/2), \ldots, \log(y_n + 1/2)\}$ 的样本方差。因为对于线性回归模型，$\boldsymbol{\beta}$ 的后验方差接近 $\sigma^2 (X^\top X)^{-1}$, 其中 σ^2 是 Y 的方差。如果该分布产生的 Markov chain 的接受比率过高或过低，总可以相应调整 proposal distribution 中的方差。实现本例 Metropolis 算法的 R code 如下。

```
n <- length(y)
p <- dim(X)[2]

pmn.beta <- rep (0 ,p) #prior expectation
psd.beta <- rep (10 ,p) #prior sd

var.prop <- var ( log (y+1/2) ) * solve ( t(X) %*% X ) #proposal var

S <- 10000 #length of Markov chain
beta <- rep(0,p) #initial value
acs <- 0 #number of acceptances
BETA <- matrix(0, nrow=S, ncol=p)

# Metropolis algorithm
```



```
set.seed (1)
for(s in 1:S){
  beta.p <- t( rmvnorm(1, beta, var.prop) )

  log_ar <- sum( dpois(y, exp(X %*% beta.p), log=T) ) -
            sum( dpois(y, exp(X %*% beta), log=T) ) +
            sum( dnorm(beta.p, pmn.beta, psd.beta, log=T) ) -
            sum( dnorm(beta, pmn.beta, psd.beta, log =T) )

  if( log(runif(1)) < log_ar ){
    beta <- beta.p
    acs <- acs+1
  }

  BETA[s,] <- beta
}
```

代入雌雀的数据，上述 Metropolis 算法得到的 Markov chain 的接受比率为 43%. 图3的左图展示了 β_3 对应的 Markov chain 的 traceplot. 可以看到该 Markov chain 很快从初始值 0 移动到 posterior mode 附近。图3中间的图展示了该 Markov chain 的 autocorrelation function (ACF), 可以看到相邻样本的自相关系数很高。如果我们从该 Markov chain 上每 10 步取一个样本组成一条 *thinned* Markov chain，图3展示了新的 Markov chain 的 ACF, 可以看到此时样本间的相关性很小。经过 thinning 的 Markov chain 只保留了原 Markov chain 上 10000 个样本中的 1000 个，但这 1000 个样本接近相互独立，它们的 ESS 为 726. 这些样本对于估计本例的后验分布是足够的，图4的左图用虚线展示了由离散网格法估计的 β_3 的边际后验分布的 pdf，这与从上述 thinned Markov chain 上估计的边际分布几乎完全相同。

最后，图4的右图展示了雌雀在年龄 x 下期望繁殖的后代数 $E(Y \mid x)$ 的 posterior median 及 95% credible interval, 这些结果反映出该雀种产生的后代数与雌雀年龄的二次函数的关系。

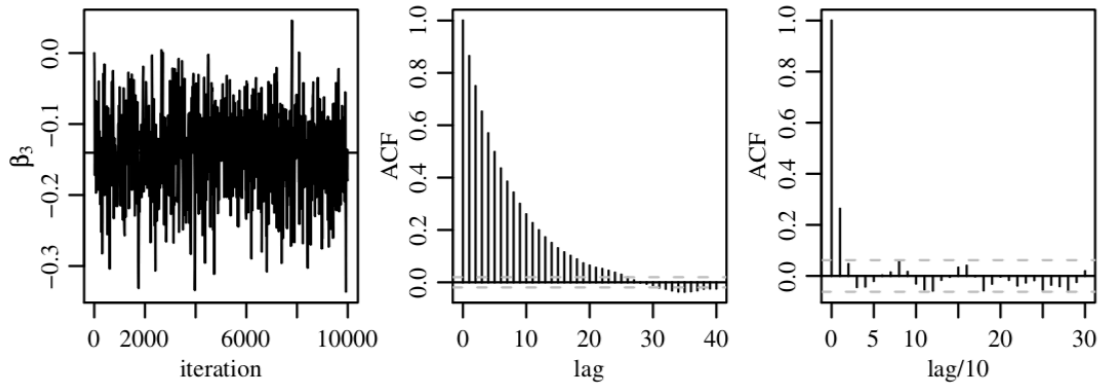

Figure 3: β_3 的 Markov chain 及其 autocorrelation functions. Picture sourse: Hoff (2009)

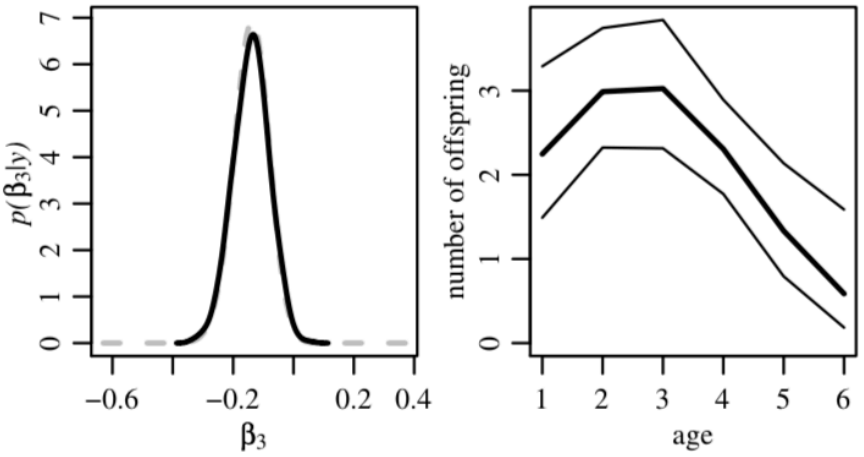

Figure 4: 左图：实线代表由 β_3 的 thinned Markov chain 估计的边际 pdf，虚线是由数值方法估计的 β_3 的边际 posterior density; 右图：从上到下的三条线分别对应 $\exp(\boldsymbol{x}^\top\boldsymbol{\beta})$ 的 2.5%, 50% 和 97.5% posterior quantiles. Picture sourse: Hoff (2009)

4 Metropolis-Hastings 算法

前面介绍的 Gibbs sampler 和 Metropolis 算法是两种使用 Markov chain 近似目标分布的方法。事实上，它们是更一般的 Metropolis-Hastings (M-H) 算法的两个特例。M-H 算法与 Metropolis 算法很相似，它也需要在每步迭代中从 proposal distribution 产生一个候选样本，然后按照一定规则接受或拒绝该样本，但是 M-H 算法允许任何形式的 proposal distribution, 不一定是对称的条件分布。

假设我们的目标分布是参数 $\boldsymbol{\theta}$ 的后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$. 选取 proposal distribution $\tilde{g}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ 后，

M-H 算法如下产生样本 $\boldsymbol{\theta}^{(t+1)}$:

1. 抽取 $\boldsymbol{\theta}^\star \sim \tilde{g}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$

2. 计算接受比率

$$r_t = \frac{p(\boldsymbol{\theta}^\star \mid \boldsymbol{y})}{p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})} \cdot \frac{\tilde{g}(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^\star)}{\tilde{g}(\boldsymbol{\theta}^\star \mid \boldsymbol{\theta}^{(t)})} = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta}^\star)p(\boldsymbol{\theta}^\star)\tilde{g}(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^\star)}{p(\boldsymbol{y} \mid \boldsymbol{\theta}^{(t)})p(\boldsymbol{\theta}^{(t)})\tilde{g}(\boldsymbol{\theta}^\star \mid \boldsymbol{\theta}^{(t)})}.$$

3. 令

$$\boldsymbol{\theta}^{(t+1)} = \begin{cases} \boldsymbol{\theta}^\star & \text{概率 } \min(r_t, 1) \\ \boldsymbol{\theta}^{(t)} & \text{概率 } 1 - \min(r_t, 1) \end{cases}$$

M-H 算法的适用范围更广，因为对称的 proposal distribution 有时并不合理，比如对于方差参数，对称的 proposal distribution 可能无法保证产生的样本为正。与 Metropolis 算法类似，我们可以用 detail balance 条件证明 M-H 算法产生的 Markov chain 收敛到参数的后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$。

- 如果 $\boldsymbol{\theta}^{(t+1)} \neq \boldsymbol{\theta}^{(t)}$，从 $\boldsymbol{\theta}^{(t)}$ 到 $\boldsymbol{\theta}^{(t+1)}$ 的 transition density 为

$$\begin{aligned} p(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) &= \tilde{g}(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) P(\boldsymbol{\theta}^{(t+1)} \text{ is accepted}) \\ &= \tilde{g}(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) \min\left\{ \frac{p(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{y})}{p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})} \cdot \frac{\tilde{g}(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t+1)})}{\tilde{g}(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)})}, 1 \right\} \end{aligned}$$

  此时

$$p(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y}) = \min\{p(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{y})\tilde{g}(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t+1)}),\, p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})\tilde{g}(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)})\} \quad (7)$$

  注意到(7)的右端关于 $(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)})$ 对称，所以

$$p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t+1)}) p(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{y}) = p(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y}), \quad \forall \boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t+1)}$$

  即 detail balance 条件关于后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 成立。

- 如果 $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)}$，不论 transition density 是何种形式，detail balance 条件对后验分布 $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ 总是成立的，因为 $p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y}) = p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) p(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{y})$.

Remarks

1. 在 M-H 算法中，如果选取的 proposal distribution $\tilde{g}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$ 是一个对称的条件分布，则算法退化为 Metropolis 算法。与 Metropolis 算法类似，M-H 算法产生的 Markov chain 也可能出现重复的样本。

2. 如果 M-H 算法的目标分布 $\pi(\boldsymbol{\theta})$ 可以分解为如下两部分：

$$\pi(\boldsymbol{\theta}) \propto \alpha(\boldsymbol{\theta}) g(\boldsymbol{\theta})$$

其中 $g(\boldsymbol{\theta})$ 在 $\pi(\boldsymbol{\theta})$ 中占主导地位且对应一个容易抽样的分布。可将 proposal distribution 选为 $g(\boldsymbol{\theta})$, 此时 M-H 算法的接受比率简化为

$$r_t = \frac{\pi(\boldsymbol{\theta}^\star)}{\pi(\boldsymbol{\theta}^{(t)})} \cdot \frac{g(\boldsymbol{\theta}^{(t)})}{g(\boldsymbol{\theta}^\star)} = \frac{\alpha(\boldsymbol{\theta}^\star)}{\alpha(\boldsymbol{\theta}^{(t)})}.$$

References

Hoff, P. D. (2009). *A first course in Bayesian statistical methods.* Springer Science & Business Media.

