The Application of MCMC in Bayesian Estimation

Chen Zhiyuan
School of Statistics, Beijing Normal University

December 2021

目录

摘要

本文主要介绍了 MCMC 在贝叶斯统计中的应用。贝叶斯统计离不开后验分布期望的计算，但当参数是多维时积分难以计算。于是想到用蒙特卡洛方法：从后验分布中抽取，再用均值作为期望的估计。而马尔可夫链的性质保证了算法收敛。

显然，越耦合，越复杂，越多维的问题越能体现 MCMC 方法在贝叶斯统计应用中的优势和典型。本文选择截断泊松分布——即若干段强度不同的泊松分布作为问题背景进行模拟。

模拟一共使用了三种方法：带舍选抽样法的 Gibbs 算法、带 MH 算法的 Gibbs 算法和 Hybrid MCMC 算法。三种方法是循序渐进的，前两种方法是一般性的思路，但囿于细节不能解决这个问题，第三种方法给出了正确的参数估计，并且运算速度很快，估计方差很小。

1 引言

在数理统计中，总体 $X \sim f(X,\theta)$，有 i.i.d 样本 $X_1, X_2...X_n$，如何估计 θ?

我们有两种方法。频率学派的做法是极大似然估计：认为 θ 是常数，$\hat{\theta}_{mle} = argmaxL(\theta, x)$。而贝叶斯学派的做法是贝叶斯估计：认为 $\theta \sim h(\theta)$，$f(\theta, x) = f(x|\theta)$，故：

$$\begin{aligned} k(\theta|x_1...x_n) &= \frac{g(\theta, x_1...x_n)}{f(x_1...x_n)} \\ &= \frac{\prod_1^n f(x_i|\theta)h(\theta)}{\int \prod_1^n f(x_i|\theta)h(\theta)dx} \end{aligned}$$

然后选择合适的损失函数积分，比如 $E(\theta|x)$ 是 θ 的一个好的估计。

但是当 $\theta = (\theta_1...\theta_d)$ 且每个分布都不同时，高维积分的计算是十分复杂的，无论是边缘分布还是后续的期望计算都很困难。这时我们自然想到用统计模拟的方法来近似计算积分：用不同方法有效地从后验概率 $k(\theta|x)$ 中抽样，再用样本均值作为期望的估计。

$$\frac{1}{n}\sum_{i=1}^{n}\theta_i \to E(\theta|x)$$

$$\frac{1}{n}\sum_{i=1}^{n}L(\theta_i) \to \int L(\theta|x)f(\theta|x)dx$$

2 问题

截断泊松分布在生活中非常常见，它指所观测的随机变量在不同时段可能服从强度不同的泊松分布。例如，单位时间内来到食堂的人数服从泊松分布。那么在 11：40 下课后，去食堂的人会陡然增多，单位时间内来到食堂的人数将服从一个强度更大的泊松分布。如何估计各段的强度呢?

若使用矩估计的办法，则需要先确定断点的位置，再分别计算两段的均值即可作为强度的估计。矩估计方便又有效，但在有先验信息的时候不如贝叶斯估计更高效。

指定真实值为 $\theta_{true} = 3$，$\lambda_{true} = 3$，$true = 76$。设观测变量 $Y = (Y_1...Y_n)$ 有如下分布：

$$f(y_i|k,\theta,\lambda) = \frac{\theta^{y_i}e^{-\theta}}{y_i!} \quad , \quad i = 1,2,...,k$$
$$f(y_i|k,\theta,\lambda) = \frac{\lambda^{y_i}e^{-\lambda}}{y_i!} \quad , \quad i = k+1,k+2,...,n$$

容易看出，Y 服从截断的泊松分布：在 k 之前强度为 θ，在 k 之后强度为 λ。

设各参数的先验分布为：

$$\theta|b_1 \sim \Gamma(0.5, b_1)$$
$$\lambda|b_2 \sim \Gamma(0.5, b_2)$$
$$h(b_1) = IG(1,1) = \frac{e^{-\frac{1}{b_1}}}{b_1^2}$$
$$h(b_2) = IG(1,1) = \frac{e^{-\frac{1}{b_2}}}{b_2^2}$$
$$k \sim U(1,2,...,n)$$
$$IG(\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha x^{\alpha+1}}e^{-\frac{1}{\beta x}}$$

3 MCMC

3.1 方法论

无论用 MH 或 Gibbs 方法从后验概率中抽样，首先要计算目标分布函数。同时注意到：分母上的边缘分布对于后验概率而言是常数。

$$\begin{aligned} k(\theta|x) &= \frac{g(\theta, x_1...x_n)}{f(x_1...x_n)} \\ &\propto g(\theta, x_1...x_n) \\ &= cg(\theta, x) \end{aligned}$$

当我们以 $k(\theta|x)$ 为目标构造马尔可夫链时，使其满足细致平稳条件的接受率为：（与 1 取 min 省略）

$$\begin{aligned} \alpha(i,j) &= \frac{q(j)\,P(j|i)}{q(i)\,P(i|j)} \\ &= \frac{k(\theta_j|x)\,P(j|i)}{k(\theta_i|x)\,P(i|j)} \\ &= \frac{c\,g(\theta_j|x)\,P(j|i)}{c\,g(\theta_i|x)\,P(i|j)} \\ &= \frac{g(\theta_j|x)\,P(j|i)}{g(\theta_i|x)\,P(i|j)} \end{aligned}$$

由于。结果等同于以联合分布 $g(\theta|x)$ 为目标抽取样本，而联合分布只需要简单的乘法就可得到，这就不需要计算后验概率的具体表达式，避免了复杂积分。此时已经可以直接使用 MH 方法抽取 θ。但对于这个问题，估计多维参数时 MH 方法会很低效，故选择 Gibbs 抽样。

3.2 算法

3.2.1 后验分布与条件分布的推导

计算得到联合分布为:

$$\begin{aligned} g(Y,k,\theta,\lambda) &= \prod_1^n f(y_i|\theta,k,\lambda)h(\theta)h(\lambda)p_k \\ &= \prod_1^n f(y_i|\theta,k,\lambda)h(\theta|b_1)h(b_1)h(\lambda|b_2)h(b_2)p_k \\ &= \prod_1^k \frac{\theta^{y_i}e^{-\theta}}{y_i}\cdot\prod_{k+1}^n \frac{\lambda^{y_i}e^{-\lambda}}{y_i}\cdot\frac{\theta^{-0.5}e^{-\frac{\theta}{b_1}}}{\Gamma(0.5)b_1^{\frac{1}{2}}}\cdot\frac{\lambda^{-0.5}e^{-\frac{\lambda}{b_2}}}{\Gamma(0.5)b_2^{\frac{1}{2}}}\cdot\frac{e^{-\frac{1}{b_1}}e^{-\frac{1}{b_2}}}{b_1^2b_2^2}\cdot\frac{1}{n} \end{aligned}$$

联合分布是我们 MCMC 的目标，但操作 Gibbs 抽样还需要得到全部的条件分布，这只需将目标视作自变量而条件视作参数（即联合分布中目标的函数结构与何种分布相符合，参数配平，视为常数者舍去），比如:

$$\begin{aligned} f(\theta|k,\lambda,b_1,b_2,Y) &\propto g(Y,k,\theta,\lambda) \\ &\propto \theta^{\sum_1^k y_i}e^{-k\theta}\theta^{-0.5}e^{-\frac{\theta}{b_1}} \\ &= \theta^{\sum_1^k y_i-0.5}e^{-\frac{kb_1+1}{b_1}\theta} \\ &\propto \frac{\theta^{\sum_1^k y_i-0.5}e^{-\frac{kb_1+1}{b_1}\theta}}{\Gamma(\sum_1^k y_i+0.5)(\frac{b_1}{kb_1+1})^{\sum_1^k y_i+0.5}} \\ &\propto \Gamma(\sum_1^k y_i+0.5,\frac{b_1}{kb_1+1}) \end{aligned}$$

同理，其他条件分布为:

$$\begin{aligned} f(\lambda|k,\theta,b_1,b_2,Y) &\propto \Gamma(\sum_{k+1}^n y_i+0.5,\frac{b_2}{(n-k)b_2+1}) \\ f(b_1|k,\theta,\lambda,b_2,Y) &\propto IG(1.5,\frac{1}{1+\theta}) \\ f(b_2|k,\theta,b_1,\lambda,Y) &\propto IG(1.5,\frac{1}{1+\lambda}) \\ p_{k|\lambda,\theta,b_1,b_2,Y} &= \theta^{\sum_1^k y_i}\lambda^{\sum_{k+1}^n y_i}e^{(\lambda-\theta)k-n\lambda} \end{aligned}$$

3.2.2 生成随机数

需要抽取的分布中有些并不是我们平时常用的分布，需要自行模拟。

3.2.2.1 逆变换法 用逆变换法模拟生成 k 的初始值，因为 k 服从离散均匀分布，分布函数较好计算。

算法 3.1: 生成服从离散均匀分布的 k

$$gen\ \ u\sim U(0,1)$$
$$k=[n*u]+1$$

3.2.2.2 随机变量的映射 b_1, b_2 服从逆伽马分布：$IG(\alpha, \beta)$。可以证明，对随机变量 $x \sim IG(\alpha, \beta)$ 有 $\frac{1}{x} \sim Gamma(\alpha, \beta)$

算法 3.2: 生成密度为 $IG(\alpha, \beta)$ 的 b_1, b_2

$$gen\ x \sim Gamma(\alpha, \beta)$$
$$b_i = \frac{1}{x}, \quad i = 1, 2$$

3.2.2.3 舍选抽样法 $p_{k|\lambda,\theta,b_1,b_2,Y} \propto \theta^{\sum_1^k y_i} \lambda^{\sum_{k+1}^n y_i} e^{(\lambda-\theta)k-n\lambda}$ 不是常见离散分布的密度，也难以计算其分布函数，故用舍选抽样法模拟。

令 $p_{k|\lambda,\theta,b_1,b_2,Y} = \theta^{\sum_1^k y_i} \lambda^{\sum_{k+1}^n y_i} e^{(\lambda-\theta)k}$ 为目标密度，其中 $\lambda - \theta$ 必须小于 0，则 $e^{\lambda-\theta} < 1$。由于其形式与几何分布的密度接近，选取辅助密度服从参数 $q = e^{\lambda-\theta-n\lambda}$ 的几何分布，$q_k = (1 - e^{(\lambda-\theta)})e^{(\lambda-\theta)k}$。计算 $c = max\frac{p_k}{q_k}$：

$$\frac{p_k}{q_k} = \frac{\theta^{\sum_1^x y_i} \lambda^{\sum_{1+x}^n y_i}}{(1 - e^{\lambda-\theta})e^{n\lambda}} \leq \frac{\theta^{\sum_1^n y_i}}{(1 - e^{\lambda-\theta})e^{n\lambda}}$$

由于上式的最大值较为难求，且每轮转移 λ, θ 的值都会变化，故采用放缩法令 $c = \frac{\theta^{\sum_1^n y_i}}{(1 - e^{\lambda-\theta})e^{n\lambda}}$

算法 3.3: 生成 $k|\lambda, \theta, b_1, b_2, Y$

$$if\ \lambda - \theta < 0: \quad continue$$
$$else \quad regen\ \lambda, \theta$$
$$repeat:$$
$$gen\ k \sim geom(e^{\lambda-\theta}); \quad gen\ u \sim U(0, 1)$$
$$p_k = \theta^{\sum_1^k y_i} \lambda^{\sum_{k+1}^n y_i}; \quad q_k = (1 - e^{(\lambda-\theta)})e^{(\lambda-\theta)k}$$
$$if\ u < \frac{p_k}{cq_k} = (\frac{\lambda}{\theta})^{\sum_{k+1}^n y_i}: \quad return\ k,\ break$$
$$else \quad continue$$

3.3 模拟

首先，从先验分布中抽取一组 $(\lambda^0, k^0, \theta^0, b_1^0, b_2^0)$，然后再根据上述条件分布（由 $(\lambda^t, k^t, \theta^t, b_1^t, b_2^t)$ 和样本 $(Y_1 ... Y_n)$ 计算）依次抽取 $(\lambda^{t+1}, k^{t+1}, \theta^{t+1}, b_1^{t+1}, b_2^{t+1})$，舍去未收敛的样本即可。但在实际抽取时遇到了问题：

（1）程序运行太慢，无法得到结果。经过检查，原因可能是舍选抽样法抽 k 的效率太低，因为表达式带指数形式，很容易出现数值非常小超出 R 语言精度，被自动压缩为 0 的情况。但就 k 密度的复杂表达式而言也很难有比几何分布更好的选择。

（2）抽出的参数可能超出本身的取值范围，于是需要在程序中加入判定条件如 $\lambda < \theta$ 和 $k \leq n$，这实际上是增加了条件，改变了分布。在这种情况下 Gibbs 抽样能否正确收敛既没有理论保证，也难以模拟验证。

4 带 MH 的 Gibbs 抽样

用 MH 算法抽取 k 和用舍选抽样法抽取 k 本质相同，但不用计算常数 c，故尝试以 $q = \frac{1}{n}$ 作为转移概率，即建议分布还是离散均匀分布。

4.1 算法

算法 4.1: 用 MH 算法抽取 k

$$gen\ k_{new} \sim U(1, ..., n)$$

$$gen\ u \sim U(0, 1)$$

$$log(\rho) = log(\theta)(\sum_{1}^{k_{new}} y_i - \sum_{1}^{k} y_i) + log(\lambda)(\sum_{k_{new}+1}^{n} y_i - \sum_{k+1}^{n} y_i) + (\lambda - \theta)(k_{new} - k)$$

$$log(\alpha) = min(log(\rho), 0)$$

$$if\ log(u) < \alpha: \quad k = k_{new}$$

由于密度中指数运算多，故对接受率 α 取对数。

4.2 模拟

模拟速度非常快，结果如下：

结果 4.1: MH-Gibbs 的参数估计结果

	mean	var
θ	0.1370541	5.5046896
λ	0.1370541	5.5046896

结果却很奇怪，不仅估计不准确，而且两个参数的估计是一样的。查看向量发现结果并未收敛，且将重复次数提升到 5000 后仍没有收敛迹象。猜测原因为 k 的抽取还是没有收敛，新换的建议分布并没有解决问题。而且可能由于建议分布是均匀的导致样本被平滑了，所以两个参数的估计相同。可是为什么结果比两个真实值都小呢？这一点还未想明白。

5 Hybrid MCMC

之所以称这个新方法为“杂交马尔可夫链蒙特卡洛”，是打算用其他统计方法先给出 k 的估计，并将其作为常数参数输入后验分布中（与样本 $(Y_1...Y_n)$ 的角色相似），再执行 MCMC 抽样。

5.1 k 的确定

5.1.1 画图初筛

已知 k 是分割两个不同强度的泊松分布的断点，那么以 k 为分界前后的数据均值应该有显著差异。绘制前 m 个样本的均值曲线如图所示：

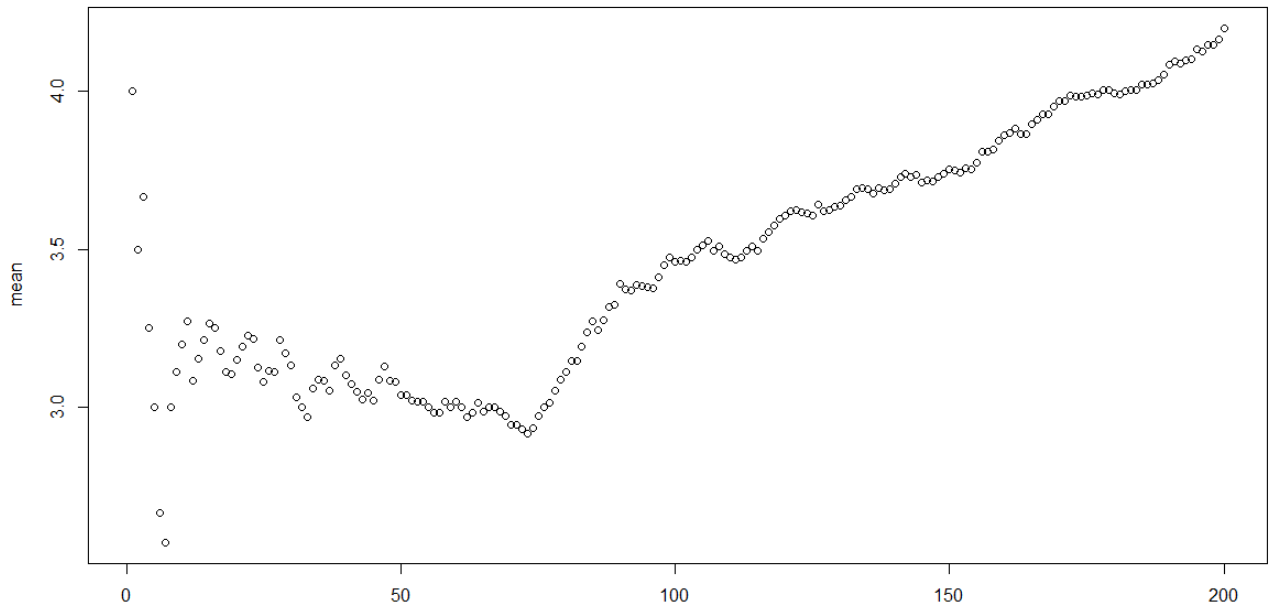

图 1: 前 m 个样本的均值

可以看到曲线的拐点大概发生在 70 附近，初步猜测从拐点开始泊松分布的强度变大了导致新数据加入后整个均值开始不断上升。

5.1.2 非参数检验

已经确定最低点 $m = 74$，现在检验最低点左右数据的均值是否有显著差异。如果有显著差异则可以认为两部分数据来自不同的泊松分布，即 74 是 m 的一个好的估计。

由于数据不是来自正态分布，用方差分析的 F 检验可能有谬误。使用非参数的 Kruskal-Wallis 检验，结果如下：

结果 5.1: 非参数检验的结果

Kruskal-Wallis rank sum test

Chi-square	df	p-value
45.891	1	$1.25 * 10^{-11}$

p 值非常非常小，显著拒绝两总体相同的原假设，我们即可认为 $\hat{m} = 74$ 是可信的。

5.2 Gibbs 抽样

重新回到 MCMC 的框架下，我们可以直接将样本分为两组 $Y_1, ..., Y_k$ 和 $Y_{k+1}, ..., Y_n$，这样只需进行两个泊松分布的参数估计，每次估计两个参数。

计算：

$$S_1 = \sum_{i=1}^{k} Y_i$$
$$S_2 = \sum_{i=k+1}^{n} Y_i$$

从先验分布中抽取一组 $(\theta^0, b_1^0), (\lambda^0, b_2^0)$，结合常量 $S-1, S_2, k$，按下算法迭代至收敛：

算法 5.1: 生成 $\theta|b_1, S_1, k$

$$repeat\ for\ 1000\ times:$$
$$gen\ b_1^{(t+1)} \sim IG(1.5, \frac{1}{1+\theta^{(t)}})$$
$$gen\ \theta^{(t+1)} \sim \Gamma(S_1 + 0.5, \frac{b_1^{(t)}}{kb_1^{(t)}+1})$$

算法 5.2: 生成 $\lambda|b_2, S_2, k$

$$repeat\ for\ 1000\ times:$$
$$gen\ b_2^{(t+1)} \sim IG(1.5, \frac{1}{1+\lambda^{(t)}})$$
$$gen\ \lambda^{(t+1)} \sim \Gamma(S_2 + 0.5, \frac{b_2^{(t)}}{(n-k)b_2^{(t)}+1})$$

模拟结果为：

结果 5.2: Hybrid Gibbs 的参数估计结果

	mean	var
θ	2.92405224	0.03874388
λ	4.93003518	0.03937806

运算速度很快，参数估计结果非常接近真实值，而且方差很小。

6 结论

MCMC 方法可以有效解决贝叶斯统计中后验概率积分难算的问题，运算速度并不显著输于矩估计和极大似然估计，但随着迭代次数的增加方差可以逐渐减小，变得越来越精确。在复杂的，多维的参数估计应用中非常有效。

但是，我们也应该注意到一些问题。后验分布并不一定是常见的，或者是容易用逆变换法生成的分布，这时我们不得不依赖 MH 或者舍选抽样法来单独抽取这个分布。这时建议分布的选取十分重要！虽然马尔可夫链理论上保证一定收敛，但计算机有精度限制，极低的抽取效率会造成程序事实上的无效。

当然，如果像在这个实例中遇到的让人头疼的变点 k 一样，很难找到令人满意的建议分布时，可以考虑使用“杂交的 MCMC”方法。即用贝叶斯框架以外的方法先识别变点的位置，然后对剩下的参数执行 MCMC。结果证明这种思路是可行的。我很喜欢这种灵活的思路，它使得泛化性极强的 MCMC 方法更加“圆滑”了，能够在更具体的场景发挥最大效率。但我并不推荐首先使用这种方法，因为最终结果的的精度高度依赖变点识别的精度，而后者有经验的成分存在（即使在变点识别这个问题上有非常的的统计方法可以参考，大家好像也更倾向于作经验判断）。

7 参考文献

[1] Sheldon M.Ross. Simulation. 2006

[2] Hogg, McKean, Craig. Introduction to Mathematical Statistics. 2014

[3] 李勇. 概率论. 2013

[4] 韦来生. 贝叶斯统计. 2016

[5] 肖枝洪, 朱强. 统计模拟及其 R 实现. 2010

[6] John D. Cook. Inverse Gamma Distribution. http://www.johndcook.com/inversegamma.pdf, 2008