Statistics and Application 统计学与应用, 2018, 7(2), 247-255
Published Online April 2018 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2018.72029

Hans汉斯

Application of Hierarchical Bayesian Model Based on Gibbs Sampling in Statistical Inference of Fire Occurrences

Kang Cao

Shanghai Maritime University, Shanghai
Email: 347572701@qq.com

Received: Apr. 3rd, 2018; accepted: Apr. 21st, 2018; published: Apr. 28th, 2018

Abstract

Gibbs sampling method is the most widely used method in MCMC algorithm. The basic idea of Gibbs sampling is to construct the Markov chain by the conditional distribution family of the components of the parameter vector when the high-dimensional parameters are posteriorly inferred, so that its invariant distribution is the target distribution. This topic is based on the method to determine the parameters of the model, which can be based on existing information to estimate the number of years, the number of fire occurred in the region and the estimated confidence interval of the parameters.

Keywords

Gibbs Sampling, Hierarchical Bayesian Model, Markov Chain, Fire

基于Gibbs抽样的分层贝叶斯模型在火灾发生次数统计推断中的应用

曹　康

上海海事大学，上海
Email: 347572701@qq.com

收稿日期：2018年4月3日；录用日期：2018年4月21日；发布日期：2018年4月28日

摘　要

Gibbs抽样是MCMC抽样算法中应用最广泛的方法之一，其核心思想是对高维参数进行后验推断时，通过

文章引用：曹康. 基于 Gibbs 抽样的分层贝叶斯模型在火灾发生次数统计推断中的应用[J]. 统计学与应用, 2018, 7(2): 247-255. DOI: 10.12677/sa.2018.72029

参数向量的分量的条件分布族来构造Markov链，使其不变分布为目标分布。本文利用Gibbs抽样方法结合分层贝叶斯模型，对我国各地区火灾发生次数进行了回测，结果显示，相比于传统的poisson分布刻画方法，基于Gibbs抽样的分层贝叶斯方法充分利用了历史信息使结果更具可信度。

关键词

Gibbs抽样，分层贝叶斯模型，马尔科夫链，火灾

1. 引言

近年来，由于我国在城镇化进程中的快速推进，城市规模和人口越来越多，因此而产生的城市问题也越来越多，火灾即是其中之一。据公安部消防局公布的数据，2016 年全国共接报火灾 31.2 万起，造成死亡 1582 人，伤 1065，直接财产损失 37.2 亿元。可见，火灾对市民人身和财产安全造成了巨大的危害，如何寻找出火灾易发区域，采取有效措施及时防范已是一项十分艰巨和重要的任务。

影响火灾发生率的因素有很多，如 Schaeman [1] (1997)认为经济越发达的地区，火灾发生次数可能越少；同样，杨立中[2] (2003)也研究得出经济水平欠发达地区发生火灾的次数相应也较高。此外，也有很多学者也关注于对火灾影响因素进行量化分析，如邓欧[3] (2012)等建立了 Logistic 全局火灾预报回归模型；Bisquert [4]等(2012)利用人工神经网络对火灾发生率进行建模；Dlamini [5] (2011)也建立了基于贝叶斯网络的火灾发生率模型。总的来说，目前已有研究中无论是 Logistic 回归、人工神经网络还是贝叶斯都关注的是传统影响因素下模型精度的问题，对火灾发生率的空间分布考虑较少。

本文即是在此背景下，从统计抽样的角度，利用贝叶斯思想，建立分层贝叶斯模型，再利用 Gibbs 抽样方法得到不同地区火灾发生次数的 Markov 链，借此分析了不同地区火灾发生次数的分布特点。

2. 分层贝叶斯模型

分层贝叶斯模型[6]是贝叶斯模型的一种，用来为具有不同水平的问题进行建模，通过贝叶斯方法估计后验分布的参数。很多时候模型会具有多个参数，这些参数也有可能具有结构性的联系。例如在研究第 i 次射击命中 10 环的概率 λ_i，很显然我们预估 λ_i 是相互联系的，利用贝叶斯方法，将 λ_i 看做总体分布的一个样本，观测数据 y_{ij}，其中 j 为第 j 次试验，利用观测数据可以用来估计 λ_i 的分布。这样就构成了一个分层的贝叶斯模型。分层贝叶斯模型是通过计算参数在已知观恶量下的条件后验概率，推到过程为：

1) 写出联合后验密度 $p(\theta,\varphi|y)$，其非正规化的形式是超先验分布 $p(\varphi)$、总体分布 $p(\theta|\varphi)$和似然函数 $p(y|\theta)$的乘积。

2) 在给定超参数 φ 的情况下，确定 θ 的条件后验密度，固定观测值 y 的情况下，它是 φ 的函数 $p(\theta|\varphi, y)$。

3) 使用贝叶斯分析范例估计 φ，也就是要获取边缘后验分布 $p(\varphi|y)$。

3. Gibbs 抽样

3.1. Gibbs 抽样原理

Gibbs 抽样[7]简单、应用最广泛的 MCMC 抽样方法之一，应用该抽样方法的前提是要分布 $\pi(\boldsymbol{x})$的满

条件分布已知，即对于任意 i，在已知 x 的第 i 个分量以外其他分量值的条件下，第 i 个分量的条件分布已知。在给定初值点 $x(0)=\left(x(0)_1,\cdots,x(0)_p\right)$ 后，假定第 t 次迭代值为 $\boldsymbol{x}(t)$，则第 $t+1$ 次迭代分为如下 p 步(这里 p 表示 x 共有 p 个分量)。具体算法为：

给定 $x^{(t)}=\left(x_1^t,x_2^t,\cdots,x_p^t\right)$，

(1) 生成 $X_1^{(t+1)}\sim\pi\left(x_1\mid x_2^{(t-1)},\cdots,x_p^{(t-1)}\right)$，

…

(i) 生成 $X_1^{(t+1)}\sim\pi_i\left(x_i\mid x_1^{(t)},\cdots,x_{i-1}^{(t)},x_{i+1}^{(t-1)},\cdots,x_p^{(t-1)}\right)$，

…

(p) 生成 $X_p^{(t+1)}\sim\pi_p\left(x_p\mid x_1^{(t)},\cdots,x_{p-1}^{(t)}\right)$

3.2. Gibbs 抽样的具体实现方法

设 $\theta=\left(\theta_1,\theta_2,\cdots,\theta_p\right)'$ 是 p 维参数向量，$\pi\left(\theta\mid D\right)$ 是观察到数据集 D 后 θ 的后验分布。Gibbs 抽样方法[8]如下：

第 0 步.任意选取一个初始点 $\theta^{(0)}=\left(\theta_{1,0},\theta_{2,0},\cdots,\theta_{p,0}\right)'$，并置 $i=0$；

第 1 步.按下列方法生成 $\theta^{(i+1)}=\left(\theta_{1,i+1},\theta_{2,i+1},\cdots,\theta_{p,i+1}\right)'$：

生成 $\theta_{1,i+1}\sim\pi\left(\theta_1\mid\theta_{2,i},\cdots,\theta_{p,i},D\right)'$，

生成 $\theta_{2,i+1}\sim\pi\left(\theta_2\mid\theta_{1,i+1},\theta_{3,i},\cdots,\theta_{p,i},D\right)'$，

…

生成 $\theta_{p,i+1}\sim\pi\left(\theta_p\mid\theta_{1,i+1},\theta_{2,i+1},\cdots,\theta_{p-1,i+1},D\right)$；

第 2 步。置 $i=i+1$，并返回到第 1 步.

在这个算法过程中，θ 的每一个分量按照自然顺序生成，每一个循环需要生成 p 个随机变量。

4. 应用实例

4.1. 案例介绍

本文根据 2012 年全国各地区发生火灾的次数(见表 1，数据来源于公安部消防局中国火灾消防统计年鉴)为样本，使用分层贝叶斯方法对各地区的火灾发生次数进行统计推断，具体数据如表 1 所示。

4.2. 模型建立

假设第 i 地区发生火灾的次数服从参数为 $\lambda_i\left(1\le i\le 31\right)$ 的 poisson 分布。对于观察时间 t_i，发生火灾次数 X_i 服从参数为 $\lambda_i t_i\left(1\le i\le 31\right)$ 的 poisson 分布 $P\left(\lambda_i t_i\right)$。考虑到 gamma 分布是 poisson 分布和 gamma 分布的共轭先验分布，故取参数 λ_i 和 β 服从 gamma 分布。综上，考虑下列分层贝叶斯模型：

$$
\begin{aligned}
&X_i\sim P\left(\lambda_i t_i\right),i=1,2,\cdots,31,\\
&\lambda_i\sim\Gamma\left(\alpha,\beta\right),i=1,2,\cdots,31,\\
&\beta\sim\Gamma\left(\gamma,\delta\right).
\end{aligned}
$$

则，各层的条件密度分别为：

Table 1. Number of fire occurrences by region in 2012
表 1. 2012 年各地区火灾发生次数数据

	次数(千次)	地区	次数(千次)
北京	3.409	湖北	4.962
天津	2.213	湖南	10.399
河北	5.012	广东	8.154
山西	3.897	广西	1.386
内蒙古	7.545	海南	0.688
辽宁	8.265	重庆	3.758
吉林	5.652	四川	6.899
黑龙江	5.794	贵州	0.959
上海	4.469	云南	1.251
江苏	7.739	西藏	0.192
浙江	3.5	陕西	7.857
安徽	5.653	甘肃	4.434
福建	5.698	青海	1.054
江西	3.79	宁夏	3.304
山东	11.918	新疆	7.286
河南	5.11		

$$f\left(x_i \mid \lambda_i, t_i\right)=\frac{\left(\lambda_i t_i\right)^{x_i}}{x_i!}\mathrm{e}^{-\lambda_i t_i}, i=1,\cdots,31, x_i=0,1,2,\cdots,$$

$$\pi\left(\lambda_i \mid \alpha,\beta\right)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda_i^{\alpha-1}\mathrm{e}^{-\beta\lambda_i}, i=1,\cdots,31,$$

$$\pi\left(\beta \mid \gamma,\delta\right)=\frac{\delta^{\gamma}}{\Gamma(\gamma)}\beta^{\gamma-1}\mathrm{e}^{-\delta\beta}.$$

参数 $\left(\lambda_1,\cdots,\lambda_{10},\beta\right)$ 联合后验分布为：

$$\begin{aligned}&\pi\left(\lambda_1,\cdots,\lambda_{10},\beta \mid t_1,\cdots,t_{10},p_1,\cdots,p_{10}\right)\\ &\propto\prod_{i=1}^{10}\left\{\left(\lambda_i t_i\right)^{x_i}\mathrm{e}^{-\lambda_i t_i}\lambda_i^{\alpha-1}\mathrm{e}^{-\beta\lambda_i}\right\}\beta^{10\alpha}\beta^{\gamma-1}\mathrm{e}^{-\delta\beta}\\ &\propto\prod_{i=1}^{10}\left\{\lambda_i^{x_i+\alpha-1}\mathrm{e}^{-\left(t_i+\beta\right)\lambda_i}\right\}\beta^{10\alpha+\gamma-1}\beta^{\gamma-1}\mathrm{e}^{-\delta\beta}.\end{aligned}$$

各参数的全条件后验分布为：

$$\lambda_i \mid \beta,t_i,x_i \sim \Gamma\left(x_i+\alpha,t_i+\beta\right), i=1,\cdots,31,$$

$$\beta \mid \lambda_1,\cdots,\lambda_{10} \sim \Gamma\left(31\alpha+\gamma,\delta+\sum_{i=1}^{31}\lambda_i\right).$$

4.3. 算法的实现与结果

该 Gibbs 抽样是直接从各参数的全条件后验中进行抽样的，取超参数 $\alpha=0.001,\gamma=0.001,\delta=1.1$，进行 10,000 次迭代，结果见图 1、图 2 (由于地区较多，这里只给出了表 1 中的前 10 个地区的 λ 及 β)。

由上可以看出参数大部分都是在迭代 2000 次后趋于稳定，故将前 2000 次作为预迭代剔除出去，得到的结果见图 3。

剔除前 2000 次最后得到的参数估计值(均值)及 95%置信区间见表 2。

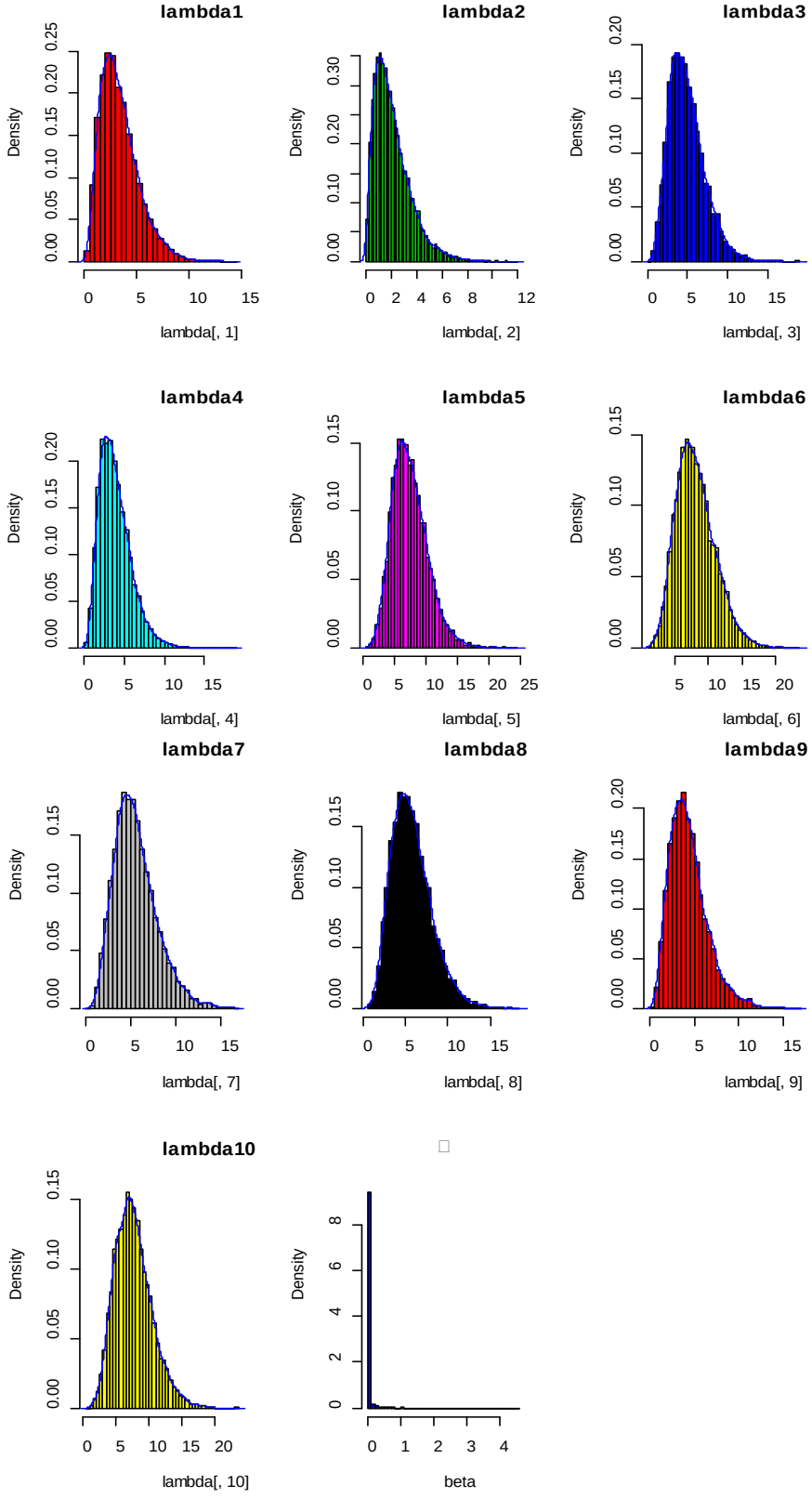

Figure 1. Some parameters posterior histogram and density curve
图 1. 部分参数后验直方图和密度曲线图

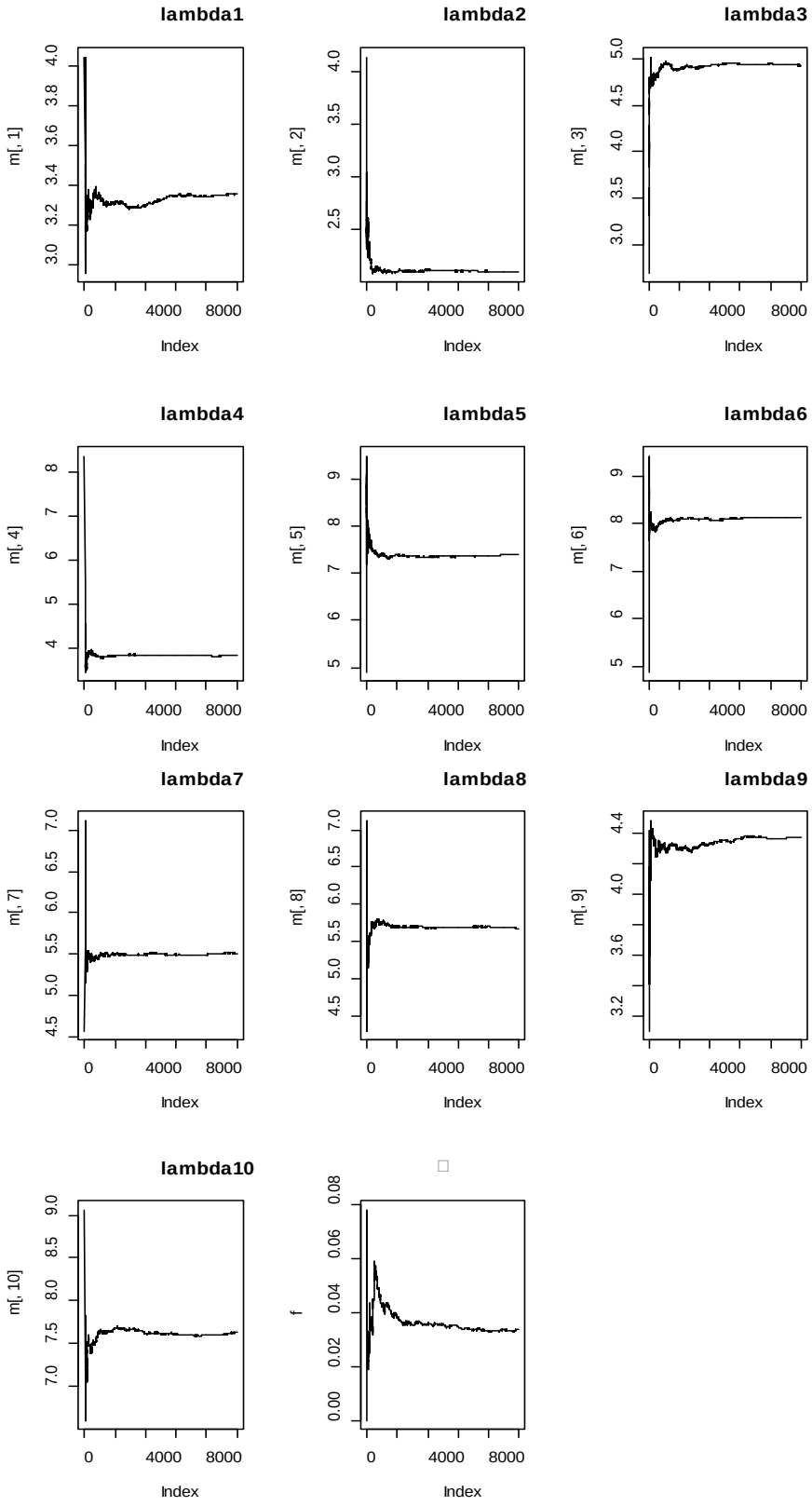

Figure 2. Iterate over 10,000 times the dynamic average of some parameters
图 2. 迭代 10,000 次的部分参数的动态均值

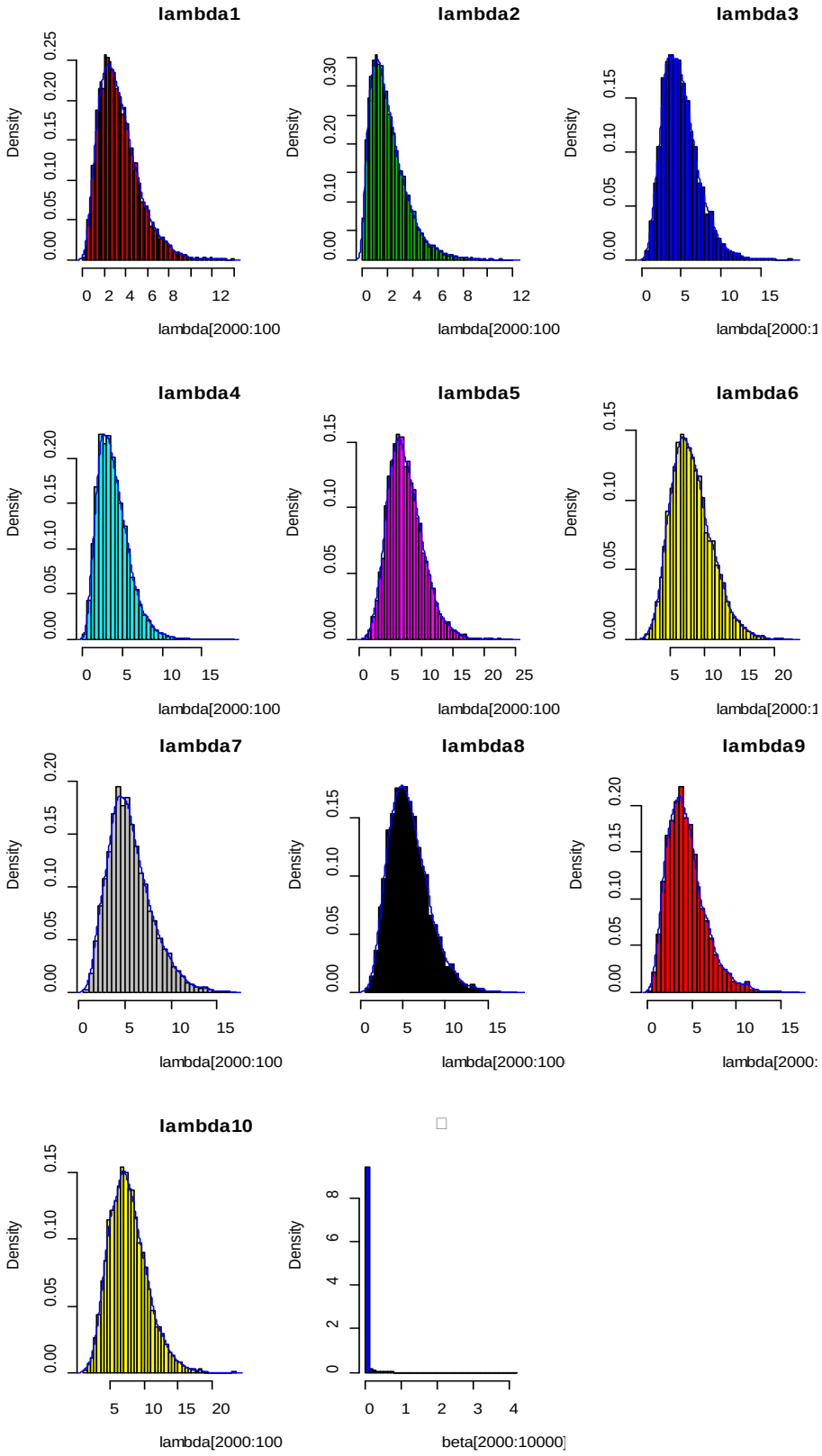

Figure 3. Remove the first 2000 posterior histograms and density plots
图 3. 剔除前 2000 次的后验直方图和密度曲线图

Table 2. Parameter confidence interval
表 2. 参数置信区间

序号	地区	参数	均值	置信下限	置信上限
1	北京	λ_1	3.366557	3.31342	3.360541
2	天津	λ_2	2.078503	2.028595	2.080018
3	河北	λ_3	4.918985	4.927619	4.967373
4	山西	λ_4	3.814735	3.827368	3.909014
5	内蒙古	λ_5	7.434227	7.414996	7.466077
6	辽宁	λ_6	8.087923	8.110538	8.18734
7	吉林	λ_7	5.565465	5.485664	5.569156
8	黑龙江	λ_8	5.697199	5.648096	5.697268
9	上海	λ_9	4.358082	4.346526	4.399555
10	江苏	λ_{10}	7.589888	7.510398	7.579255
11	浙江	λ_{11}	3.440375	3.361206	3.431547
12	安徽	λ_{12}	5.584512	5.536405	5.574634
13	福建	λ_{13}	5.596041	5.595939	5.651254
14	江西	λ_{14}	3.721809	3.716697	3.756724
15	山东	λ_{15}	11.721	11.67724	11.75821
16	河南	λ_{16}	5.028443	4.929083	5.017305
17	湖北	λ_{17}	4.857016	4.856308	4.91549
18	湖南	λ_{18}	10.1198	10.14167	10.30233
19	广东	λ_{19}	7.984993	7.971372	8.048027
20	广西	λ_{20}	1.342791	1.350906	1.380346
21	海南	λ_{21}	0.688424	0.654358	0.68352
22	重庆	λ_{22}	3.72425	3.698123	3.726004
23	四川	λ_{23}	6.787429	6.686947	6.772387
24	贵州	λ_{24}	0.941195	0.934009	0.958631
25	云南	λ_{25}	1.23728	1.221238	1.240721
26	西藏	λ_{26}	0.191603	0.183452	0.193099
27	陕西	λ_{27}	7.665271	7.666456	7.76611
28	甘肃	λ_{28}	4.348795	4.323805	4.351662
29	青海	λ_{29}	1.04032	1.042944	1.076705
30	宁夏	λ_{30}	3.2596528	3.2189546	3.2548818
31	新疆	λ_{31}	7.1295013	7.1438106	7.2328882
32		β	0.03262869	0.03287894	0.03697632

5. 结论

以上得到了 31 个地区的 λ、β 及其 95%置信区间。在此基础上可对 2012 年以后火灾发生次数进行统计推断。以北京地区为例，其火灾发生次数 x_i 概率为 $f(1)=\frac{(3.366557)^1}{1!}e^{-3.366557}\approx 0.116174$。计算北京地区 2013 发生 1 千次火灾的概率为 c 由于火灾发生次数数据的不可重复性，传统经典统计方法在样本量较

小的情形下很难得到具有说服力的结论，而贝叶斯方法充分利用了历史数据中所包含的信息，通过 Gibbs 抽样，可推断下一期的火灾发生次数的概率，为消防部门的工作安排提供理论指导，具有很强的实际意义。

参考文献

[1] Schaenman, P.S., Hall, J. and Schainblatt, A. (1977) Procedures for Improving the Measurement of Local Fire Protection Effectiveness.

[2] 杨立中, 江大白. 中国火灾与社会经济因素的关系[J]. 中国工程科学, 2003, 5(2): 62-67.

[3] 邓欧, 李亦秋, 冯仲科, 张冬有. 基于空间Logistic的黑龙江省林火风险模型与火险区划[J]. 农业工程学报, 2012, 28(8): 200-205.

[4] Bisquert, M., Caselles, E., Sánchez, J.M., *et al*. (2012) Application of Artificial Neural Networks and Logistic Regression to the Prediction of Forest Fire Danger in Galicia Using MODIS Data. *International Journal of Wildland Fire*, **21**, 1025-1029. https://doi.org/10.1071/wf11105

[5] Dlamini, W.M. (2011) Application of Bayesian Networks for Fire Risk Mapping Using GIS and Remote Sensing Data. *Geojournal*, **76**, 283-296. https://doi.org/10.1007/s10708-010-9362-x

[6] 王宁. 基于分层贝叶斯分析的城镇登记失业率估计方法[D]: [硕士学位论文]. 天津: 天津财经学, 2012

[7] 张颖, 傅强. GJR-CAViaR 模型的贝叶斯分位数回归——基于 Gibbs 抽样 MCMC 算法实现[J]. 中央财经大学学报, 2017(7): 87-95.

[8] 马跃渊, 徐勇勇. Gibbs 抽样算法及软件设计的初步研究[J]. 计算机应用与软件, 2005(2): 124-126.