HW5MuyangShi

Muyang Shi

2024-03-21

Note: the cpp source code to this document can be found on my Github, listed as `MCMC.cpp`, here.

Problem

Assume the model

$$X_j \sim \begin{cases} \text{Poisson}(\lambda_1), & j = 1, ..., \theta \\ \text{Poisson}(\lambda_2), & j = \theta + 1, ..., 112 \end{cases}$$

Assume $\lambda_i|\alpha \sim \text{Gamma}(3, \alpha)$ for $i = 1, 2$ where $\alpha \sim \text{Gamma}(10, 10)$ and assume θ follows a discrete uniform distribution over $\{1, ..., 111\}$. We would like to estimate the posterior distribution of the model parameters via a Gibbs sampler.

(a) Derive the conditional distributions necessary to carry out Gibbs sampling for the change-point model.

The joint likelihood is:

$$\begin{aligned} p(\boldsymbol{\Theta}|\boldsymbol{X}) &\propto p(\boldsymbol{X}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}) \\ &= p(\boldsymbol{X}|\lambda_1, \lambda_2, \theta) \cdot p(\lambda_1|\alpha)p(\lambda_2|\alpha) \cdot p(\alpha)p(\theta) \\ &= \prod_{j=1}^{j=\theta} \frac{\lambda_1^{X_j} e^{-\lambda_1}}{X_j!} \prod_{j=\theta+1}^{j=112} \frac{\lambda_2^{X_j} e^{-\lambda_2}}{X_j!} \cdot \frac{\alpha^3}{\Gamma(3)} \lambda_1^{3-1} e^{-\alpha\lambda_1} \frac{\alpha^3}{\Gamma(3)} \lambda_2^{3-1} e^{-\alpha\lambda_2} \cdot \frac{10^{10}}{\Gamma(10)} \alpha^{10-1} e^{-10\alpha} \frac{1}{111} \mathbf{1}(\theta \in \{1, ..., 111\}) \end{aligned}$$

And the full conditionals are:

For λ_1:

$$\begin{aligned} p(\lambda_1|\boldsymbol{X}, \lambda_2, \alpha, \theta) &\propto p(\boldsymbol{X}|\lambda_1, \lambda_2, \alpha, \theta)p(\lambda_1|\alpha) \\ &\propto \prod_{j=1}^{j=\theta} \frac{\lambda_1^{X_j} e^{-\lambda_1}}{X_j!} \lambda_1^{3-1} e^{-\alpha\lambda_1} \\ &\propto \lambda_1^{\sum_{j=1}^{\theta} X_j + 3 - 1} e^{-(\theta+\alpha)\lambda_1} \\ &\sim \text{Gamma}(3 + \sum_{j=1}^{\theta} X_j, \theta + \alpha) \end{aligned}$$

Similarly, for λ_2:

$$\begin{aligned} p(\lambda_2|\boldsymbol{X}, \lambda_1, \alpha, \theta) &\propto \lambda_2^{\sum_{j=\theta+1}^{112} X_j + 3 - 1} e^{-(112-\theta+\alpha)\lambda_2} \\ &\sim \text{Gamma}(3 + \sum_{j=\theta+1}^{112} X_j, 112 - \theta + \alpha) \end{aligned}$$

And for α:

$$
\begin{aligned}
p(\alpha|\boldsymbol{X}, \lambda_1, \lambda_2, \theta) &\propto \alpha^3 e^{-\alpha\lambda_1}\alpha^3 e^{-\alpha\lambda_2}\alpha^{10-1}e^{-10\alpha} \\
&= \alpha^{16-1}e^{-(10+\lambda_1+\lambda_2)\alpha} \\
&\sim \text{Gamma}(16, 10+\lambda_1+\lambda_2)
\end{aligned}
$$

Finally for θ:

$$
\begin{aligned}
p(\theta|\boldsymbol{X}, \lambda_1, \lambda_2, \alpha) &\propto \prod_{j=1}^{j=\theta} \frac{\lambda_1^{X_j}e^{-\lambda_1}}{X_j!} \prod_{j=\theta+1}^{j=112} \frac{\lambda_2^{X_j}e^{-\lambda_2}}{X_j!} \mathbf{1}(\theta \in \{1, ..., 111\}) \\
&\propto \lambda_1^{\sum_{j=1}^{\theta} X_j} e^{-\theta\lambda_1} \lambda_2^{\sum_{j=\theta+1}^{112} X_j} e^{-(112-\theta)\lambda_2} \mathbf{1}(\theta \in \{1, ..., 111\})
\end{aligned}
$$

Note that we can't see the distribution of θ through just eyeballing the equation. We'd be using a Metropolis approach – 1.) proposing θ from the discrete Uniform between 1 and 111, 2.) calculate the likelihood (equation above) ratio to decide whether we accept or reject. (Also note that this proposal is symmetric, so there's no "Hasting" ratio.)

(b) Implement the Gibbs sampler. Use a suite of convergence diagnostics to evaluate the convergence and mixing of your sampler.

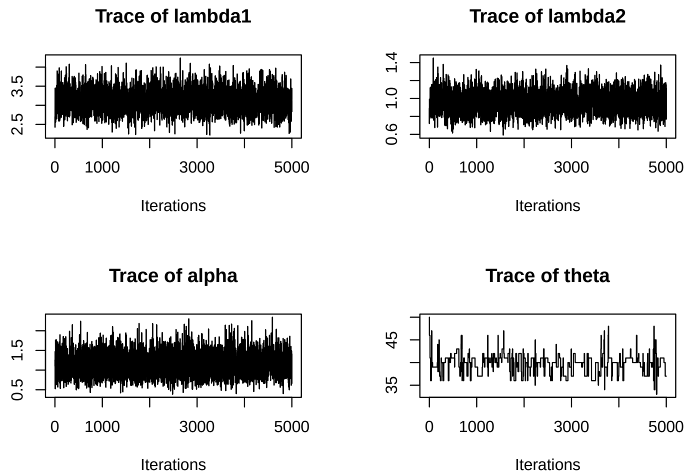
The Gibbs sampler is implemented in `RcppArmadillo` with the function name `MCMC`. We ran four chains (for the Gelman diagnostic), and for inference we will use the first chain.

Note that, as we plot below, where

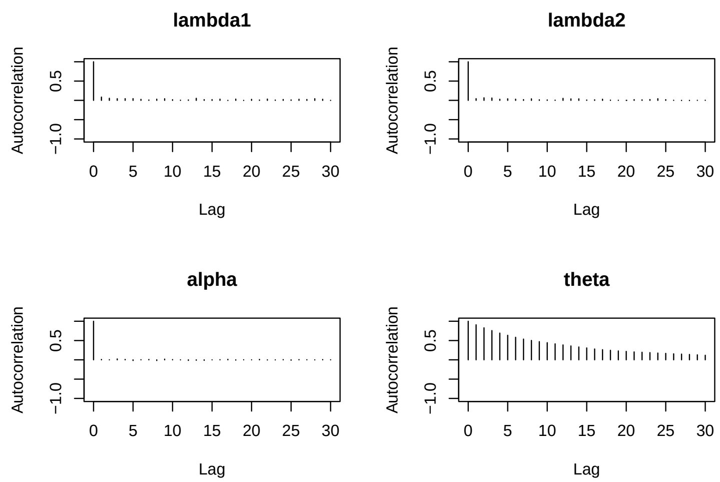
- the traceplot looks fine in that the chain seems to have converged and the mixing is good,
- the autocorrelation plot also seems fine in that we have decent effective sample size; θ is discrete so the traceplot for it actually doesn't look that bad
- and the Gelman diagnostic R is very close to 1 which is good (indicate that the chains have converged).

```r
coal <- read.table('coal.dat', skip = 1) # load data

# initial values
theta_0 <- 50
alpha_0 <- 1.1
lambda1_0 <- mean(coal[1:theta_0,2])
lambda2_0 <- mean(coal[theta_0:nrow(coal),2])

# run a couple chains using Rcpp
set.seed(2345)
chain1 <- MCMC(5000, coal$V2, lambda1_0, lambda2_0, alpha_0, theta_0)
chain2 <- MCMC(5000, coal$V2, lambda1_0, lambda2_0, alpha_0, theta_0)
chain3 <- MCMC(5000, coal$V2, lambda1_0, lambda2_0, alpha_0, theta_0)
chain4 <- MCMC(5000, coal$V2, lambda1_0, lambda2_0, alpha_0, theta_0)
```

```r
plot(chain1.mcmc, density = FALSE)
```




```r
autocorr.plot(chain1.mcmc)
```

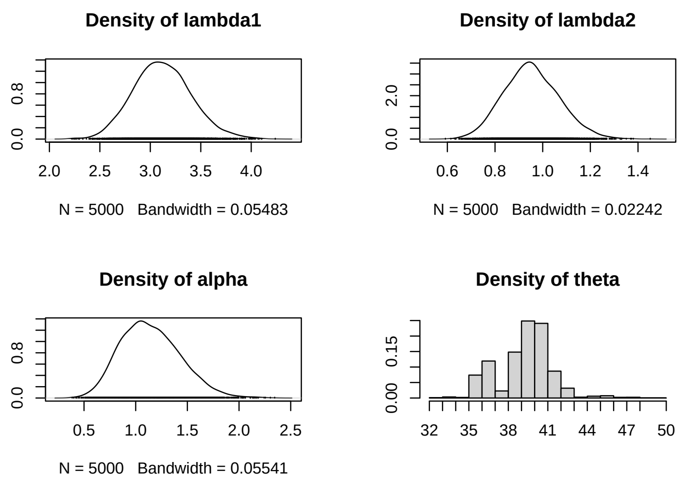



```r
gelman.diag(mcmc_list)
```

```
## Potential scale reduction factors:
##
##          Point est. Upper C.I.
## lambda1        1.00       1.00
## lambda2        1.00       1.00
## alpha          1.00       1.00
## theta          1.01       1.02
##
## Multivariate psrf
##
## 1.01
```

(c) Construct density histograms and a table of summary statistics for the approximate posterior distributions of θ, λ_1, and λ_2. Are symmetric HPD intervals appropriate for all of these parameters?

Here is the density histograms for the four parameters, $\lambda_1, \lambda_2, \alpha$ and θ:



Here is a table of summary statistics for the approximate posterior distributions of $\theta, \lambda_1, \lambda_2$:

```
##
## Iterations = 1:5000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 5000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean     SD Naive SE Time-series SE
## lambda1  3.120 0.2865 0.004051       0.005849
## lambda2  0.953 0.1162 0.001643       0.002437
## alpha    1.145 0.2871 0.004061       0.004061
## theta   39.746 2.0683 0.029250       0.137149
##
## 2. Quantiles for each variable:
##
##            2.5%     25%     50%    75%  97.5%
## lambda1  2.5928  2.9245  3.1098  3.305  3.726
## lambda2  0.7381  0.8735  0.9487  1.030  1.195
## alpha    0.6593  0.9341  1.1219  1.333  1.757
## theta   36.0000 39.0000 40.0000 41.000 43.000
```

The skewness in the distributions aren't "severe" – hence using symmetric HPD intervals are fine; however, one might argue that the distributions for λ_1, λ_2, as well as for θ are slightly right skewed, hence it might be better to use equal-tail intervals.

(d) Interpret the results in the context of the problem.

Based on the posterior distributions from the MCMC, we believe there is a 95% probability that:

- λ_1, average number of yearly coal-mining disasters during the first period, is anywhere between 2.59 to 3.73
- λ_2, average number of yearly coal-mining disasters during the second period, is anywhere between 0.74 to 1.20
- θ is anywhere between 36 to 43, i.e. the transition happened anywhere between the year of 1886 to 1893.

(e) Change the prior for λ_1 and λ_2 to a half-normal distribution $\propto N(0, \sigma^2)1_{[\lambda>0]}$ with σ^2 known. Derive the appropriate algorithm (MH or Gibbs) to carry out inference on the posterior distribution of λ_1 and λ_2.

We begin with the joint likelihood:

$$
\begin{aligned}
p(\boldsymbol{\Theta}|\boldsymbol{X}) &\propto p(\boldsymbol{X}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}) \\
&= p(\boldsymbol{X}|\lambda_1, \lambda_2, \theta) \cdot p(\lambda_1|\sigma^2)p(\lambda_2|\sigma^2) \cdot p(\theta) \\
&= \prod_{j=1}^{j=\theta} \frac{\lambda_1^{X_j} e^{-\lambda_1}}{X_j!} \prod_{j=\theta+1}^{j=112} \frac{\lambda_2^{X_j} e^{-\lambda_2}}{X_j!} \cdot \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \exp(-\frac{\lambda_1^2}{2\sigma^2}) \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \exp(-\frac{\lambda_2^2}{2\sigma^2}) \cdot \frac{1}{111} \mathbf{1}(\theta \in \{1, ..., 111\}) \\
&\propto \prod_{j=1}^{j=\theta} \lambda_1^{X_j} e^{-\lambda_1} \prod_{j=\theta+1}^{j=112} \lambda_2^{X_j} e^{-\lambda_2} \cdot \frac{1}{\sigma^2} \exp(-\frac{\lambda_1^2}{2\sigma^2}) \exp(-\frac{\lambda_2^2}{2\sigma^2}) \cdot \mathbf{1}(\theta \in \{1, ..., 111\})
\end{aligned}
$$

And the full conditionals,

For λ_1:

$$p(\lambda_1|\boldsymbol{X}, \lambda_2, \sigma, \theta) \propto \lambda_1^{\sum_{j=1}^{\theta} X_j} \exp(-\theta\lambda_1 - \frac{\lambda_1^2}{2\sigma^2})$$

For λ_2:

$$p(\lambda_2|\boldsymbol{X}, \lambda_1, \sigma, \theta) \propto \lambda_1^{\sum_{j=\theta+1}^{112} X_j} \exp(-(112-\theta)\lambda_1 - \frac{\lambda_2^2}{2\sigma^2})$$

And for θ:

$$p(\theta|\boldsymbol{X}, \lambda_1, \lambda_2, \sigma) \propto \lambda_1^{\sum_{j=1}^{\theta} X_j} e^{-\theta\lambda_1} \lambda_2^{\sum_{j=\theta+1}^{112} X_j} e^{-(112-\theta)\lambda_2} \mathbf{1}(\theta \in \{1, ..., 111\})$$

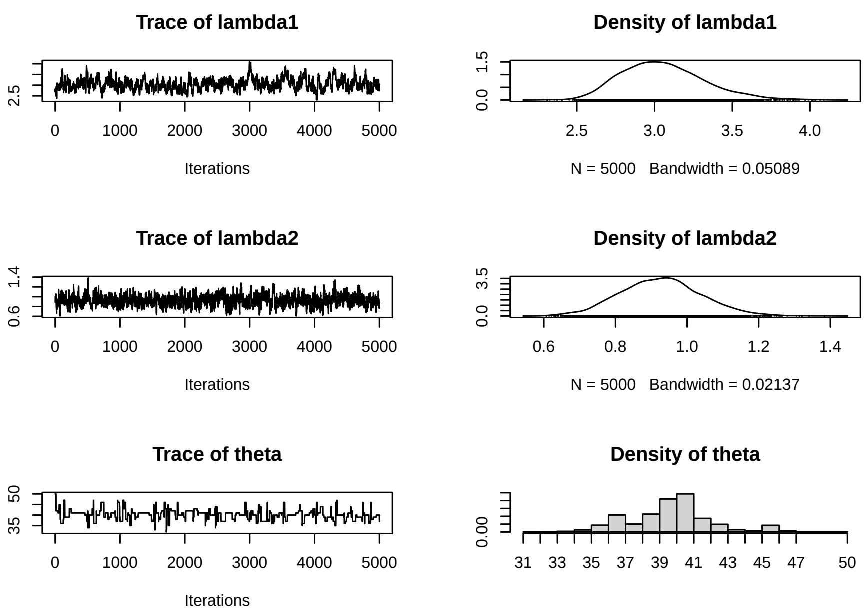
We will be using **random-walk metropolis** to update the parameters λ_1 and λ_2 (i.e. immediately reject if λ_1 and λ_2 falls out of support, so by doing it this way the hasting ratio will be 1).

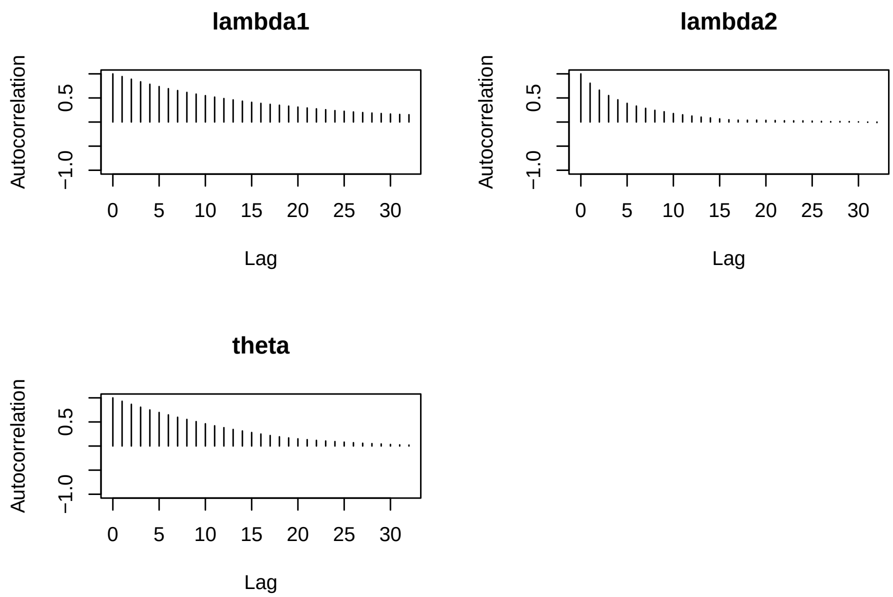
(f) Implement your MCMC algorithm from part (e). Try several values of σ^2 based on your understanding of the problem. (What is a reasonable variance for λ based on what you know about the problem?).

The MH sampler is implemented in `RcppArmadillo` with the function name `MCMC_halfnorm`. We ran four chains (for the Gelman diagnostic) across three values of $\sigma \in \{0.5, 1.5, 2.5\}$, and for inference we will use the first chain for each σ.

```r
sigmas <- c(0.5,1.5,2.5)
set.seed(2345)

chain1.1 <- MCMC_halfnorm(5000, coal$V2, lambda1_0, lambda2_0, sigmas[1], theta_0)
chain1.2 <- MCMC_halfnorm(5000, coal$V2, lambda1_0, lambda2_0, sigmas[1], theta_0)
chain1.3 <- MCMC_halfnorm(5000, coal$V2, lambda1_0, lambda2_0, sigmas[1], theta_0)
chain1.4 <- MCMC_halfnorm(5000, coal$V2, lambda1_0, lambda2_0, sigmas[1], theta_0)

chain2.1 <- MCMC_halfnorm(5000, coal$V2, lambda1_0, lambda2_0, sigmas[2], theta_0)
chain2.2 <- MCMC_halfnorm(5000, coal$V2, lambda1_0, lambda2_0, sigmas[2], theta_0)
chain2.3 <- MCMC_halfnorm(5000, coal$V2, lambda1_0, lambda2_0, sigmas[2], theta_0)
chain2.4 <- MCMC_halfnorm(5000, coal$V2, lambda1_0, lambda2_0, sigmas[2], theta_0)

chain3.1 <- MCMC_halfnorm(5000, coal$V2, lambda1_0, lambda2_0, sigmas[3], theta_0)
chain3.2 <- MCMC_halfnorm(5000, coal$V2, lambda1_0, lambda2_0, sigmas[3], theta_0)
chain3.3 <- MCMC_halfnorm(5000, coal$V2, lambda1_0, lambda2_0, sigmas[3], theta_0)
chain3.4 <- MCMC_halfnorm(5000, coal$V2, lambda1_0, lambda2_0, sigmas[3], theta_0)
```

We looked at the traceplot, autocorrelation plots, and the Gelman potential scale reduction factors, and decide that the model with $\sigma = 1.5$ yields the best result. Below are the traceplots, density plots, autocorrelation plots, and the Gelman PSRF for the model with $\sigma = 1.5$:






```
## Potential scale reduction factors:
##
##         Point est. Upper C.I.
## lambda1       1.00       1.01
## lambda2       1.00       1.01
## theta         1.01       1.02
##
## Multivariate psrf
##
## 1
```

(g) Provide a few key results to compare the inferences that you'd make based on the two models. Which model do you prefer and why?

Here is a summary of posterior for the halfnorm model with $\sigma = 1.5$:

```
##
## Iterations = 1:5000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 5000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##            Mean     SD Naive SE Time-series SE
## lambda1  3.0551 0.2657 0.003757       0.021709
## lambda2  0.9274 0.1130 0.001599       0.005055
## theta   40.1808 2.4859 0.035156       0.184296
##
## 2. Quantiles for each variable:
```

```
##
##             2.5%     25%     50%     75%  97.5%
## lambda1  2.6047  2.8659  3.0339  3.2192  3.634
## lambda2  0.7168  0.8501  0.9272  0.9984  1.157
## theta   36.0000 39.0000 40.0000 41.0000 46.000
```

Based on the posterior distributions from the MCMC, we believe there is a 95% probability that:

- λ_1, average number of yearly coal-mining disasters during the first period, is anywhere between 2.60 to 3.63
- λ_2, average number of yearly coal-mining disasters during the second period, is anywhere between 0.72 to 1.16
- θ is anywhere between 36 to 46, i.e. the transition happened anywhere between the year of 1886 to 1896.

Note that qualitatively, the results is similar to the Gamma-Poisson model – estimates for λ_1 being larger than λ_2 suggests that the reduction in yearly coal-mining disasters; the estimates for θ (change point) also being similar between the two models.

However, I would prefer the Gamma-Poisson model, because:

- Gibbs sampling does not involve the Metropolis-Hasting step, so mixing would not be an issue as we have the closed form posterior.

- Due to not having the Metropolis-Hasting update, the Gibbs sampler is also marginally faster than the MH.

(h) Run both models in `Nimble`. Compare the results and speed from your code and Nimble's. Briefly discuss.