Final STAT544 Spring2014

Yet Tien Nguyen

May 13, 2014

1 Introduction

In this project, we will analyze RNA-Seq data from RFI (Residual Feed Intake) project from ISU animal science research group. The data consists of a count table with ten thousand rows and 24 columns, and a RFI covariate. Each column is an experimental unit. Those 24 experimental units are in one of the two lines: half of them is in HRFI (High Residual Feed Intake) line, the other half is LRFI (Low Residual Feed Intake) line, denoted as line 1, and 2, correspondingly. The RFI value is the associated continuous covariate with each experimental unit. We want to know which genes are differentially expressed between two lines in the presence of the associated covariate RFI.

2 Model

Let y_{ijk} be the RNA expression level of gene $j \in \{1, \ldots, J\}$ from experimential unit $k \in \{1, \ldots, 12\}$ in group $i \in \{1, 2\}$. Let x_{ik} be the RFI value corresponding to the experimential unit k in group i. Suppose the count data for each gene have a Poisson distribution with the log link function of the mean as a linear combination of line effect and RFI covariate effects. In particular,

$$\begin{aligned}
& y_{ijk} \sim Poisson(\lambda_{ijk}) \\
& log(\lambda_{ijk}) = \alpha_j + (-1)^i \tau_j + \beta_j x_{ik} \\
& \alpha_j \sim N(0, 100^2) \\
& \tau_j \sim \pi_\tau \delta_0 + (1 - \pi_\tau) N(0, \sigma_\tau^2) \quad \text{equivalently, } \tau_j = (1 - \pi_{\tau_j}) * N(0, \sigma_\tau^2), \pi_{\tau_j} \sim Bern(\pi_\tau) \\
& \beta_j \sim \pi_\beta \delta_0 + (1 - \pi_\beta) N(0, \sigma_\beta^2) \quad \text{equivalently, } \beta_j = (1 - \pi_{\beta_j}) * N(0, \sigma_\beta^2), \pi_{\beta_j} \sim Bern(\pi_\beta) \qquad (1) \\
& \pi_\tau \sim Beta(8, 1) \\
& \pi_\beta \sim Beta(8, 1) \\
& \sigma_\tau \sim Unif(0, 100) \\
& \sigma_\beta \sim Unif(0, 100).
\end{aligned}$$

here δ_0 is the Dirac probability at 0. Note that the specification of $\pi_\tau \sim Beta(8, 1), \pi_\beta \sim Beta(8, 1)$ is based on Section 2.2 of the paper "An Exploration of Aspects of Bayesian Multiple Testing" by JG Scott, JO Berger - Journal of Statistical Planning and Inference, 2006

- Elsevier where we *believe* that the proportion of signals differentiating lines and covariate are small (around 1%-20%). Another option for the priors of parameters τ_j and β_j can be horseshoe prior

$$\begin{aligned}
&\tau_j \sim N(0, \sigma^2_{\tau_j}) \\
&\sigma_{\tau_j} \sim Ca^+(0, \sigma_\tau) \\
&\sigma_\tau \sim Ca^+(0, 1) \\
&\beta_j \sim N(0, \sigma^2_{\beta_j}) \\
&\sigma_{\beta_j} \sim Ca^+(0, \sigma_\beta) \\
&\sigma_\beta \sim Ca^+(0, 1).
\end{aligned} \tag{2}$$

At the end, the inference goal is to estimate the posterior probability of signals (differential expression) $P(\tau_j \text{ is signal}|y)$. If $P(\tau_j \text{ is signal } |y) > 0.5$, then the gene j is called *signal*, i.e., differentially expressed.

We propose two metrices to estimate the posterior probability of signals.

- Metric 1: We estimate the posterior probability of signal of gene j for point mass mixture prior model (1) by

  $$P(\tau_j \text{ is signal } |y) = P(|\tau_j| > 0|y) \approx \frac{1}{M}\sum_{i=1}^{M} I(\pi^{(i)}_{\tau_j} = 0) \tag{3}$$

  where $(\pi^{(i)}_{\tau_j}, i = 1, \ldots, M)$ is the posterior MCMC sample of π_{τ_j}.

  On the other hand, we estimate the posterior probability of signal of gene j for the horseshoe prior model (2) by

  $$P(\tau_j \text{ is signal } |y) \approx \frac{1}{M}\sum_{i=1}^{M} \big(1 - \frac{1}{1 + \sigma^{2(i)}_{\tau_j}}\big) \tag{4}$$

  where $(\sigma^{2(i)}_{\tau_j}, i = 1, \ldots, M)$ is the posterior MCMC sample of $\sigma^2_{\tau_j}$.

  This metric rule is a modified threshold rule motivated from section 3.4 Thresholding in the paper "The Horseshoe Estimator for Sparse Signals" by Carlos M. Carvalho and Nicholas G. Polson, Biometrika (2010), 97 ,2, pp. 465480. Note that the paper use the normal distribution setting which is different from the Poission distribution setting of this project.

- Metric 2: We estimate the posterior probability of signal of gene j for both models (1) and (2) by

  $$P(|\tau_j| > \varepsilon) \approx \frac{1}{M}\sum_{i=1}^{M} I(|\tau^{(i)}_j| > \varepsilon) \tag{5}$$

  for given threshold $\varepsilon > 0$.

We will use these two metrics to evaluate performance of detecting signals in simulated dataset. Based on the simulation results, we will pick up the better model to analyze RFI data. For the purpose of this project, we will analyze first 200 genes of the original RNA-Seq dataset.

3 Evaluation two models using simulation data

We simulate data to evaluate the performance of the two models in detecting true signals. For simplicity, we simulate a count table with dimension 100×24 where each row corresponding to one gene, first 12 columns corresponding to line 1, the next 12 colums corresponding to line 2. Also, we simulate 24 covariates $x_k \sim N(0,1)$ for $k = 1, \ldots, 24$. The $(j,k)-$ element of the count table is simulated from $Poisson(\lambda_{jk})$ distribution, where

$$\begin{aligned} \lambda_{jk} &= \alpha_j - \tau_j + \beta_j x_k, \quad \text{for} \quad k = 1, \ldots, 12 \\ \lambda_{jk} &= \alpha_j + \tau_j + \beta_j x_k, \quad \text{for} \quad k = 13, \ldots, 24 \\ \alpha_j &\sim N(3, 2^2) \quad \text{for} \quad j = 1, \ldots, 50 \\ \tau_j &= (1 - Bern(0.8)) * N(\mu_\tau, 0.25^2) \quad \mu_\tau \in \{0.5, 1, 2\} \\ \beta_j &= (1 - Bern(0.8)) * N(\mu_\beta, 0.25^2) \quad \mu_\beta \in \{0.5, 1, 2\}. \end{aligned} \tag{6}$$

In summary, we have total $3 \times 3 = 9$ different simulation scenarios. Each simulation scenario corresponds to one pairs of value

$$(\mu_\tau, \mu_\beta) \in \{(0.5, 0.5), (0.5, 1), (0.5, 2), (1, 0.5), (1, 1), (1, 2), (2, 0.5), (2, 1), (2, 2)\}.$$

For each simulation scenario, we simultaneously run a MCMC for the model (1) which is the point mass mixture prior model and the model (2) which is the horseshoe prior model. Then, we calculate the estimated posterior probability of signal of τ_j for each j based on two different evaluation metrices introduced in Section 2. In particular, with respect to metric 2, we use threshold $\varepsilon = 2\mu_\tau/3$ where μ_τ is the corresponding μ_τ of the simulation scenario. The ability to detect true signals of each model with specific metric is summarized in Table 1.

	(0.5,0.5)	(0.5,1)	(0.5,2)	(1,0.5)	(1,1)	(1,2)	(2,0.5)	(2,1)	(2,2)
(1)(3)	19(16)	20(20)	20(20)	26(26)	28(27)	24(24)	19(19)	16(16)	17(17)
(2)(4)	2(2)	0(0)	0(0)	6(6)	9(9)	7(7)	19(19)	15(15)	17(17)
(1)(5)	15(14)	14(14)	15(15)	23(23)	23(23)	20(20)	19(19)	15(15)	17(17)
(2)(3)	12(12)	12(12)	14(14)	21(21)	23(23)	20(20)	19(19)	15(15)	17(17)
true signals	20	22	22	26	27	25	19	16	17

Table 1: Results of detecting signals of two models (1) and (2) in 9 simulation scenarios.

The explanation of the Table 1 is as below. Each column is one simulation scenario corresponding to the value of (μ_τ, μ_β). Each row is the specification of the model and metric. Each element of the table indicates the number of detected signals and the number of correctly detected signals by the model and metric specified by its row name. For example, the element 19(16) at the first row "(1)(3)" and the first column "(0.5, 0.5)" means that: for the simulation scenario (6), model (1) with metric 3 detects 19 signals, 16 signals of which are correct signals, and there is total 20 true signals (the number at the last row in the first column).

Table 1 shows that the point mass mixture prior model (1) predicts very well the true signals between two lines, and indeed it predicts better than the horseshoe prior model (2) does. However, when the signals are large, the horseshoe prior model (2) performs well also. Furthermore, by checking the plots of posterior densitiies of τ_j, we also have the same conclusion. The Figures 1, 2, 3, 4 are the plots obtained from an analysis based on a simulation scenario (6) with $(\mu_\tau, \mu_\beta) = (1, 1)$ and 4 genes $j = 1, \ldots, 4$, the simulated parameters $\tau_3 = 1.49$, and $\tau_j = 0$ for all $j = 1, 2, 4$.

4 Analyze RFI data

As mentioned above, we only analyze 200 genes in the RNASeq - RFI data set for the simplicity. Based on the simulation results, we will use the point mass mixture prior Poission model (1) with two metrics (3) and (5) to see how they perform in analyzing this real dataset. Figures 5 and 6 show posterior samples and acf dianogstic plots of $\tau_1, \tau_2, \tau_3, \tau_4$,

As it turns out, the model (1) with metric (3) gives 184 signals which is a very high number and is not appropriate in reality that we *believe* the signals are sparse. On the other hand, the model (1) with metric (5) and $\varepsilon = 0.5$ gives 3 signals

```
##    line gene26 gene81 gene142
## 1     1      0      1       3
## 2     1      0      0      11
## 3     1      0      0      12
## 4     1      0      0       1
## 5     1      2      0      27
## 6     1      0      1       3
## 7     1      1      0       6
## 8     1      0      0       6
## 9     1      0      0      21
## 10    1      1      3      11
## 11    1      1      1      54
## 12    1      0      0       1
## 13    2      0      1       8
## 14    2      0      2       0
## 15    2      0      0     218
## 16    2      0      3     335
## 17    2      0     24      14
## 18    2     85      5      26
## 19    2      0      2      15
## 20    2      0      2      19
## 21    2      0      0       7
## 22    2      0      2       9
## 23    2      0      1      10
## 24    2      0      0      15
```

5 Conclusion

In this project, we develop a full Bayesian approach to analyze RNASeq data. We consider two models with different prior specifications. The first one is point mass mixture prior model (1). The second ons is the horsehoe prior model (2). We conduct a simulation study to evaluate the efficiency of these models in predicting the true differences between two lines. The simulation results suggest that the point mass mixture prior model (1) outperforms the horseshoe prior model (2) for the simulation scenarios (6). Then, we use the point mass mixture prior model (1) to analyze the RNASeq data and obtain a list of 3 signal genes (differentiallly expressed genes) in total 200 genes. However, the full RNASeq data consists of about 14000 genes, and it takes about 8 hours to run MCMC posterior samples on my computer, also the size of posterior samples for the parameters $(\tau_j, \alpha_j, \beta_j)$ is about 6GB, which is very big. One important point to note is that using metric (5), the number of signal genes depends on the threshold value ε. Therefore, more work needs to be done to account for the false discovery rate when changing the threshold value ε.

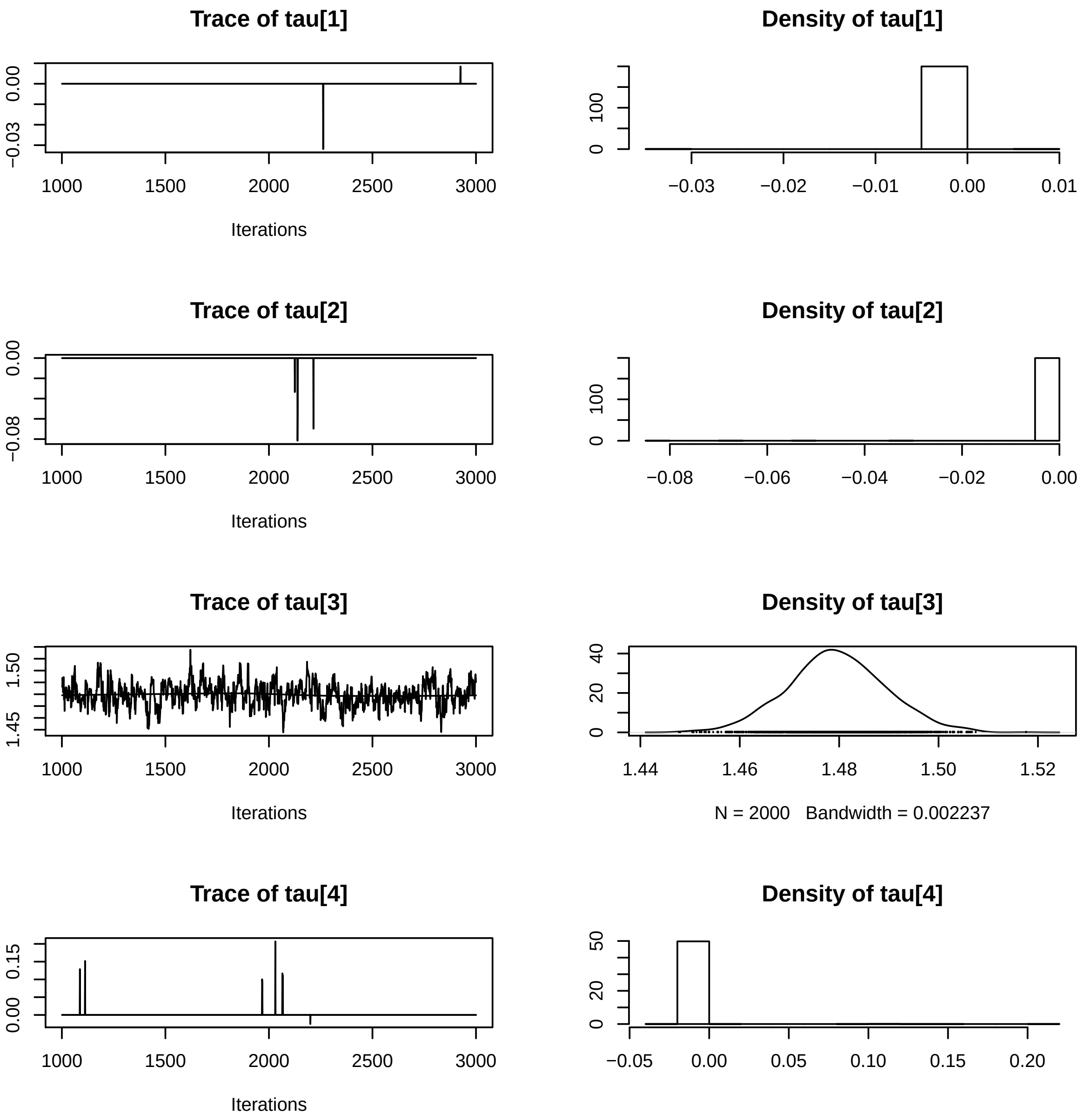

Figure 1: Posterior samples of $\tau_j, j = 1, \ldots, 4$ for the point mass mixture prior model (1).

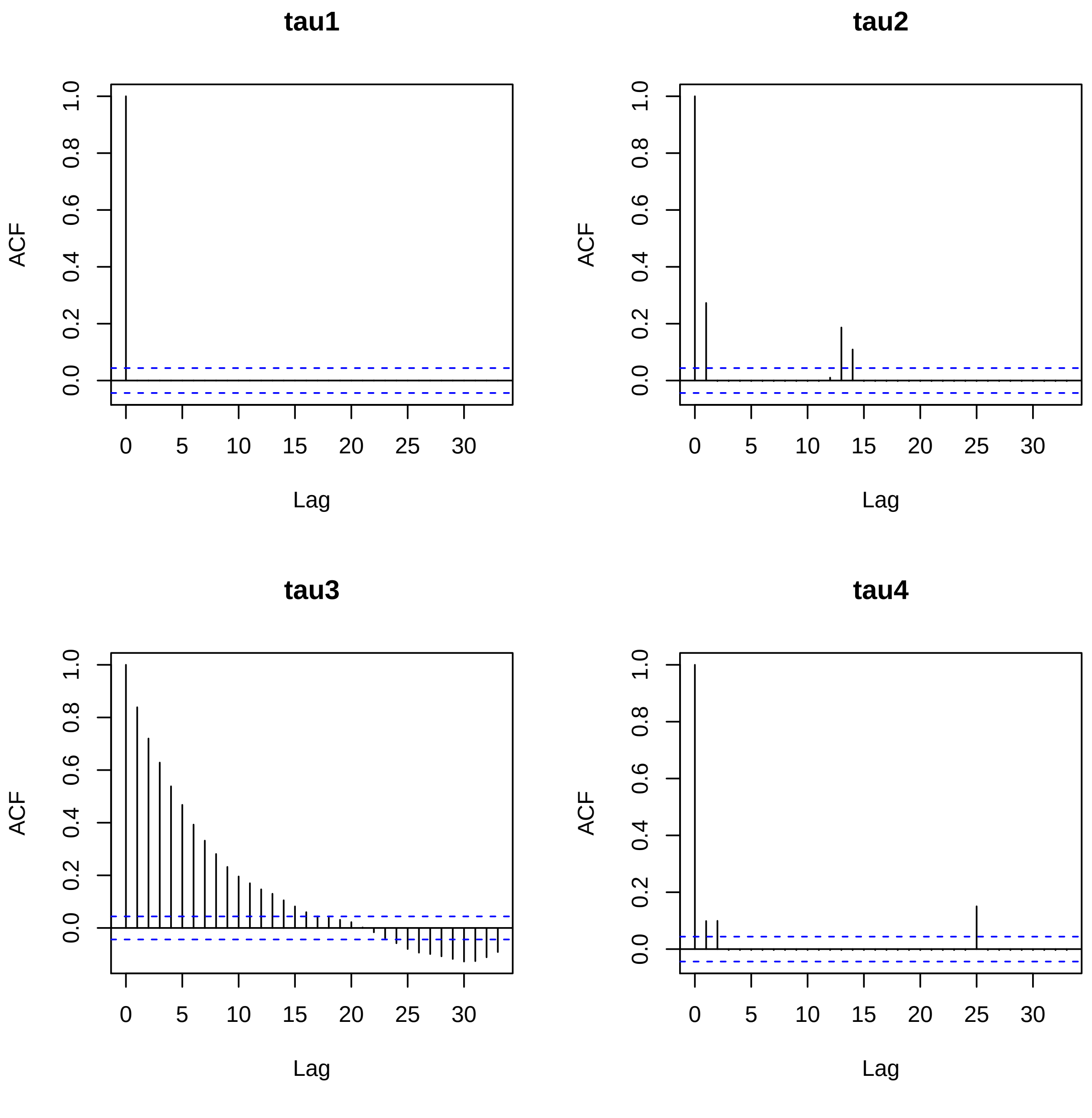

Figure 2: The dianogstic plots acf of the posterior MCMC samples of $\tau_j, j = 1, \ldots, 4$ under the point mass mixture prior model (1).

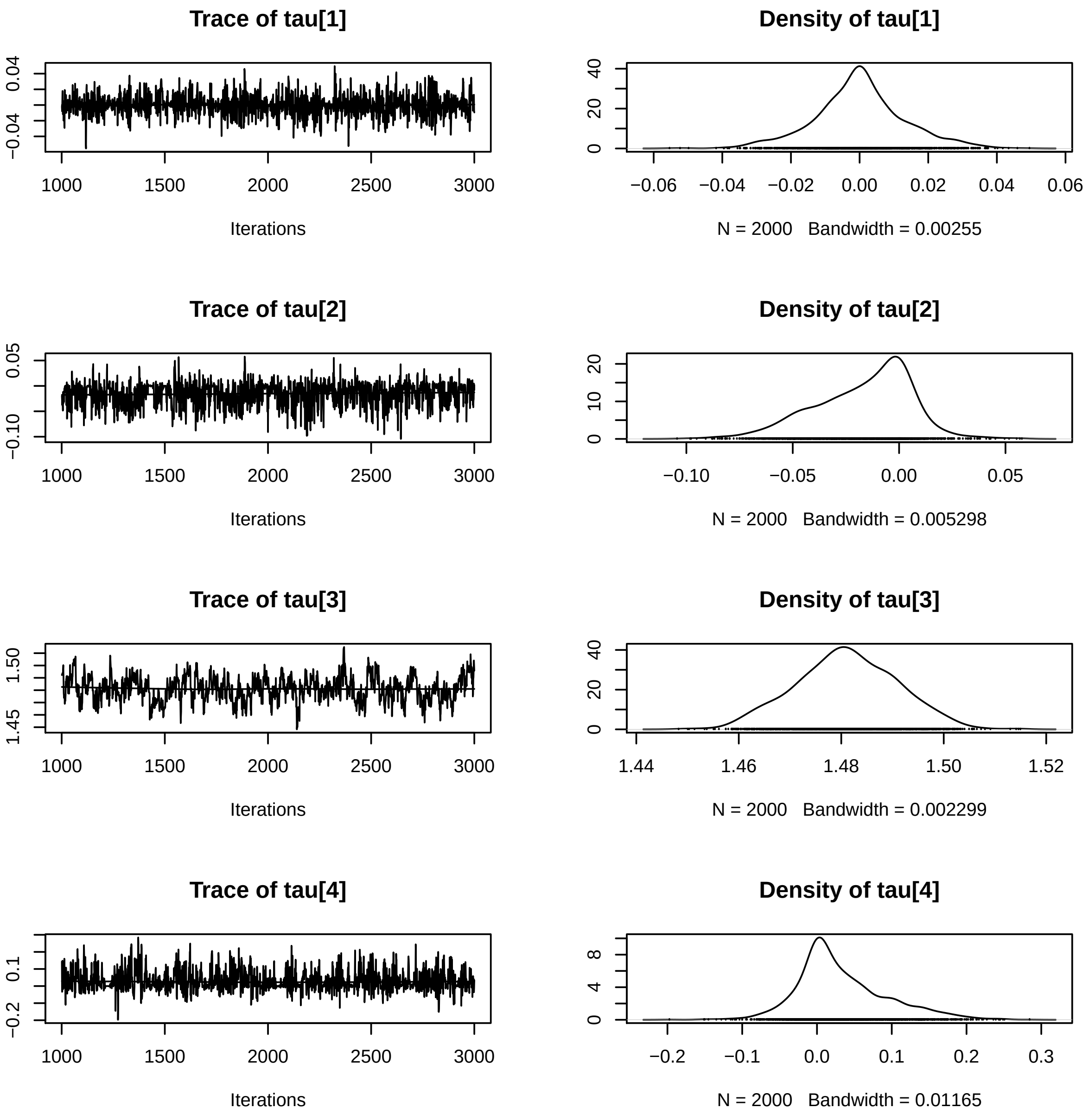

Figure 3: Posterior samples of $\tau_j, j = 1, \ldots, 4$ for the horseshoe prior model (2).

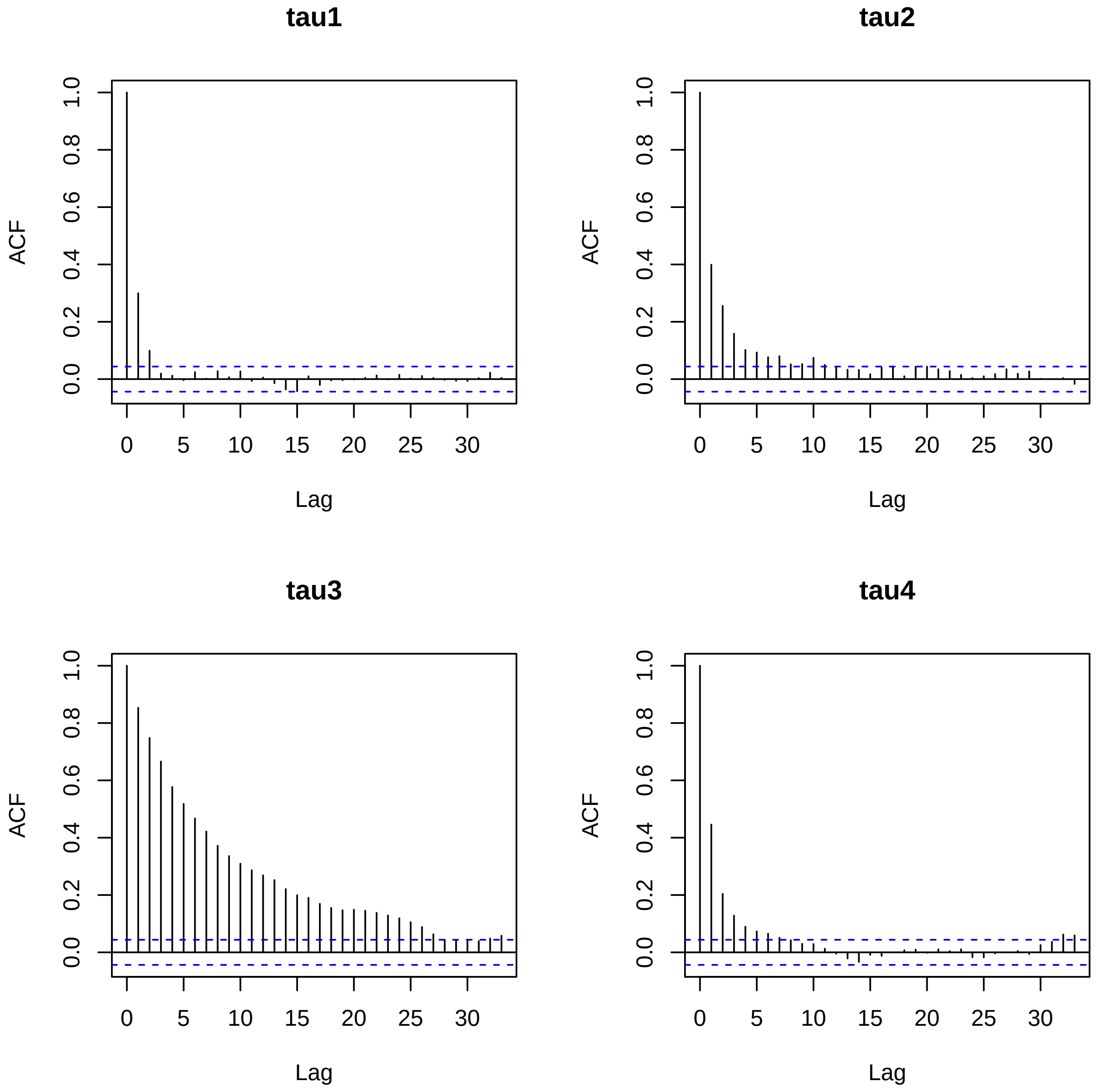

Figure 4: The dianogstic plots acf of the posterior MCMC samples of $\tau_j, j = 1, \ldots, 4$ under the horseshoe prior model (2).

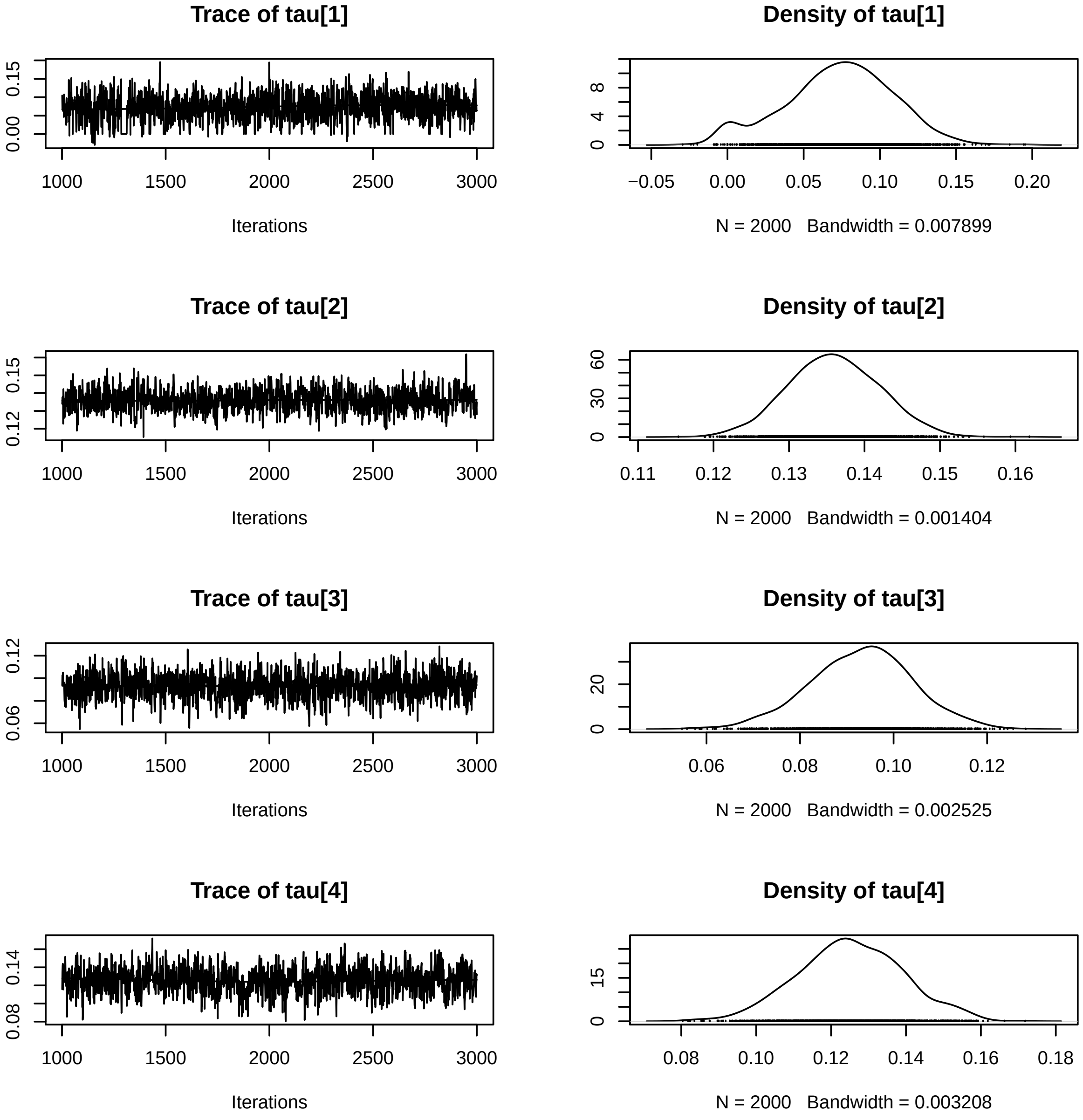

Figure 5: Posterior samples of $\tau_j, j = 1, \ldots, 4$ for the real data RNA-Seq under the point mass mixture prior model (1).

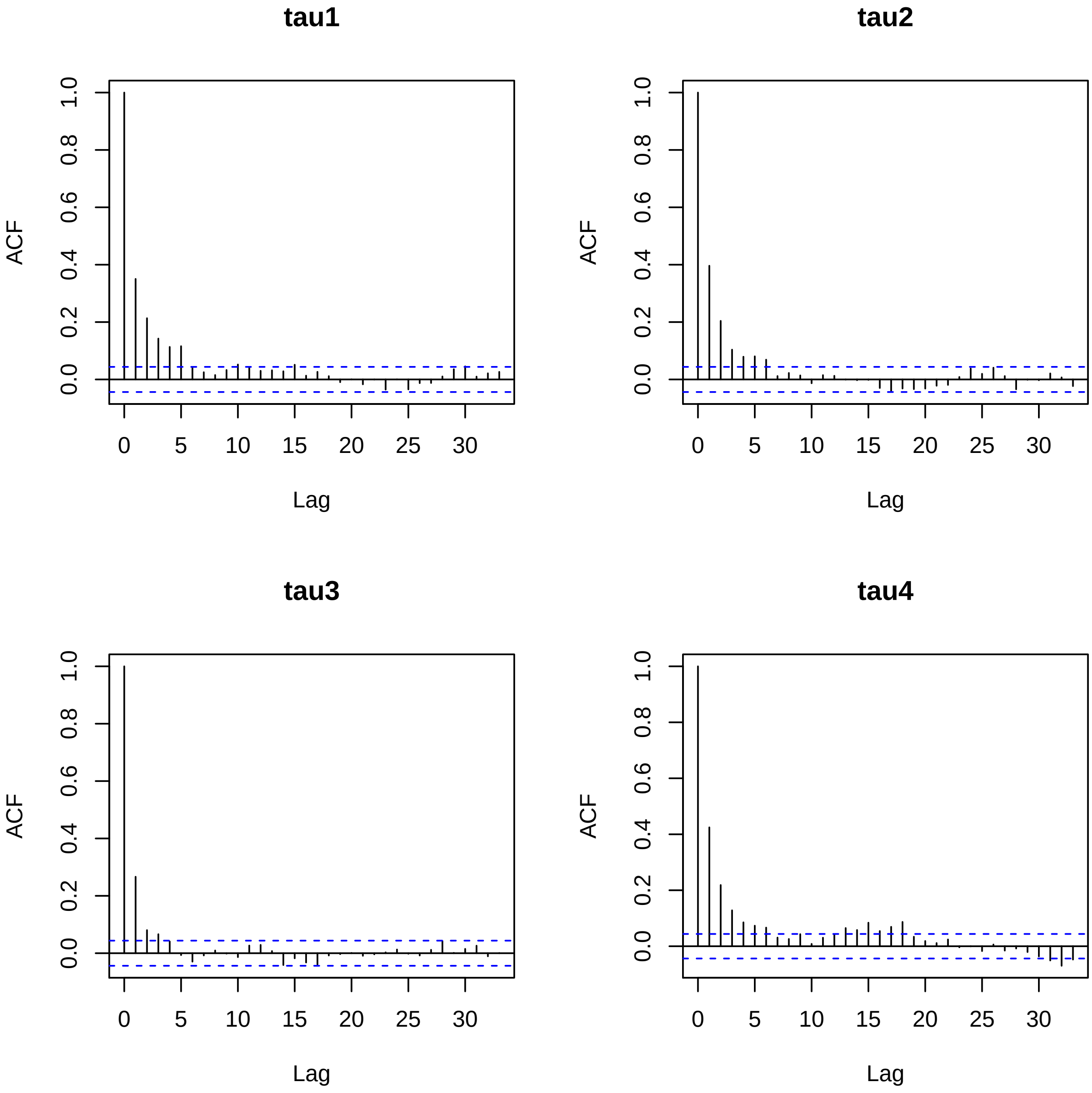

Figure 6: The dianogstic plots acf of the posterior MCMC samples of $\tau_j, j = 1, \ldots, 4$ for the real RNA-Seq data under the point mass mixture prior model (1).