Assignment

Computational Statistics 2015

Haibo Jin
haibo.nick.jin@gmail.com
014343698

University of Helsinki
Department of Computer Science
Master Student

Problem 1: Derive the maximum likelihood estimator of α and β. Identify the full conditionals $p(\alpha|\beta,y)$ and $p(\beta|\alpha,y)$. Do they correspond to familiar distributions?

Answer: The likelihood function is

$$f(y_1,y_2,...,y_n|\alpha,\beta)=\prod_{i=1}^{n}f(y_i|\alpha,\beta)$$

So the log-likelihood function is

$$\begin{aligned}\log f(y_1,y_2,...,y_n|\alpha,\beta)&=\sum_{i=1}^{n}\log f(y_i|\alpha,\beta)\\&=\sum_{i=1}^{n}\log(\frac{\beta^\alpha}{\Gamma(\alpha)}y_i^{\alpha-1}\exp\{-\beta y_i\})\\&=\sum_{i=1}^{n}\alpha\log\beta-\log\Gamma(\alpha)+(\alpha-1)\log y_i-\beta y_i\\&=n\alpha\log\beta-n\log\Gamma(\alpha)+\sum_{i=1}^{n}(\alpha-1)\log y_i-\beta y_i\end{aligned} \tag{1}$$

Take the derivative of the log-likelihood with respect to β, and set it to zero

$$\frac{d}{d\beta}\log f(y_1,y_2,...,y_n|\alpha,\beta)=\frac{n\alpha}{\beta}-\sum_{i=1}^{n}y_i=0$$

We get the maximum likelihood estimation of β:

$$\beta_{MLE}=\frac{n\alpha}{\sum_{i=1}^{n}y_i} \tag{2}$$

which depends on α. By substituting equation (2) to equation (1), we get

$$\log f(y_1,y_2,...,y_n|\alpha,\beta)=n\alpha\log\frac{n\alpha}{\sum_{i=1}^{n}y_i}-n\log\Gamma(\alpha)+(\alpha-1)\sum_{i=1}^{n}\log y_i-n\alpha \tag{3}$$

Take the derivative of equation (3) with respect to α, and set it to zero

$$\begin{aligned}\frac{d}{d\alpha}\log f(y_1,y_2,...,y_n|\alpha,\beta)&=n\log\frac{n\alpha}{\sum_{i=1}^{n}y_i}+n\alpha\frac{n}{\sum_{i=1}^{n}y_i}\frac{\sum_{i=1}^{n}y_i}{n\alpha}-n\psi(\alpha)+\sum_{i=1}^{n}\log y_i-n\\&=n\log\frac{n\alpha}{\sum_{i=1}^{n}y_i}+n-n\psi(\alpha)+\sum_{i=1}^{n}\log y_i-n\\&=n\log\frac{n\alpha}{\sum_{i=1}^{n}y_i}-n\psi(\alpha)+\sum_{i=1}^{n}\log y_i=0\end{aligned}$$

where $\psi(\alpha)$ is the digamma function of α. There is no analytical solution for this equation, we can use numerical method to get an approximate of α. Assuming we already get α_{MLE}, we then substitute it to equation (2), we get $\beta_{MLE}=\frac{n\alpha_{MLE}}{\sum_{i=1}^{n}y_i}$.

The full conditional $p(\alpha|\beta, y_1, ..., y_n)$ is propotional to the joint $p(\alpha, \beta, y_1, ..., y_n)$:

$$\begin{aligned} p(\alpha|\beta, y_1, ..., y_n) &\propto p(\alpha, \beta, y_1, ..., y_n) = p(\alpha)p(\beta)p(y_1, ..., y_n|\alpha, \beta) \\ &= \lambda \exp\{-\lambda\alpha\}\lambda \exp\{-\lambda\beta\}\prod_{i=1}^{n}\frac{\beta^\alpha}{\Gamma(\alpha)}y_i^{\alpha-1}\exp\{-\beta y_i\} \\ &\propto \exp\{-\lambda\alpha\}\frac{\beta^{n\alpha}}{\Gamma(\alpha)^n}(\prod_{i=1}^{n}y_i)^{\alpha-1} \end{aligned}$$

So $p(\alpha|\beta, y_1, ..., y_n)$ does not give a familiar distribution.
The full conditional $p(\beta|\alpha, y_1, ..., y_n)$ is also propotional to the joint $p(\alpha, \beta, y_1, ..., y_n)$:

$$\begin{aligned} p(\beta|\alpha, y_1, ..., y_n) &\propto p(\alpha, \beta, y_1, ..., y_n) = p(\alpha)p(\beta)p(y_1, ..., y_n|\alpha, \beta) \\ &= \lambda exp\{-\lambda\alpha\}\lambda exp\{-\lambda\beta\}\prod_{i=1}^{n}\frac{\beta^\alpha}{\Gamma(\alpha)}y_i^{\alpha-1}exp\{-\beta y_i\} \\ &\propto exp\{-\lambda\beta\}\prod_{i=1}^{n}\beta^\alpha exp\{-\beta y_i\} \\ &= \beta^{n\alpha}exp\{-(\lambda + \sum_{i=1}^{n}y_i)\beta\} \\ &= Gamma(n\alpha + 1, \lambda + \sum_{i=1}^{n}y_i) \end{aligned} \tag{4}$$

So $p(\beta|\alpha, y_1, ..., y_n)$ is actually a Gamma distribution.

Problem 2: Implement a random walk Metropolis-Hastings sampler for the constrained parameters $\alpha, \beta > 0$. Immediately reject proposed values if they are not both positive. Select the covariance matrix $a\Sigma$ of the proposal distribution so that the acceptance rate becomes reasonable (10-40%). Report posterior summary statistics and numerical standard errors, produce trace, autocorrelation and cumulative average plots for the parameters (α, β) and explain how you calibrated the tuning constant a.

Answer: The posterior distribution is

$$\begin{aligned} p(\alpha, \beta|y_1, ..., y_n) &\propto p(\alpha, \beta, y_1, ..., y_n) = p(\alpha)p(\beta)p(y_1, ..., y_n|\alpha, \beta) \\ &= \lambda \exp\{-\lambda\alpha\}\lambda \exp\{-\lambda\beta\}\prod_{i=1}^{n}\frac{\beta^\alpha}{\Gamma(\alpha)}y_i^{\alpha-1}\exp\{-\beta y_i\} \\ &\propto \exp\{-\lambda(\alpha+\beta)\}\frac{\beta^{n\alpha}}{\Gamma(\alpha)^n}(\prod_{i=1}^{n}y_i)^{\alpha-1}\exp\{-\beta\sum_{i=1}^{n}y_i\} \end{aligned} \tag{5}$$

So the log-posterior distribution is

$$\log p(\alpha, \beta|y_1, ..., y_n) \propto -\lambda(\alpha+\beta) + n\alpha\log\beta - n\log\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^{n}\log y_i - \beta\sum_{i=1}^{n}y_i \tag{6}$$

For the covariance matrix $a\Sigma$ of the proposal distribution, we usually set Σ to be an approximation of the posterior covariance matrix. I use normal approximation to estimate the posterior covariance matrix. Firstly, I use grid method to get the mode of the posterior: $\hat{\alpha} = 6.9$, $\hat{\beta} = 9.4$ (see details in the script *HaiboJin_code2_COMPSTAT2015.R*). Then I calculate the Hessian matrix of the log-posterior:

$$H(\alpha,\beta) = \begin{bmatrix} -n\psi_1(\alpha) & \frac{n}{\beta} \\ \frac{n}{\beta} & -\frac{n\alpha}{\beta^2} \end{bmatrix}$$

where $\psi_1(\alpha)$ is the trigamma function of α. By evaluating $H(\alpha,\beta)$ at the mode, we get $H(\hat{\alpha},\hat{\beta}) = \begin{bmatrix} -7.8 & 5.32 \\ 5.32 & -3.9 \end{bmatrix}$. So we get the approximation of the posterior covariance matrix as the inverse of the negative Hessian evaluating at the mode, $\Sigma = \begin{bmatrix} 1.82 & 2.48 \\ 2.48 & 3.63 \end{bmatrix}$.

The number of samples is 10000, and the initial value of α and β is $(3,3)$.

I use *acceptance rate* as the measure to calibrate a. In this case, it is not straightforward to calculate the average of the acceptance probabilities because the proposal can be negative which is not possible to evaluate its acceptance probability. So I simply set the *acceptance rate* to be the ratio of the number of accepted proposals to the total number of proposals. Since a reasonable acceptance rate is 10-40%, I tried a series of values of a with a broad range. Table 1 shows different values of tuning constant a and its acceptance rate. According to the table, a between 2 and 20 gives a reasonable acceptance rate. For the later calculations, I choose to fix a to 5 because of its moderate and reasonable acceptance rate.

Table 1: Different values of tuning constant a and its acceptance rate for problem 2.

a	0.1	1	2	5	10	20	50
acceptance rate	84.3%	56.3%	42.7%	25.5%	16.6%	8.7%	3.7%

Figure 1 gives the trace plots of α and β. We can see that both the chains start to mix well in the begining, so I just ignore the burn-in and retain all the samples.

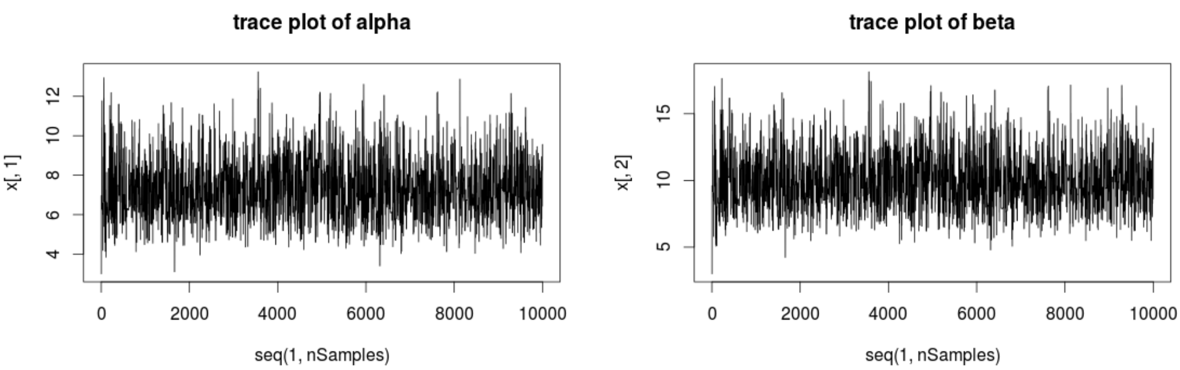

Figure 1: Trace plots of α and β for problem 2.

Table 2 gives the posterior summary statistics, with a being 5.

Table 2: Posterior summary statistics for problem 2.

	alpha	beta
Minimum	3.000	3.000
1st Quartile	6.300	8.574
Median	7.161	9.774
Mean	7.258	9.904
3rd Quartile	8.125	11.169
Maximum	12.112	16.403

The standard error of α and β are 0.015 and 0.021, respectively. Equation (7) gives the formula I use for calculating standard error

$$\mathrm{SE} = \frac{s}{\sqrt{n}} \tag{7}$$

where s is the standard deviation of samples and n is the number of samples.

Figure 2 shows the autocorrelation plots of α and β.

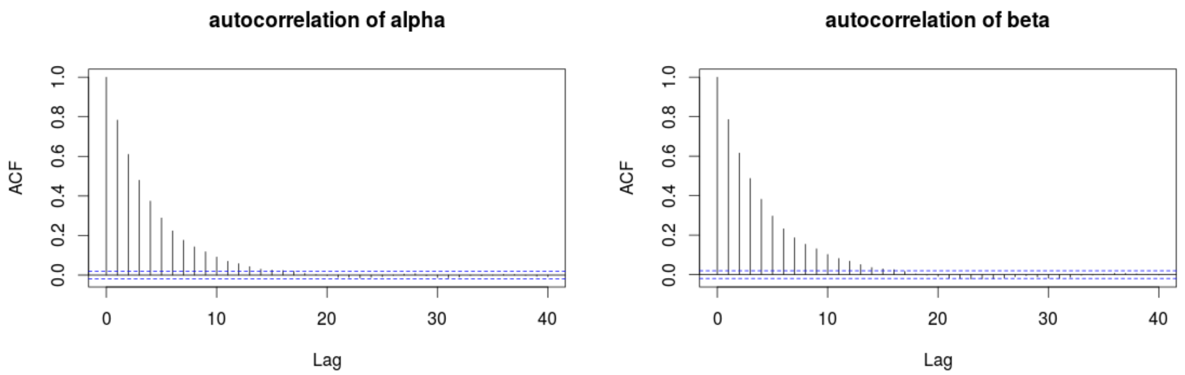

Figure 2: Autocorrelation plots of α and β for problem 2.

Figure 3 shows the cumulative average plots of α and β.

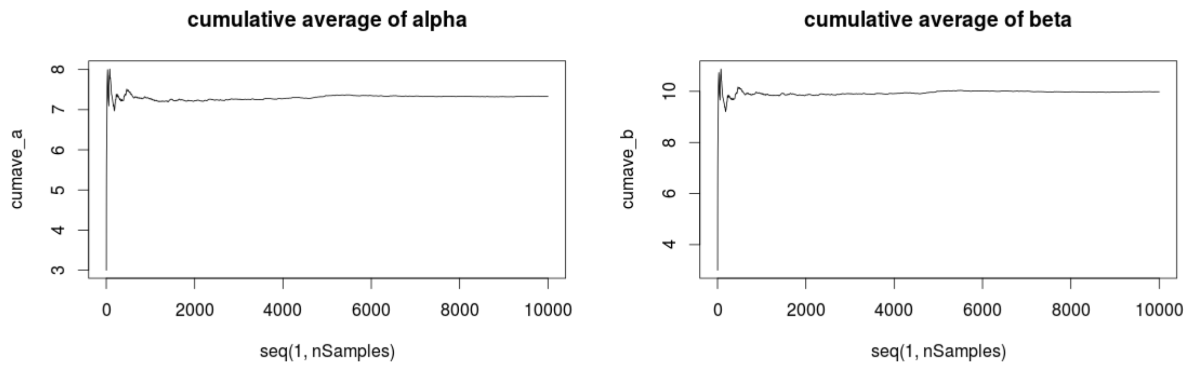

Figure 3: Cumulative average plots of α and β for problem 2.

Problem 3: Consider the following change-of-variables

$$\begin{bmatrix}\phi \\ \psi\end{bmatrix} = \mathrm{g}(\alpha,\beta) = \begin{bmatrix}\log\alpha \\ \log(\alpha/\beta)\end{bmatrix}$$

where $\mu = \alpha/\beta$ is the expected value of a random variable that follows a Gamma (α,β) distribution. Show that α,μ and therefore (ϕ,ψ) are orthogonal parameters by demonstrating that the expected Fisher information matrix $I(\alpha,\mu)$ is diagonal.

Implement a random walk Metropolis-Hastings sampler for the unconstrained parameters (ϕ,ψ). Select the covariance matrix $a\Sigma$ of the proposal distribution so that the acceptance rate becomes reasonable (10-40%). Report posterior summary statistics and numerical standard errors, produce trace, autocorrelation and cumulative average plots for the original parameters (α,β) and explain how you calibrated the tuning constant a.

Answer: The original parameters are (α,β) with Gamma distribution being its likelihood:

$$Gamma(y|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1}\exp\{-\beta y\}$$

We change variables with the following function

$$\begin{bmatrix}\alpha \\ \mu\end{bmatrix} = \mathrm{g}(\alpha,\beta) = \begin{bmatrix}\alpha \\ \alpha/\beta\end{bmatrix} \tag{8}$$

So the new likelihood function becomes

$$p(y|\alpha,\mu) = Gamma(y|\alpha,\frac{\alpha}{\mu}) = \frac{(\frac{\alpha}{\mu})^\alpha}{\Gamma(\alpha)} y^{\alpha-1}\exp\{-\frac{\alpha}{\mu}y\}$$

And the log-likelihood is

$$\log p(y|\alpha,\mu) = \alpha\log\frac{\alpha}{\mu} - \log\Gamma(\alpha) + (\alpha-1)\log y - \frac{\alpha}{\mu}y \tag{9}$$

If two parameters θ_i and θ_j are orthogonal, then the element of the ith row and jth column of the expected Fisher information matrix is zero. The element of the ith row and jth column of the expected Fisher information matrix is defined as follows

$$(I(\theta))_{i,j} = -E\left[\frac{d^2}{d\theta_i d\theta_j}\log f(Y|\theta)\right] \tag{10}$$

where the parameter θ is (α,μ) in our case. To get the expected Fisher information matrix, we first take derivative of equation (9) with respect to α

$$\frac{d}{d\alpha}\log p(y|\alpha,\mu) = \log\frac{\alpha}{\mu} + 1 - \psi(\alpha) + \log y - \frac{y}{\mu} \tag{11}$$

where $\psi(\alpha)$ is the digamma function of α. Then we can get the element of first row, first column of Fisher information matrix by taking derivative of equation (11) with respect to α

$$\frac{d^2}{d\alpha^2}\log p(y|\alpha,\mu) = \frac{1}{\alpha} - \psi_1(\alpha) \tag{12}$$

where $\psi_1(\alpha)$ is the trigamma function of α. To get the element of first row, second column as well as second row, first column, we take derivative of equation (11) with respect to μ

$$\frac{d^2}{d\alpha d\mu}\log p(y|\alpha,\mu) = -\frac{1}{\mu}+\frac{y}{\mu^2} \tag{13}$$

Finally, we take the second derivative of equation (9) with respect to μ to get the element of second row, second column

$$\frac{d^2}{d\mu^2}\log p(y|\alpha,\mu) = \frac{\alpha}{\mu^2}-\frac{2\alpha}{\mu^3}y \tag{14}$$

So the expected Fisher information matrix is

$$I(\alpha,\mu) = -E\begin{bmatrix}\frac{1}{\alpha}-\psi_1(\alpha) & -\frac{1}{\mu}+\frac{y}{\mu^2}\\ -\frac{1}{\mu}+\frac{y}{\mu^2} & \frac{\alpha}{\mu^2}-\frac{2\alpha}{\mu^3}y\end{bmatrix} \tag{15}$$

Notice that $E[y]=\mu$, so we simplify the matrix to the following

$$I(\alpha,\mu) = -\begin{bmatrix}\frac{1}{\alpha}-\psi_1(\alpha) & 0\\ 0 & -\frac{\alpha}{\mu^2}\end{bmatrix} = \begin{bmatrix}-\frac{1}{\alpha}+\psi_1(\alpha) & 0\\ 0 & \frac{\alpha}{\mu^2}\end{bmatrix} \tag{16}$$

Since the elements on the non-diagonal are zero, the parameters (α,μ) and therefore (ϕ,ψ) are orthogonal.

Now we make change-of-variables as follows

$$\begin{bmatrix}\phi\\ \psi\end{bmatrix} = \mathrm{g}(\alpha,\beta) = \begin{bmatrix}\log\alpha\\ \log(\alpha/\beta)\end{bmatrix} \tag{17}$$

So the inverse function is

$$\begin{bmatrix}\alpha\\ \beta\end{bmatrix} = \mathrm{h}(\phi,\psi) = \begin{bmatrix}e^{\phi}\\ e^{\phi-\psi}\end{bmatrix} \tag{18}$$

The Jecobian determinant of function h is

$$J_h = \det\begin{bmatrix}e^{\phi} & 0\\ e^{\phi-\psi} & -e^{\phi-\psi}\end{bmatrix} = -e^{2\phi-\psi}$$

Then the posterior of (ϕ,ψ) is

$$p(\phi,\psi|y_1,...,y_n) = p(\alpha,\beta|y_1,...,y_n)|J_h|$$

We can just write down the log-posterior:

$$\begin{aligned}\log p(\phi,\psi|y_1,...,y_n) &= \log p(\alpha,\beta|y_1,...,y_n)+\log|J_h|\\ &= \log p(\alpha,\frac{\alpha}{\mu}|y_1,...,y_n)+\log|J_h|\\ &= -\lambda(e^{\phi}+e^{\phi-\psi})+ne^{\phi}(\phi-\psi)-n\log\Gamma(e^{\phi})\\ &+(e^{\phi}-1)\sum_{i=1}^{n}\log y_i - e^{\phi-\psi}\sum_{i=1}^{n}y_i+2\phi-\psi\end{aligned} \tag{19}$$

After change-of-variables, the posterior seems a little complicated, which requires more work to get a normal approximation. Luckily, we have demonstrated that ϕ and ψ are orthogonal, so the posterior covariance matrix Σ is a diagonal matrix. I simply assume $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and it forms the covariance matrix of the proposal $a\Sigma$, where a is the tuning constant.

The number of samples is again 10000, and the initial value of ϕ and ψ is $(1,1)$.

I use *acceptance rate* as the measure to calibrate a. In order to be consistent to problem 2, I also define the *acceptance rate* here to be the ratio of the number of accepted proposals to the total number of proposals. Since a reasonable acceptance rate is 10-40%, I tried a series of values of a with a broad range. Table 3 shows different values of tuning constant a and its acceptance rate. According to the table, a between 0.02 and 0.2 gives a reasonable acceptance rate. For the later calculations, I choose to fix a to 0.1.

Table 3: Different values of tuning constant a and its acceptance rate for problem 3.

a	0.001	0.01	0.02	0.1	0.2	0.5
acceptance rate	79.5%	46.6%	34.9%	13.5%	8.3%	3.7%

Figure 4 gives the trace plots of α and β. We can see that both the chains start to mix well in the begining, so I just ignore the burn-in and retain all the samples.

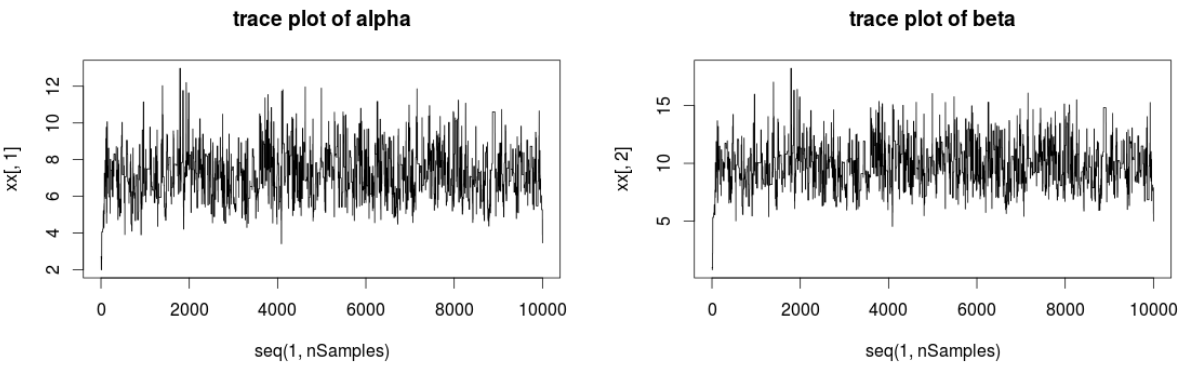

Figure 4: Trace plots of α and β for problem 3.

Table 4 gives the posterior summary statistics, with a being 0.1.

Table 4: Posterior summary statistics for problem 3.

	alpha	beta
Minimum	1.999	0.796
1st Quartile	6.226	8.475
Median	7.154	9.756
Mean	7.254	9.893
3rd Quartile	8.161	11.110
Maximum	12.975	18.206

Through equation (7), I get the standard error of α and β are 0.014 and 0.020, respectively.

Figure 5 shows the autocorrelation plots of α and β.

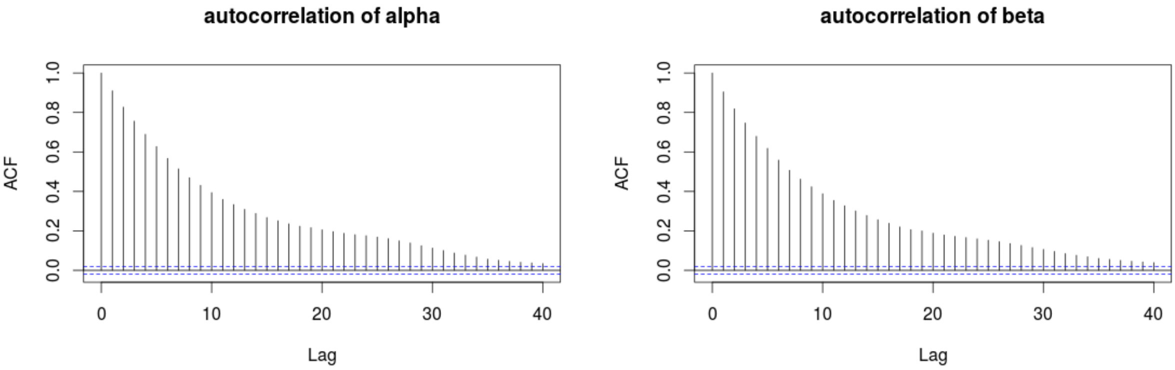

Figure 5: Autocorrelation plots of α and β for problem 3.

Figure 6 shows the cumulative average plots of α and β.

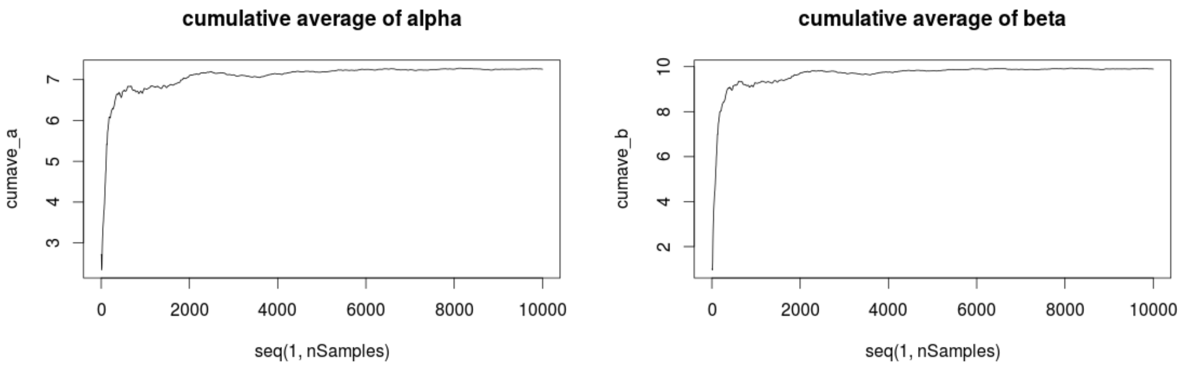

Figure 6: Cumulative average plots of α and β for problem 3.

Problem 4: Implement a hybrid sampler by taking advantage of conditional conjugacy. Update $\phi = g(\alpha) = \log\alpha$ with a random walk Metropolis-Hasting step and β by drawing from its full conditional distribution conditionally on the proposed value of α. Accept or reject the proposed pair jointly by using the ordinary Metropolis-Hastings acceptance rule. Tune the proposal distribution so that the acceptance rate becomes reasonable (10-40%). Report posterior summary statistics and numerical standard errors, produce trace, autocorrelation and cumulative average plots for the parameters (α, β) and explain how you calibrated the tuning constant.

Answer: For the previous two problems, the proposal distributions are symmetric, thus the MH ratio is simply $r = \frac{\pi(\theta')}{\pi(\theta)}$ where π denotes the target posterior. In this problem, the proposal is not symmetric, so the MH ratio is

$$r = \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)} \tag{20}$$

where q denotes the proposal distribution. In our case, the specific proposal distribution $q(\theta'|\theta)$ is

$$q(\alpha',\beta'|\alpha,\beta) = N(\log\alpha'|\log\alpha,\sigma^2)Gamma(\beta'|n\alpha'+1,\lambda+\sum_{i=1}^{n} y_i) \tag{21}$$

where N denotes normal distribution, $\log\alpha'$ and $\log\alpha$ is just ϕ' and ϕ, and the conditional Gamma is from equation (4).

Similarly, we can get $q(\theta|\theta')$

$$q(\alpha,\beta|\alpha',\beta') = N(\log\alpha|\log\alpha',\sigma^2)Gamma(\beta|n\alpha+1,\lambda+\sum_{i=1}^{n} y_i) \tag{22}$$

Because $N(\log\alpha'|\log\alpha,\sigma^2) = N(\log\alpha|\log\alpha',\sigma^2)$, so we can cancel them and simplify the MH ratio to the following equation

$$r = \frac{\pi(\theta')Gamma(\beta|n\alpha+1,\lambda+\sum_{i=1}^{n} y_i)}{\pi(\theta)Gamma(\beta'|n\alpha'+1,\lambda+\sum_{i=1}^{n} y_i)} \tag{23}$$

where the posterior π is still the same as equation (5) in problem 2.

In this problem, the random walk is on just one dimension, so its variance $a\sigma^2$ becomes a scalar. So I simply set $\sigma^2 = 1$.

Again, the number of samples is 10000, and the initial value of ϕ and ψ is $(1,1)$.

Same to the previous, I use *acceptance rate* as the measure to calibrate a, and it is defined as the ratio of the number of accepted proposals to the total number of proposals. Since a reasonable acceptance rate is 10-40%, I tried a series of values of a with a broad range. Table 5 shows different values of tuning constant a and its acceptance rate. According to the table, a between 0.5 and 2 gives a reasonable acceptance rate. For the later calculations, I choose to fix a to 1.

Table 5: Different values of tuning constant a and its acceptance rate for problem 4.

a	0.1	0.5	1	2	5
acceptance rate	84.1%	42.5%	23.5%	12.4%	4.8%

Figure 7 gives the trace plots of α and β. We can see that both the chains start to mix well in the begining, so I just ignore the burn-in and retain all the samples.

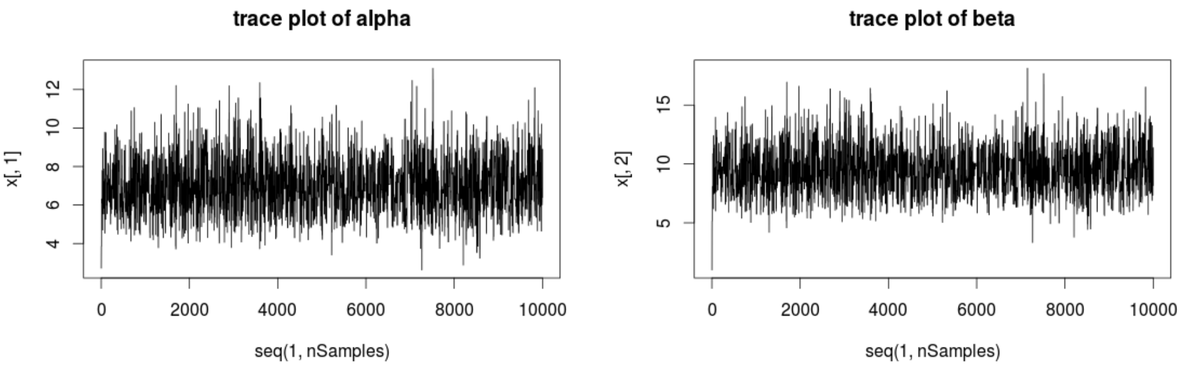

Figure 7: Trace plots of α and β for problem 4.

Table 6 gives the posterior summary statistics, with a being 1.

Table 6: Posterior summary statistics for problem 4.

	alpha	beta
Minimum	2.629	1.000
1st Quartile	6.027	8.194
Median	6.931	9.412
Mean	7.014	9.541
3rd Quartile	7.932	10.790
Maximum	13.105	18.130

Through equation (7), I get the standard error of α and β are 0.014 and 0.020, respectively. Figure 8 shows the autocorrelation plots of α and β.

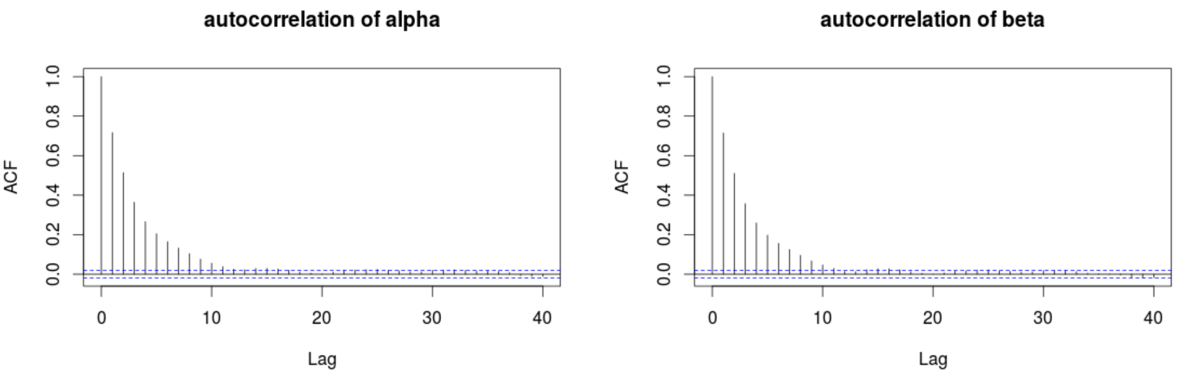

Figure 8: Autocorrelation plots of α and β for problem 4.

Figure 9 shows the cumulative average plots of α and β.

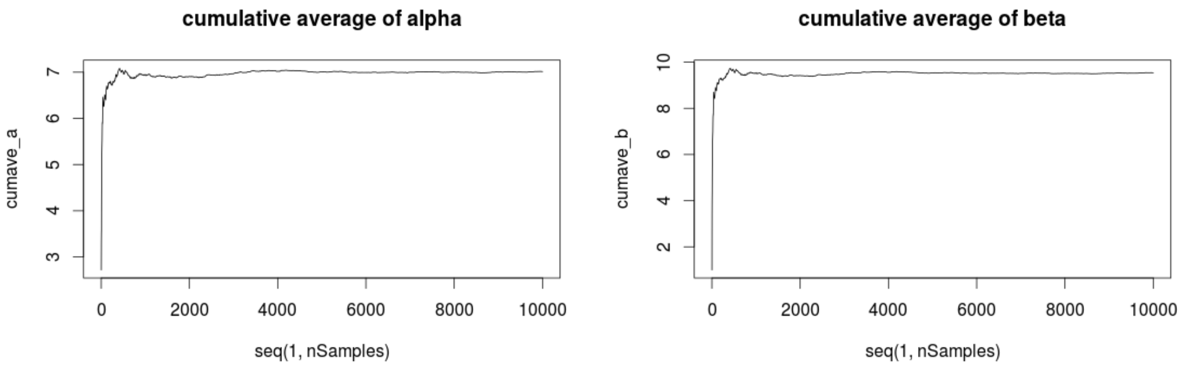

Figure 9: Cumulative average plots of α and β for problem 4.