Report: Statistical Modeling of Professors' Annual Salaries

Ziye Lin

Jul 31, 2023

Introduction

We conduct a statistical analysis to model professors' annual salaries. Specifically, we explore both the frequentist and Bayesian approach. The analysis begins with deriving analytical results, including the Fisher information matrix, to assess parameter variability under the frequentist approach. We then implement maximum likelihood estimation and construct confidence intervals. In the Bayesian framework, we specify prior distributions, derive posterior distributions, and use numerical methods in R to perform parameter estimation and uncertainty quantification. Throughout the report, we integrate theoretical derivations with computational analysis, complemented by visualizations to interpret results and evaluate model fit.

Dataset

Disclosed by University of Waterloo: https://uwaterloo.ca/about/accountability/salary-disclosure-2022

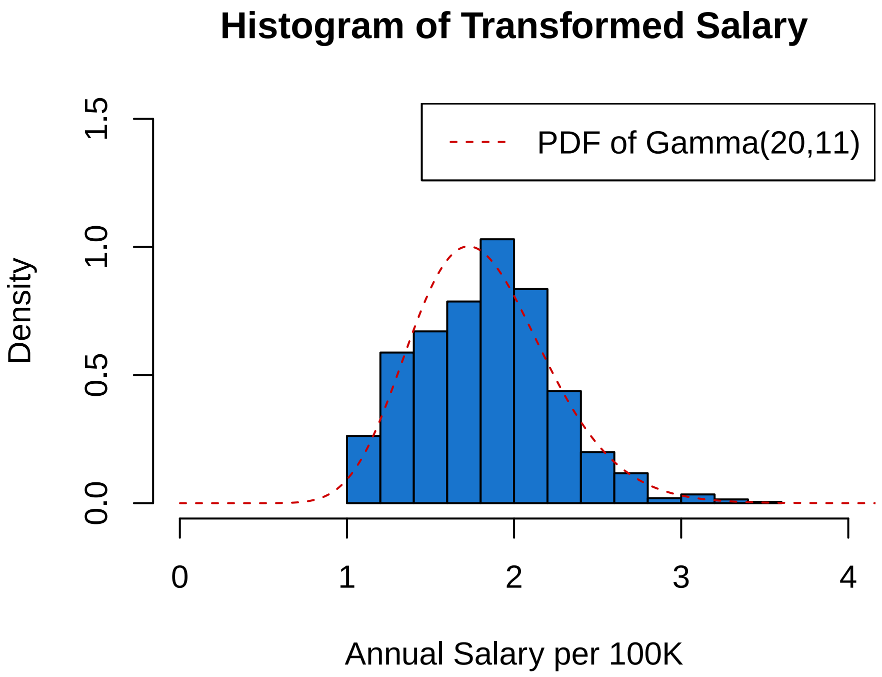
The data set contains the annual salary in excess of $100,000 and the taxable benefit of the academic staff at the University of Waterloo in 2022. We have removed the surname and the given name of all the members when importing the data into R since these are not relevant to this report. This data set can be treated as a sample to estimate the distribution of annual salaries of university professors in 2022 in Ontario, Canada.

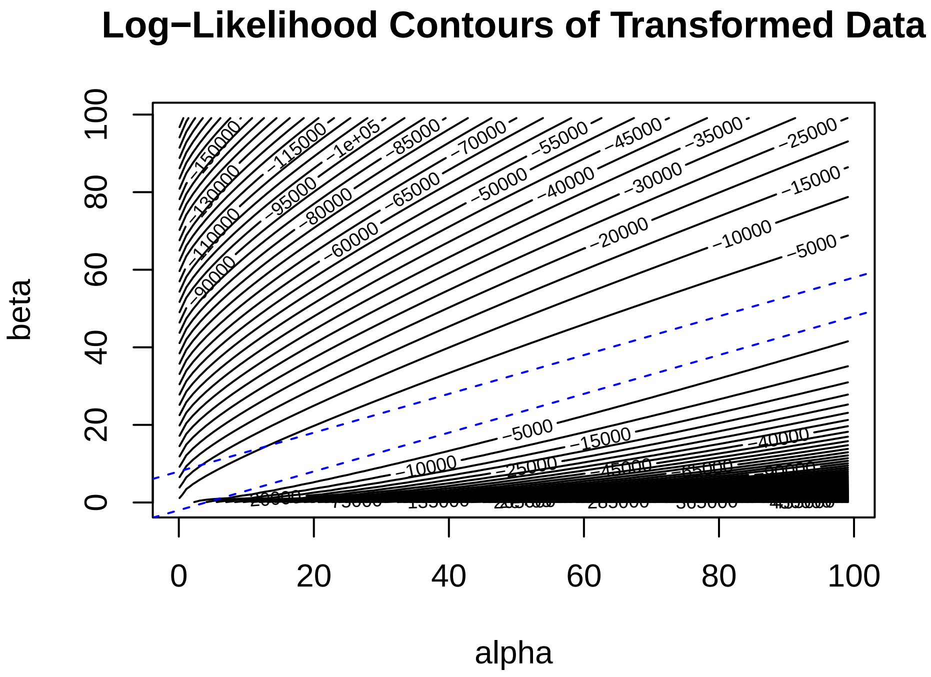
Note that professors here refer to those with position titles "assistant professor", "associate professor", and "professor". "Lecturer" is not included in the analysis.

Frequentist Inference

Since the annual salaries in the data set all exceed 100000, we decided to transform the salaries by $y_i = f(x_i) = x_i/100000$, where x_i is the annual salary so that every value is re-scaled to a number between 1 and 5.

The histogram of the transformed salary seems to resemble the shape of a gamma distribution. After guessing several times, a gamma distribution $(\alpha, \beta) = (11, 20)$ seems to work okay for now.





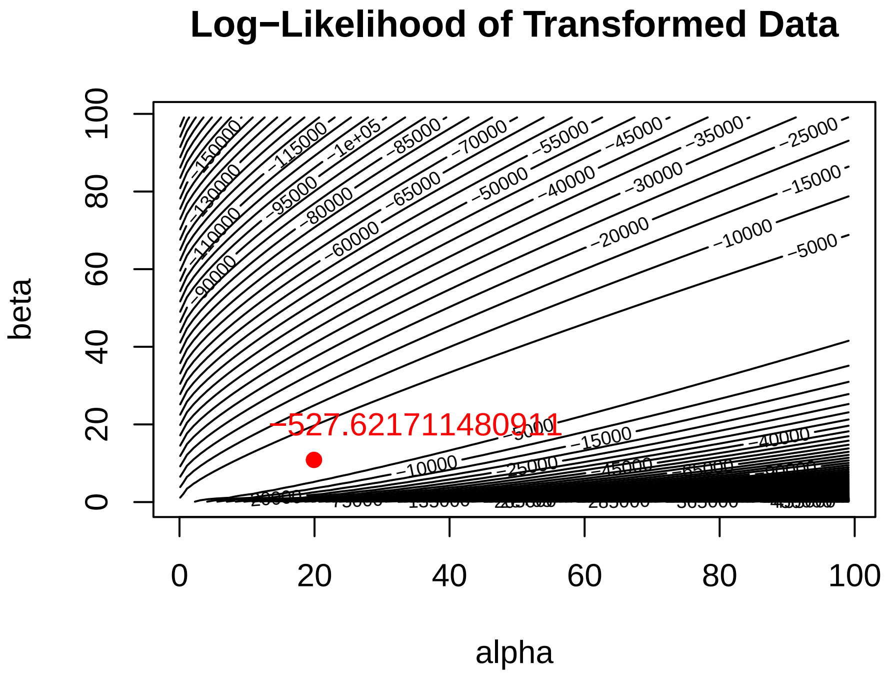
So, the MLE of α and β should be somewhere in the region enclosed by the blue dashed line above.

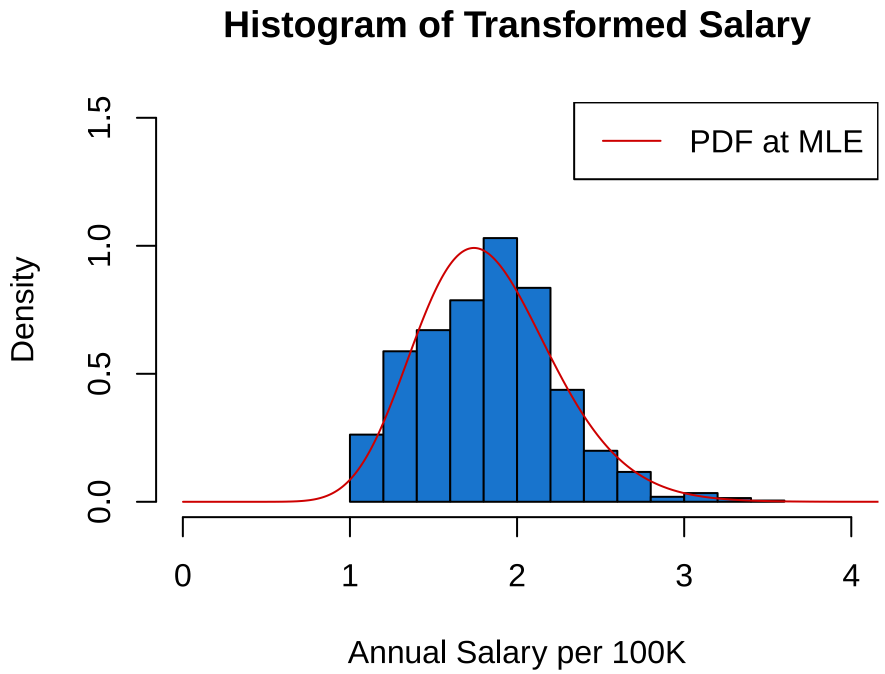
MLE of α and β

The log-likelihood function of $\Gamma(\alpha, \beta)$ is

$$l(\alpha, \beta) = n\alpha \log(\beta) - n \log(\Gamma(\alpha)) + (\alpha - 1) \sum_{i=1}^{n} \log(y_i) - \beta \sum_{i=1}^{n} y_i \ .$$

We use the optim() function in R to find the MLE of α and β:



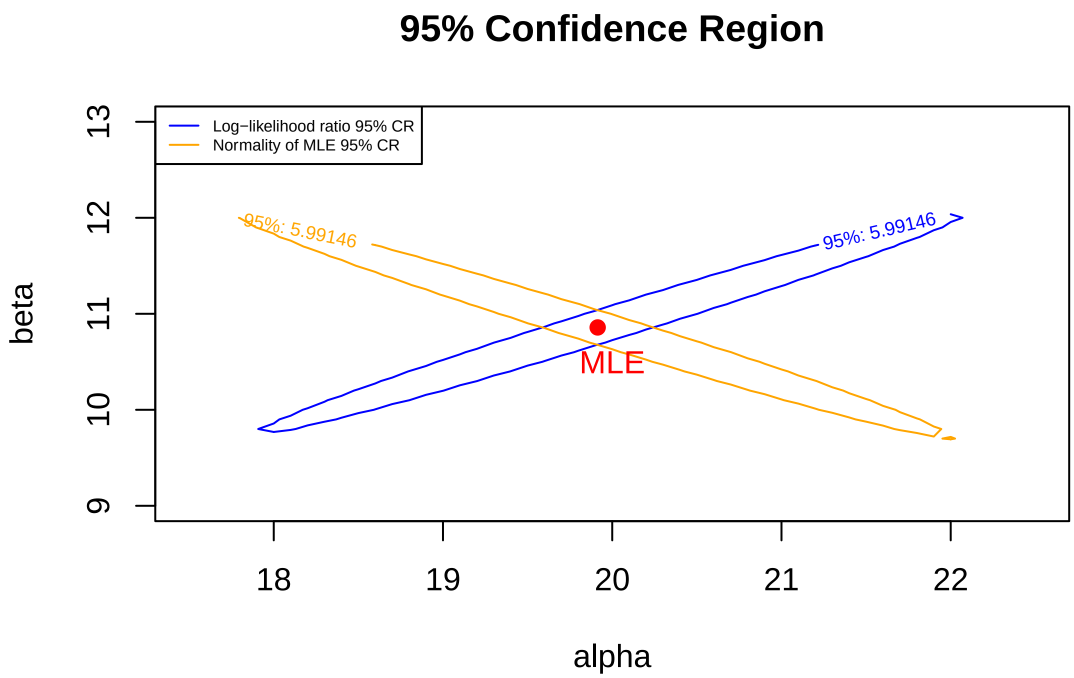


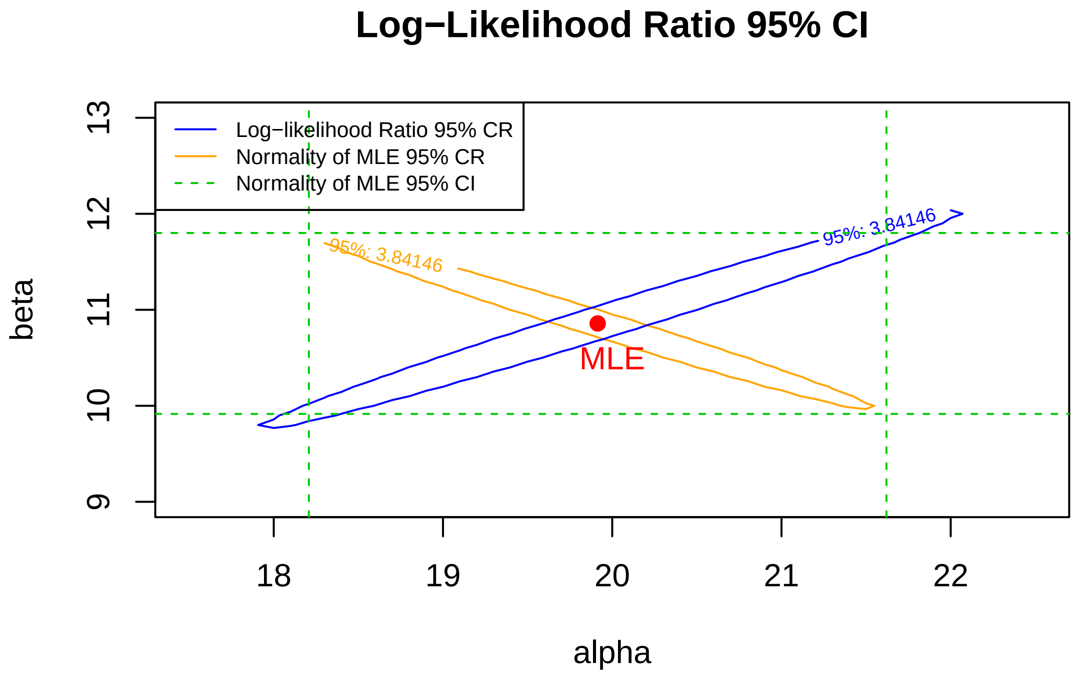
Based on the observed data, we found that $\left(\hat{\alpha},\hat{\beta}\right) = (19.91390, 10.85772)$.

Confidence Region and Interval of α and β

We can construct a 95% confidence region for $\theta = (\alpha, \beta)$ using both the normality of MLE and the log-likelihood ratio statistic:

$$R(\alpha,\beta) = 2\left[l\left(\widetilde{\alpha},\widetilde{\beta}\right) - l\left(\alpha,\beta\right)\right] \to^{\mathcal{D}} \chi^2_{(2)} \quad ; \quad \left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\right)^T \mathcal{I}(\boldsymbol{\theta})\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}\right) \quad \to^{\mathcal{D}} \chi^2_{(2)}\ .$$





Interestingly, the two statistics result in confidence regions that “point in different directions”.

We can find the marginal confidence interval for *alpha* and β using the normality of MLE:

$$\widetilde{\boldsymbol{\theta}}_i \pm Z_{1-\alpha/2}\sqrt{\left[\mathcal{I}(\widetilde{\boldsymbol{\theta}}^{-1}\right]_{ii}}$$

In R:

```
## [1] "Marginal 95% confidence interval for alpha (normality of MLE): [18.20739,21.62042]"
## [1] "Marginal 95% confidence interval for beta (normality of MLE): [9.91547,11.79997]"
```

We saw that a marginal 95% confidence interval for α is $[18.20739, 21.62042]$ and a marginal 95% confidence interval for β is $[9.91547, 11.79997]$ based on the normality of MLE.

Based on the contour plot above, if we instead construct the marginal confidence interval based on the log-likelihood ratio statistic, it should give a very similar result.

Remark

The distribution of professors' annual salary per 100K seems to follow a gamma distribution. The maximum likelihood estimates of the parameters are $\left(\hat{\alpha}, \hat{\beta}\right) = (19.91390, 10.85772)$ with 95% confidence intervals $\alpha \in [18.20739, 21.62042]$ and $\beta \in [9.91547, 11.79997]$.

Bayesian Inference

The result from the frequentist approach shows that the data exhibits the characteristics of a gamma distribution. The maximum likelihood estimates of the parameters are $\left(\hat{\alpha}, \hat{\beta}\right) = (19.91390, 10.85772)$.

Since α and β can only take positive values, let's assume a "truncated bivariate-normal" prior with probability density

$$\pi(\alpha, \beta) = \frac{2}{\pi} e^{-(\alpha^2+\beta^2)/2} \ , \ \alpha > 0 \ , \ \beta > 0 \ .$$

Then, we have the following posterior for α and β:

$$\pi(\alpha, \beta \mid \mathbf{y}) \propto e^{-(\alpha^2+\beta^2)/2} \prod_{i=1}^{n} \frac{\beta^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} e^{-\beta y_i} = \left(\frac{\beta^\alpha}{\Gamma(\alpha)}\right)^n \left(\prod_{i=1}^{n} y_i\right)^{\alpha-1} e^{-\left((\alpha^2+\beta^2)/2+\beta\sum_{i=1}^{n} y_i\right)} \ , \ \alpha > 0 \ , \ \beta > 0 \ ,$$

which I don't know if it has a name.

So, let's use the Metropolis–Hastings algorithm to sample from the posterior.

We can use a bivariate-normal proposal with means equal to their maximum likelihood estimates and variances equal to the diagonals of the inverse of their Fisher information matrix.

We also assume the covariances are 0 in this proposal:

$$(\alpha, \beta) \sim \mathcal{BVN}\left(\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} \ , \begin{bmatrix} \mathcal{I}^{-1}\left(\hat{\alpha}, \hat{\beta}\right)_{11} & 0 \\ 0 & \mathcal{I}^{-1}\left(\hat{\alpha}, \hat{\beta}\right)_{22} \end{bmatrix} \right) \ ;$$

$$q(\alpha^{(t+1)} \mid \alpha^{(t)}, \beta^{(t)}) = \frac{1}{2\pi} \exp\left\{ -\frac{1}{2} \left(\frac{\left(\alpha^{(t+1)} - \hat{\alpha}\right)^2}{\mathcal{I}^{-1}\left(\hat{\alpha}, \hat{\beta}\right)_{11}} + \frac{\left(\beta^{(t)} - \hat{\beta}\right)^2}{\mathcal{I}^{-1}\left(\hat{\alpha}, \hat{\beta}\right)_{22}} \right) \right\} \ ;$$

$$q(\beta^{(t+1)} \mid \alpha^{(t+1)}, \beta^{(t)}) = \frac{1}{2\pi} \exp\left\{ -\frac{1}{2} \left(\frac{\left(\alpha^{(t+1)} - \hat{\alpha}\right)^2}{\mathcal{I}^{-1}\left(\hat{\alpha}, \hat{\beta}\right)_{11}} + \frac{\left(\beta^{(t+1)} - \hat{\beta}\right)^2}{\mathcal{I}^{-1}\left(\hat{\alpha}, \hat{\beta}\right)_{22}} \right) \right\} \ ;$$

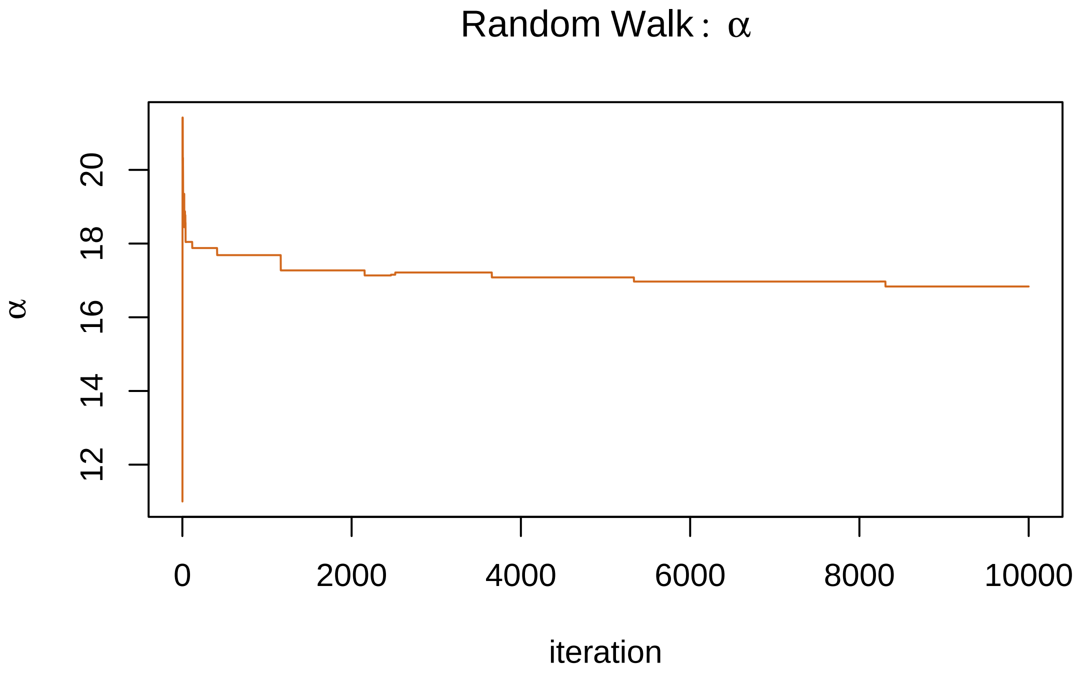
So, the acceptance probabilities are

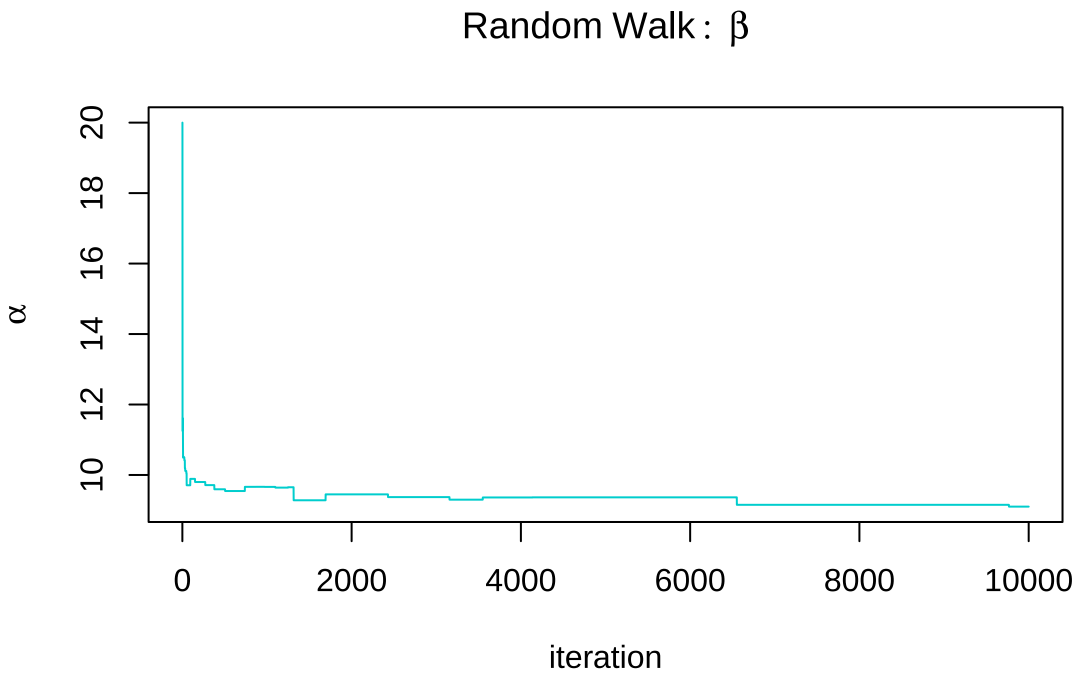
$$\mathbb{P}_\alpha = \min\left\{1\ ,\ \frac{\left(\frac{{\beta^{(t)}}^{\alpha^{(t+1)}}}{\Gamma(\alpha^{(t+1)})}\right)^n \left(\prod_{i=1}^n y_i\right)^{\alpha^{(t+1)}} \exp\left\{-\frac{1}{2}\left(\left(\alpha^{(t+1)}\right)^2 + \frac{\left(\alpha^{(t)}-\hat{\alpha}\right)^2}{\mathcal{I}^{-1}\left(\hat{\alpha},\hat{\beta}\right)_{11}}\right)\right\}}{\left(\frac{{\beta^{(t)}}^{\alpha^{(t)}}}{\Gamma(\alpha^{(t)})}\right)^n \left(\prod_{i=1}^n y_i\right)^{\alpha^{(t)}} \exp\left\{-\frac{1}{2}\left(\left(\alpha^{(t)}\right)^2 + \frac{\left(\alpha^{(t+1)}-\hat{\alpha}\right)^2}{\mathcal{I}^{-1}\left(\hat{\alpha},\hat{\beta}\right)_{11}}\right)\right\}}\right\}\ ;$$

$$\mathbb{P}_\beta = \min\left\{1\ ,\ \frac{\left({\beta^{(t+1)}}^{n\alpha^{(t+1)}}\right) \exp\left\{-\left(\frac{{\beta^{(t+1)}}^2}{2} + \beta^{(t+1)}\sum_{i=1}^n y_i + \frac{\left(\beta^{(t)}-\hat{\beta}\right)^2}{2\mathcal{I}^{-1}\left(\hat{\alpha},\hat{\beta}\right)_{22}}\right)\right\}}{\left({\beta^{(t)}}^{n\alpha^{(t+1)}}\right) \exp\left\{-\left(\frac{{\beta^{(t)}}^2}{2} + \beta^{(t)}\sum_{i=1}^n y_i + \frac{\left(\beta^{(t+1)}-\hat{\beta}\right)^2}{2\mathcal{I}^{-1}\left(\hat{\alpha},\hat{\beta}\right)_{22}}\right)\right\}}\right\}\ .$$

Results

We run the MCMC with an initial value of $\left(\alpha^{(0)}, \beta^{(0)}\right) = \left(\alpha^{(0)}, \beta^{(0)}\right) = (11, 20)$, which was our initial guess of the parameter values in the frequentist approach above:





```
## [1] "Acceptance Probilities:"

## [1] "alpha:"

## [1] 0.002

## [1] "beta:"

## [1] 0.0029

## [1] "________________________________________"

## [1] "Mixing Measures:"

## [1] "alpha:"

## [1] 0.01119483

## [1] "beta:"

## [1] 0.007780716
```

It turns out the performance of the MCMC is very bad. The acceptance probability for α is 0.002, and the acceptance probability for β is 0.0029. Both α and β have very low mixing measures.

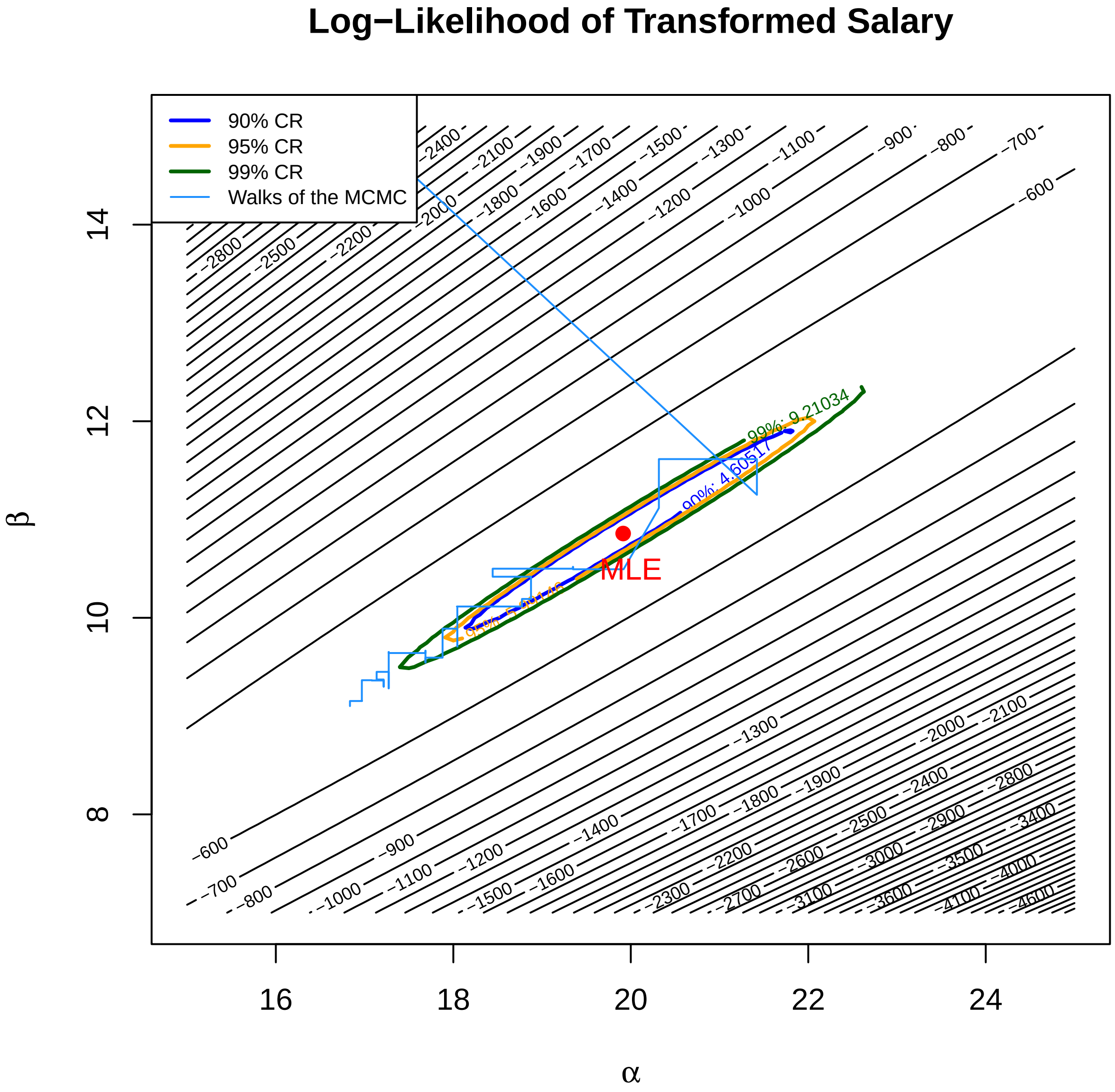
An 95% credible interval for α:

```
##     2.5%    97.5%
## 16.83527 17.87898
```

An 95% credible interval for β:

```
##      2.5%     97.5%
## 9.153408 9.712645
```

Eventually, we overlap the path of the MCMC onto the log-likelihood contours at the beginning:



Remark

It seems that the MCMC moves around slightly in the area covered by the three confidence regions. So, despite working very badly, the Bayesian framework still somewhat agrees with the Frequentist framework in terms of the inference about the parameter α and β in this case. However, since the MCMC performs very badly here, it is worth changing the prior and/or the proposal density and running the MCMC with the new density in the future to get better into the Bayesian inference for this data set.