Bayesian Clustering and Survival Analysis Based on Pro-Act Database

Berggren H. Gullbrandson M. Karmam H.

February 18, 2020

Contents

1 Introduction

1.1 Amyotrophic lateral sclerosis

Amyotrophic lateral sclerosis (ALS) is an autoimmune disease, a category of diseases where the immune system of the body starts to react against itself. Specifically, for ALS the immune system attack the nerve cells that carries electric impulses from the brain to the rest of the body. This leads to a gradually degeneration of muscle capacity to the point of complete death of the nerve cells. This development makes the disease very serious and 50% of patients do not survive more than 3 years after the first signs of symptoms. An additional 25% do not survive more than 5 years [1]. This is a consequence of the disease reaching the nerve cells that is connected to the voluntary muscles which control life supporting motions, for example breathing.

The cause of the disease is highly uncertain according to previous researches. However there are some indications that genetic and environmental factors have importance in the development of ALS. Roughly 10% of all cases of ALS are so called familial (with high uncertainty at the moment) where the patient has family history with the disease. For this type of ALS the average start of the disease occurs in the age span 47-52 [1]. In the rest of the cases the cause is unknown, factors such as smoking, moderate to severe brain injury is found in different reviews to be linked to ALS and for this type of the disease the average age for signs of the first symptoms are slightly postponed compared to familial ALS and the average start to show signs are the interval of years 58 to 63 [1].

Recently the PRO-ACT (Pooled Resource Open-Access ALS Clinical Trials) database[1] have released a large amount of de-identified cases of ALS. This have spurred an influx of research of the disease to identify for example to progression or potential interesting subgroups of the disease [2][3].

1.2 Aim

The overall goal of this study is to derive interesting conclusions by the means of Bayesian statistics. We aim to select a sizeable group of patients with a selection of lab results as well as other seemingly important information. Given such a data set we want to apply an GLMM (Generalized Linear Mixed Model) to cluster the data with the intent to create a specific survival analysis for patients in a certain cluster and compare this to a general survival analysis created on the complete data set.

[1] https://nctu.partners.org/ProACT/Home/Index

2 PRO-ACT Database

The data used throughout the project originates from the PRO-ACT database. A database with over 8500 patients diagnosed with ALS and data from clinical trials. The data consists of results from laboratory analysis, recorded survival and various well-being measurements that represent the reduction of physical capacity due to the disease. As mentioned previously this have motivated data-driven analysis of the disease, however the first step for any such analysis must be preprocessing of data. Unfortunately the amount of patients with no missing values are vastly lower than the ideal total 8500.

2.1 Selection of Data

To find a suitable subset of patients we need to limit what covariates that is used. As a starting point all patients that have no registered day of passing was removed because of an absence of actual measures to how long they live, and also given the inevitable nature of the disease such patients are simply lacking this date due to incomplete measurements or exiting the study. By removing these patients we remove any censored data and simplify future operations. From here we included ALSFRS (ALS Functional Rating Scale) that quantify a patients ability to perform certain trivial motions, FVC (Forced Vital Capacity) which measures an individuals lung capacity, age, sex and a combination of lab results. An initial selection of lab results were based on the amount of available data. However after studying the correlation between covariates it was seen that ALSFRS and all our chosen lab results except "Creatine" and "Hemoglobin" should be removed. Finally we have selected to include a binary covariate indicating whether the initial onset of the disease if on a limb or any other location of the body as this have been deemed important in previous clustering [2]. The final set of covariates are then: age, sex, onset, creatine, hemoglobin and FVC.

3 Clustering

To begin the analysis of the data from the PRO-ACT database we try to use a Bayesian approach to cluster the given data. This may be done in multiple ways but in common for all is the usage of MCMC to find partitions and finally a loss function to estimate a point estimate partition.

3.1 GLMM

The first approach to create clusters was naturally to mimic clustering done during class. We therefore first set out to create GLMM that would be able to classify patients to either of an unknown number of clusters. Let Y_i be the discrete time of survival for patient $i = 1, ..., N$, we may then write our model as:

$$\begin{aligned} Y_i &\sim \mathcal{MVN}(1, \theta_i) \\ \theta_{i,j} &= f(\beta_j \mathbf{x}_i + b_i) \\ \beta_j &\sim N(0, \sigma^2) \\ b_i | P &\sim DP(\alpha, P_0). \end{aligned} \tag{1}$$

Here $j = 1, ..., B$ represent the discrete time steps, thus $\theta_{i,j}$ each represent the probability that patient i belong to the time step j. Here $\mathcal{MVN}$ is a multinomial distribution and N is the normal distribution. The function $f(\beta_j \mathbf{x} + b_i)$ is the normalized probability given a linear regression between the patient specific covariates, $\mathbf{x}_i$, and constants β_j. Furthermore b_i sampled from a Dirichlet Process and may be written as:

$$f(\beta_j \mathbf{x}_i + b_i) = \frac{e^{\beta_j \mathbf{x}_i + b_i}}{\sum_k e^{\beta_k \mathbf{x}_i + b_i}}, \qquad \forall k \in \{1, ..., B\}.$$

This kind of transformation may be used in combination with multinomial distribution as by C. Holmes and L. Held [5], and is commonly referred to as a softmax. Furthermore we have used $\sigma^2 = 1000$ meaning that our prior for β is to be considered flat. Implementing model 1 we've used a combination of R and BUGS where the Dirichlet Process is realized by a stick-breaking process.

To find a useful partition or clustering from the many partitions generated through the MCMC we need a point estimate [6]. The general idea is to, given a loss function L and true partition $\mathbf{c}$, find a partition $\mathbf{c}^*$ such that

$$\mathbf{c}^* = \arg\min_{\hat{\mathbf{c}}} \sum_{\mathbf{c}} L(\mathbf{c}, \hat{\mathbf{c}}) p(\mathbf{c}|\mathbf{x}). \tag{2}$$

For model (1) we have used the Binder's loss. At it's core the Binder's loss,

$$B(\mathbf{c}, \hat{\mathbf{c}}) = \sum_{n<n'} l_1 \mathbf{1}(c_n = c_{n'}) \mathbf{1}(\hat{c}_n \neq \hat{c}_{n'}) + l_2 \mathbf{1}(c_n \neq c_{n'}) \mathbf{1}(\hat{c}_n = \hat{c}_{n'}), \tag{3}$$

tries to penalize either false negatives or false positives. Here l_1 and l_2 are constants and c_n is the cluster for patient n by partition $\mathbf{c}$. Now setting the

penalty for each type of error equal and to one, $l_1 = l_2 = 1$ we may significantly simplify (2) [6]. The simplified version of (2) using a Binder's loss were $l_1 = l_2$ may be written as:

$$\mathbf{c}^* = \arg\min_{\hat{\mathbf{c}}} \sum_{n<n'} |\mathbf{1}(c_n = c_{n'}) - p_{n,n'}|. \tag{4}$$

Were $p_{n,n'}$ is the posterior probability that the two points n and $n^{'}$ is in the same cluster. Implementations of a point estimation with a Binder's loss thus usually implement a so called similarity matrix for $p_{n,n'}$ which makes the actual calculations straight forward.

3.2 Results

While running the model given in (1) we monitor primarily the mixing of the MCMC model and verify by looking at the traceplots of the regression parameters, β. As described in section 2.1 six covariates is used and the regression coefficients for the discrete time step 11 is presented in figure 1. The traceplots for the rest of the time steps are included in the appendix A and where found to behave in a similar way except for β_j which showed less mixing most likely due to the fact that only four patients belonged to this group.

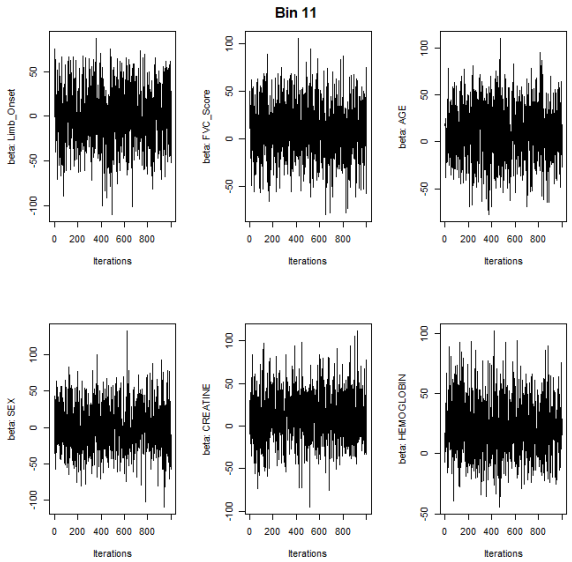


Figure 1: Traceplot of regression parameters $\beta_{11,j}$.

Furthermore the autocorrelation of the regression parameters where studied to determine the independence of the model. This plot of autocorrelation for time step eleven and its respective covariate is shown in figure 2. Again the autocorrelation for the other time steps where found to be comparable to their traceplots and are included in the appendix.

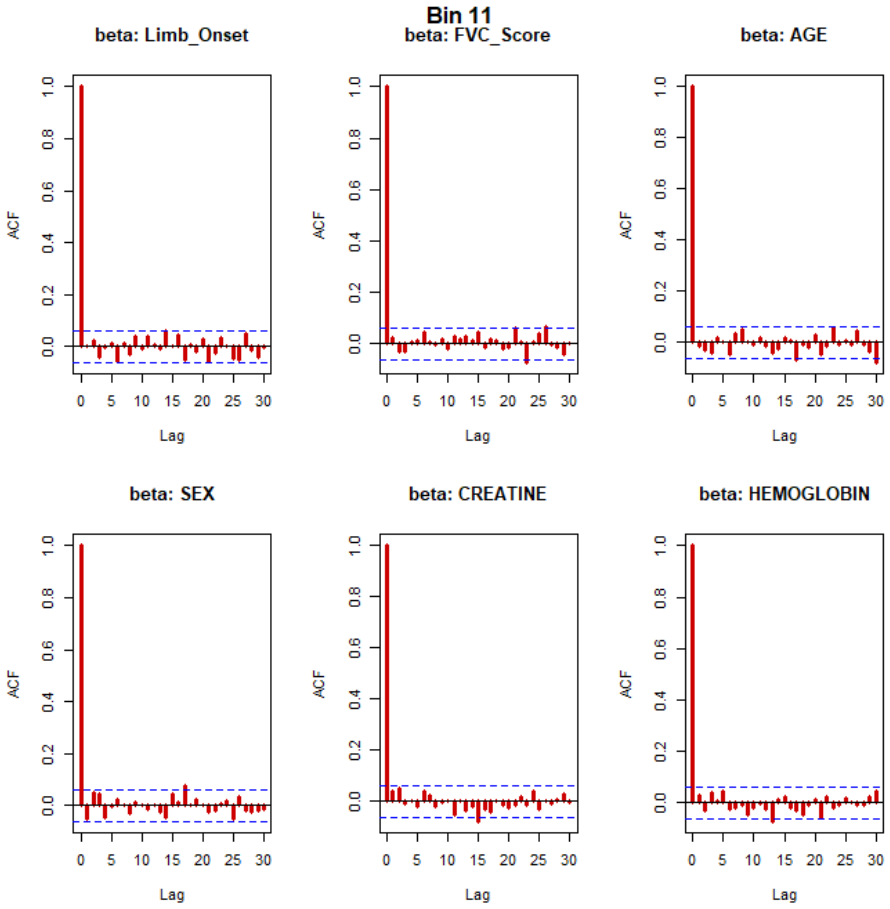


Figure 2: The autocorrelation of $\beta_{11,j}$.

To show the a posterior distribution of the regression parameters density plots were created to certify that the MCMC where run for an sufficient amount of iterations so that meaningful conclusions could be drawn. In figure 3 the density plot from corresponding time step eleven is shown for all regression parameters connected to each of the selected covariates.

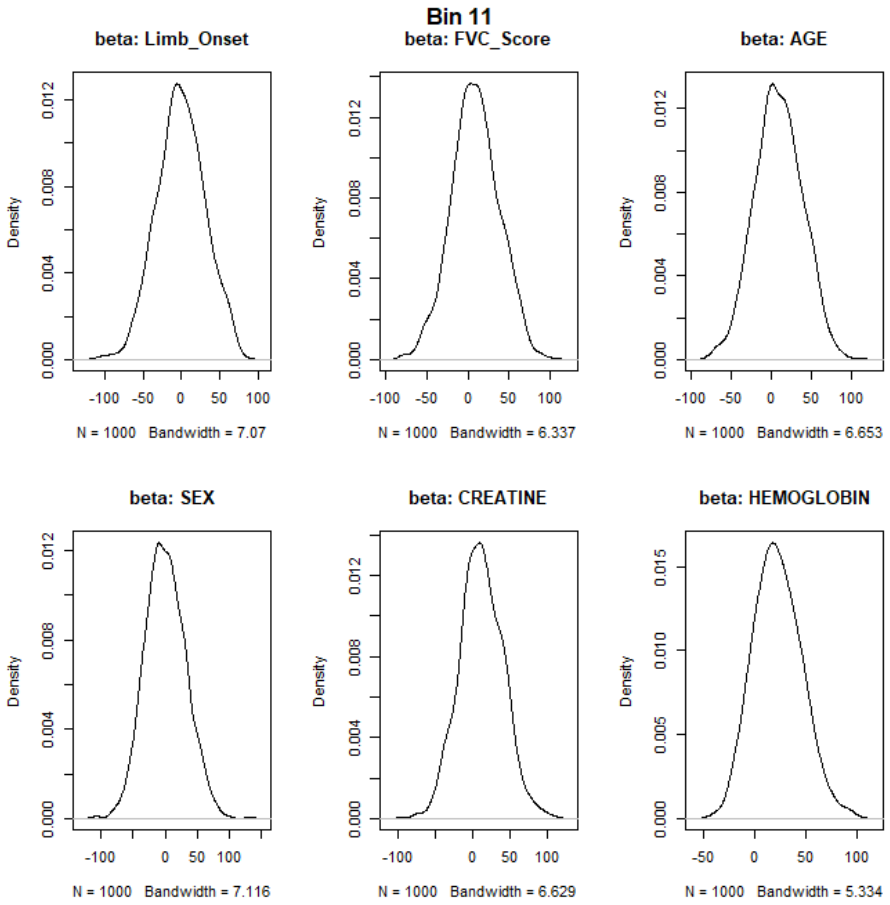


Figure 3: Density plot of $\beta_{11,j}$

3.2.1 Resulting Clusters

So far we have looked at the computational aspects of the model. The result of the model and an implementation of Binder's loss give us one partition that seem best fitted. We may summarize the sizes of clusters as well as the number of clusters in table 1 given below.

With a total amount of 24 clusters it is hard to find any inherent meaning for each cluster. We may however see that there are slight variations of mean for covariates between the clusters. For example looking at a subset of clusters, cluster 1, 2, 4 and 5, we can summarize the mean values as in table 2.

Cluster	1	2	3	4	5	6	7	8	9	10	11	12
# Members	50	64	8	38	32	10	24	2	56	31	54	34
Cluster	13	14	15	16	17	18	19	20	21	22	23	24
# Members	2	5	5	8	1	17	10	12	3	2	1	1

Table 1: Summary of the size of clusters.

	Survival Time	FVC	Age	Creatine	Hemoglobin
Cluster 1	241.6	60.0	58.6	81.9	147.3
Cluster 2	262.4	69.0	58.4	78.6	147.5
Cluster 4	261.0	64.7	61.6	76.3	147.4
Cluster 5	254.2	67.5	58.2	78.0	148.4

Table 2: Summary of the mean values for covariates.

3.3 Discussion

Given the large amount of clusters obtained it is unclear what physical, if there is one, attributes each cluster represents. Given that we do see some variations between the clusters it does not however seem unreasonable to expect our partition to carry some kind of information albeit weak. We thus think that clustering might still be possible, even if model (1) is not the desired model. One aspect that is not considered in the used model is the time dependency between the different sampled values of the multinomial distribution. Implementing the fact that a value of 1 from the multinomial distribution is temporally preceding a result of 2, might lead to more clear results.

To continue with the survival analysis, the next part of the project, we have however focused our attention on one of the larger clusters. By fitting a survival model for a certain cluster we hope to find whether the clusters and their inherent information help estimating a survival function.

4 Survival analysis

In this part of this project, we aim analyze the expected duration of time until the death of patients. As stated previously in the report ALS is a serious disease and it is thus naturally of interest to predict the survival time of patients. In the PRO-ACT database the recording of survival times is reported and this is used in the created analysis. We focus on creating a model for all the data and cluster number 2 as one of the larger clusters provided by the clustering algorithm. Via an MCMC sampling with the same covariates used in the clustering an Weibull model is used for our analysis. The background of the Weibull model an our

implementation will be covered in the following sections.

4.1 Weibull model

The Weibull model is one of the most used models in parametric survival analysis [4]. Here the idea is to use a Weibull distribution to sample survival times T_i for each patient, $i \in \{1, \ldots, n\}$. T_i is supposedly independent and identically distributed with a Weibull distribution, $T_i \sim \mathcal{W}(\alpha, \gamma_i)$ with parameters α and γ_i. In our case $\log(\gamma_i) = \beta \mathbf{x}_i$, meaning it is the link between the data, $\mathbf{x}$, and our model. The density function for any given time and patient thus is,

$$f_{T_i}(t_i \,|\alpha, \gamma_i) = \alpha t_i^{\alpha-1} \exp(\gamma_i - \exp(\gamma_i) t_i^{\alpha}). \tag{5}$$

Here t_i is the observed survival time for patient i. From the density function the survival function is derived as: $S(t_i \,|\alpha, \gamma_i) = \mathbb{P}(T_i < t_i \,|\alpha, \gamma_i) = \exp(-\exp(\gamma_i) t^{\alpha})$ [4]. To predict the survival time for the ALS patients we set up the MCMC model accordingly

$$\begin{aligned} T_i &\sim \mathcal{W}(\alpha, \gamma_i) \\ \log(\gamma_i) &= \beta \mathbf{x}_i \\ \beta &\sim N(0, \sigma^2) \\ \alpha &\sim \mathcal{G}(1, \xi). \end{aligned} \tag{6}$$

For the model we decided prior distributions for the parameters β and α with parameters $\sigma^2 = 1000$ and $\xi = 10^{-4}$. This ensures a reasonably flat prior for β, and ξ is set as proposed in [4] p. 35. The model was implemented in R and JAGS for both the entire data set but also the specified cluster.

4.2 Results

As discussed we wish to compare our model (6) trained on one cluster in specific with one trained with all available data. Again we wish to get an idea how the models are mixing and do this by monitoring β and α. In figure 4 we can see how most of the β's are indeed mixing well. Unfortunately neither versions of the model seem to be able to handle the presence of the covariate "hemoglobin".

The poor performance for the final covariate is maybe even more obvious as we in figure 5 see how the autocorrelation for this specific parameter is not up to par to the rest.

Furthermore for reference we see in figure 6 how the corresponding posterior densities for β look.

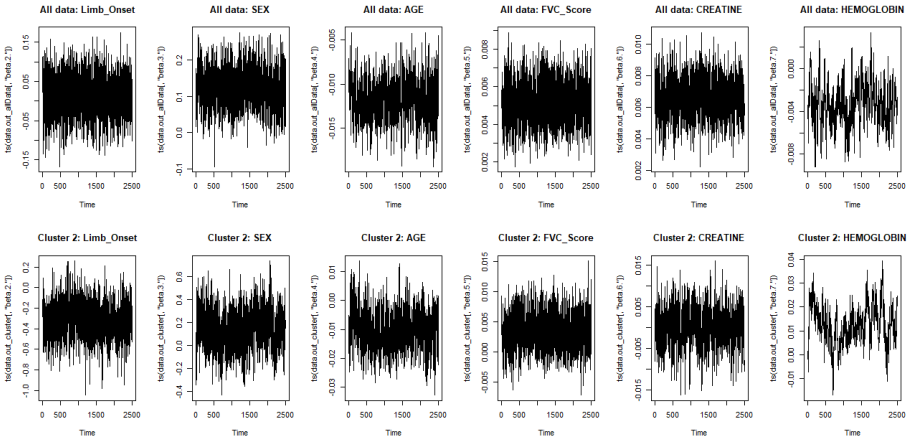


Figure 4: Traceplots of most β's for both versions of our model.

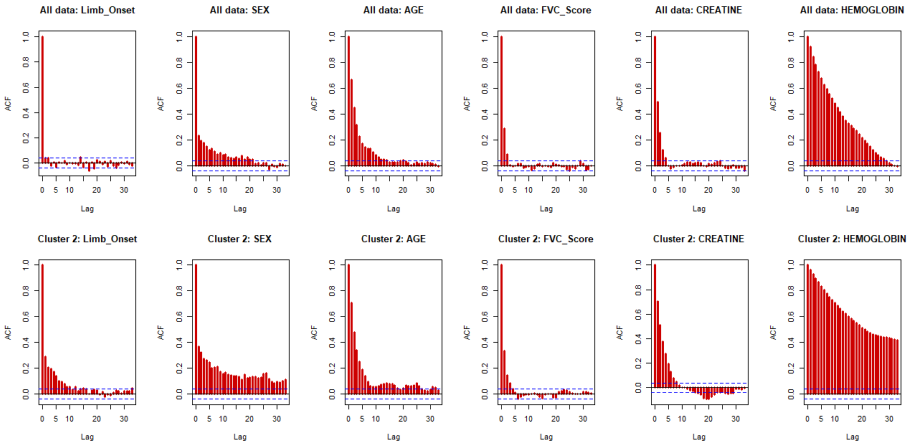


Figure 5: Autocorrelation of most β's for both versions of our model.

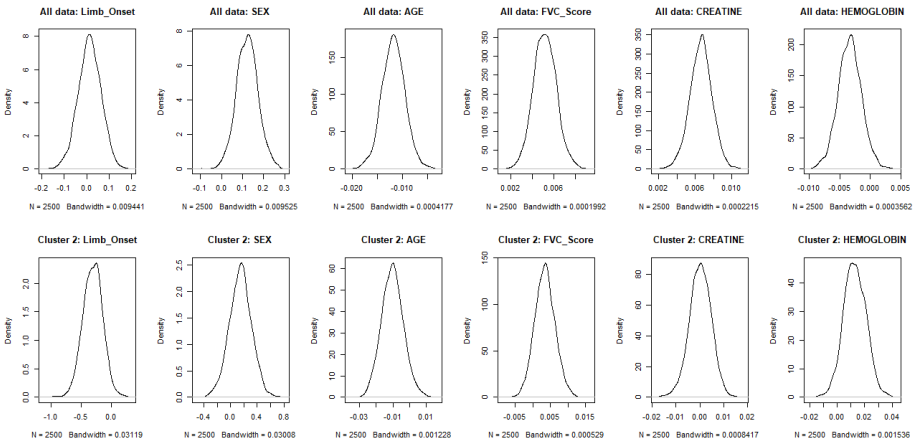


Figure 6: Posterior density of most β's for both versions of our model.

4.2.1 Resulting Survival Functions

Using the data from the MCMC-chain we may calculate the survival function for each patient. In figure 7 we see the 95% credibility interval for the survival function of both the model fitted on all data as well as the model fitted on only cluster 2.

Similar figures may be calculated for all patients in each cluster one tried to fit with model (6). Figure 8 illustrate survival functions for two other patients also in cluster 2.

4.3 Discussion

The results presented in section 4.2.1 are not very clear, the fact that the credibility intervals of the different survival functions overlap gives us no significant result. We may however notice that the credibility interval of the survival function fitted on the clustered data in general is larger. This is reasonable when considering that there is less data to use for fitting in each given cluster compared the entire data set. There nevertheless seem to be a tendency for the credibility interval of the survival function fitted on the clustered data, compared to the other model, to indicate if whether the recorded date of death is "high" or "low". Again there is no significant evidence to prove such a conclusion but may motivate further studies of clustering and the relation of clustering and survival on the PRO-ACT database.

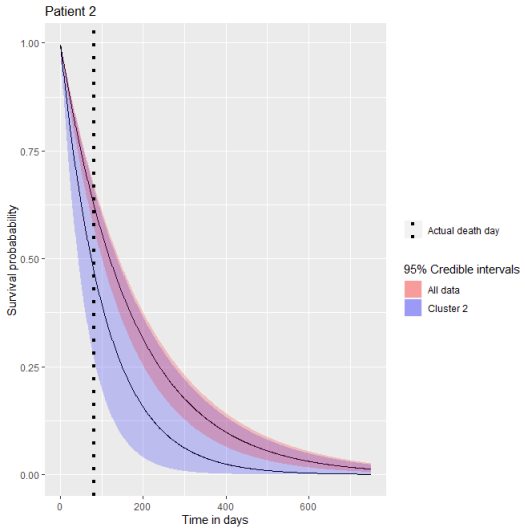
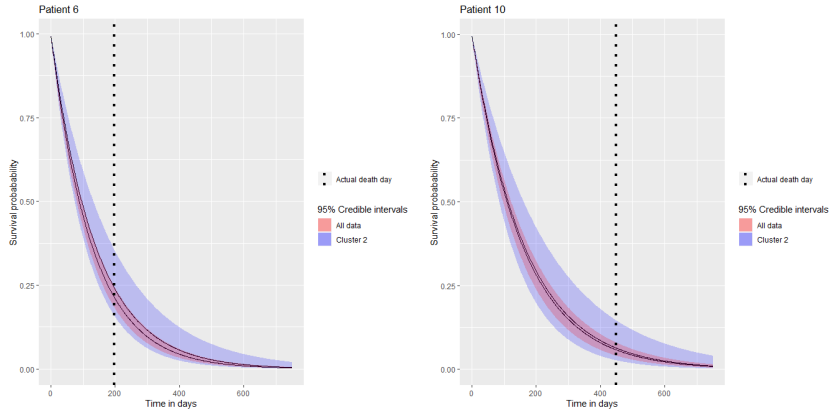


Figure 7: Survival probabilities for patient 2 both from specific survival analysis done on cluster 2 and on all data.



(a) Patient 6. (b) Patient 10

Figure 8: Survival probabilities both from specific survival analysis done on cluster 2 and on all data.

5 Conclusion

We begun the analysis on this project by trying to process a lot of data sets with medical measures and lab tests, we had to deal with incomplete and noisy data sets with a lot of inaccurate values and missing information, join measurement

on just two variables could erase all the data sets.

To choose an approach and features selection we followed statistical tools to try to find features with big variability and with intense effect of the survival of patients.

We begun the Bayesian analysis by using a GLMM model to cluster patients with the chosen covariates. The second part of the project is the Bayesian survival analysis with a Weibull model fitted both on the general data set and a specific data set. We see a kind of pattern that is different between the two survival functions in the results, but nothing significant. It does however give an indication and could be subject to further studies.

This project was important to us, because we went throw all the steps of statistical analysis from prepossessing and dealing with real data sets, analyse of data sets, choosing between covariates, building the model and check the assumptions, it teaches us also the important and the accuracy of the Bayesian statistical approach to deal with real life problematic.

References

[1] M. Kiernan, S. Vucic, B. Cheah et al.: Amyotrophic lateral sclerosis, The Lancet (2011) Available from: https://www.sciencedirect.com/science/article/pii/S0140673610611567?via%3Dihub.

[2] R. Kueffner, et. al.: Stratification of Amyotrophic Lateral Sclerorsis patients: a crowdsourcing approach. BioRxiv (2018) Available from: https://www.biorxiv.org/content/10.1101/294231v1

[3] R. Kueffner, et. al.: Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. Nature Biotechnology, volume 33, number 1 (2015).

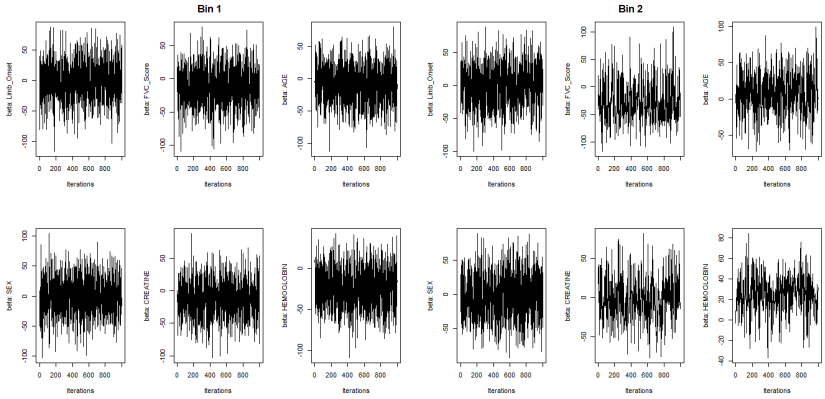
[4] J. G. Ibrahim, M. Chen, D. Sinha: Bayesian Survival Analysis. New York: Springer.

[5] C. Holmes, L Held: Bayesian Auxiliary Variable Models for Binary and Multinomial Regression. Bayesian Analysis, number 1 (2006)

[6] S. Wade, Z. Ghahramani: Bayesian Cluster Analysis: Point Estimation and Credible Balls (with Discussion). Bayesian Analysis, number 2 (2018)

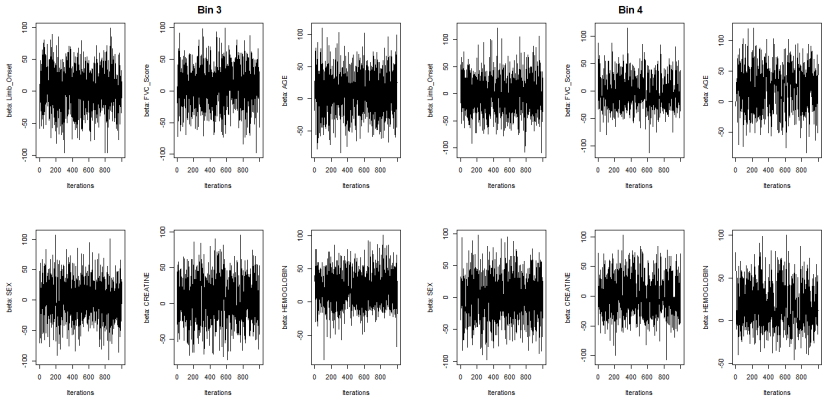
A MCMC plots for clustering

Here all the traceplots of all time steps is presented, named 1 through 20.



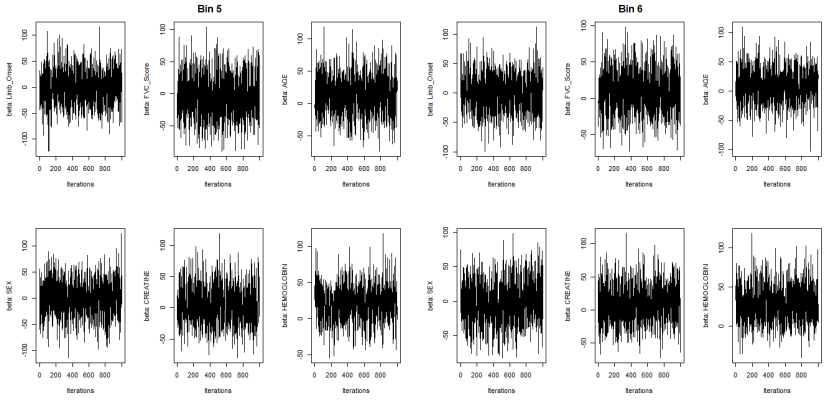
(a) Time step 1.

(b) Time step 2.



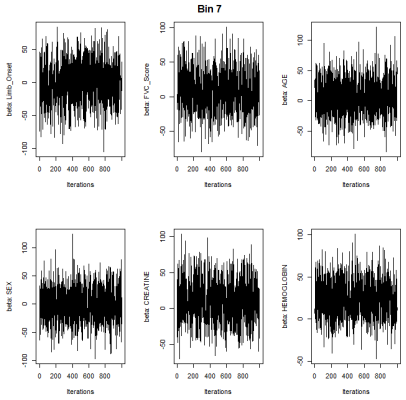
(a) Time step 3.

(b) Time step 4.

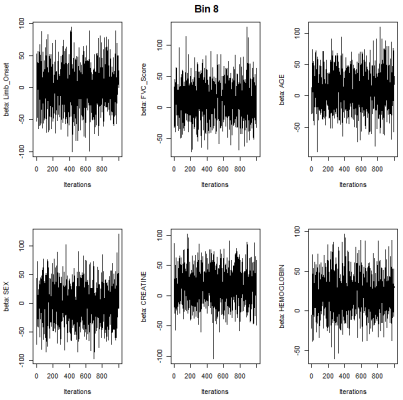


(a) Time step 5.

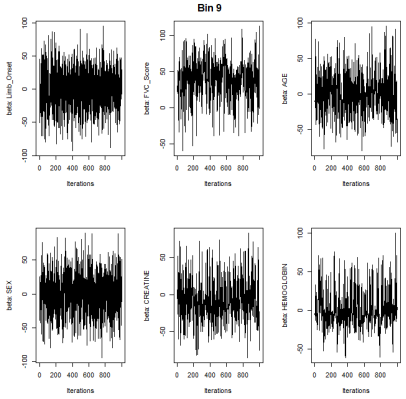
(b) Time step 6.



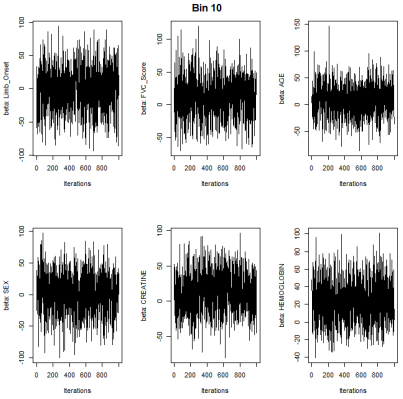
(a) Time step 7.



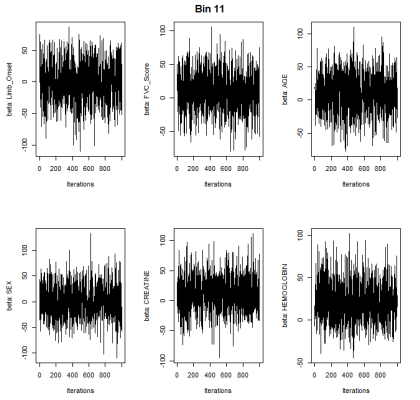
(b) Time step 8.



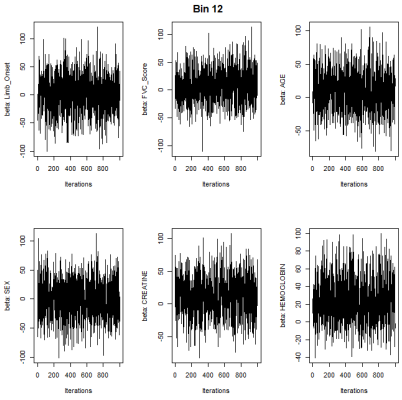
(a) Time step 9.



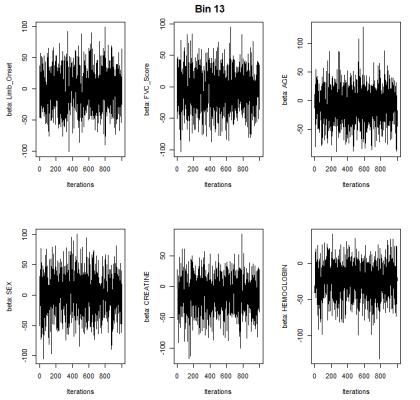
(b) Time step 10.



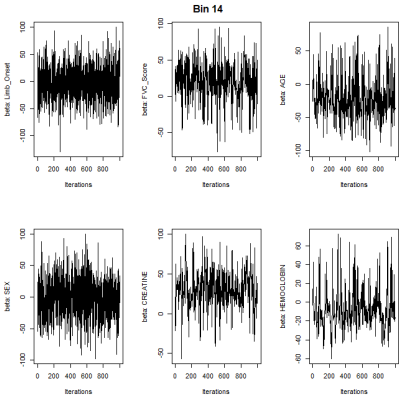
(a) Time step 11.



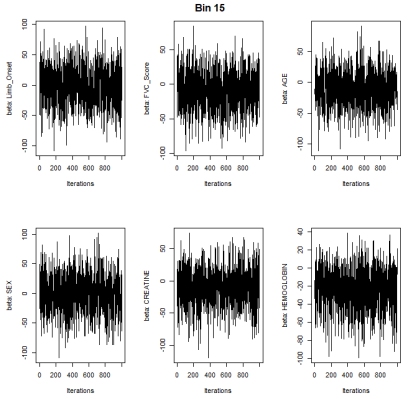
(b) Time step 12.



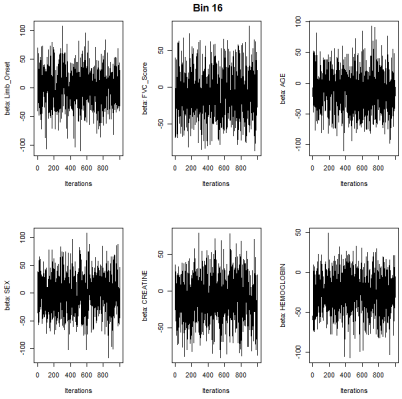
(a) Time step 13.



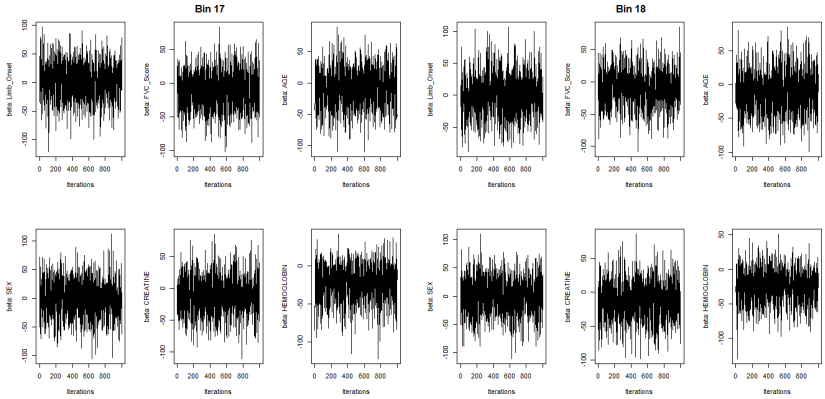
(b) Time step 14.



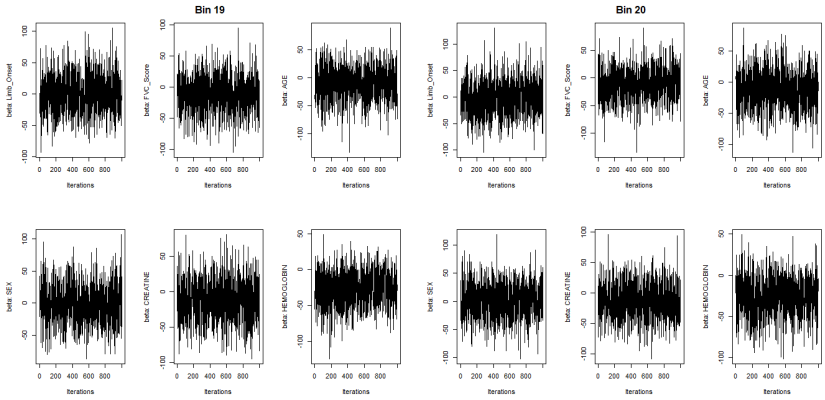
(a) Time step 15.



(b) Time step 16.

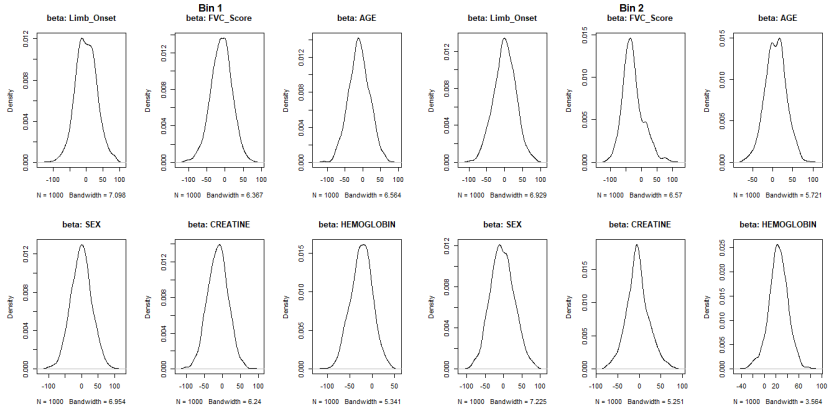


(a) Time step 17. (b) Time step 18.



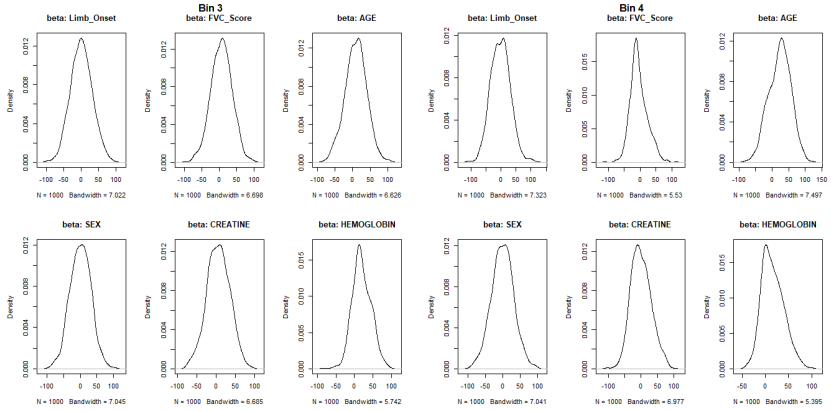
(a) Time step 19. (b) Time step 20.

The plotted density functions for all time steps, 1 through 20.



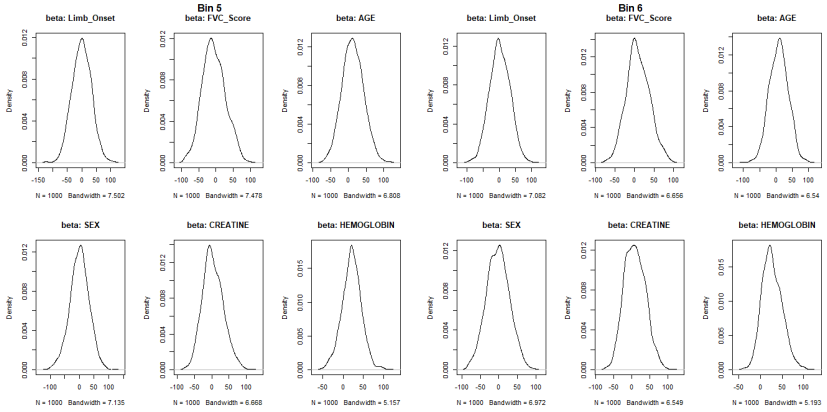
(a) Time step 1.

(b) Time step 2.

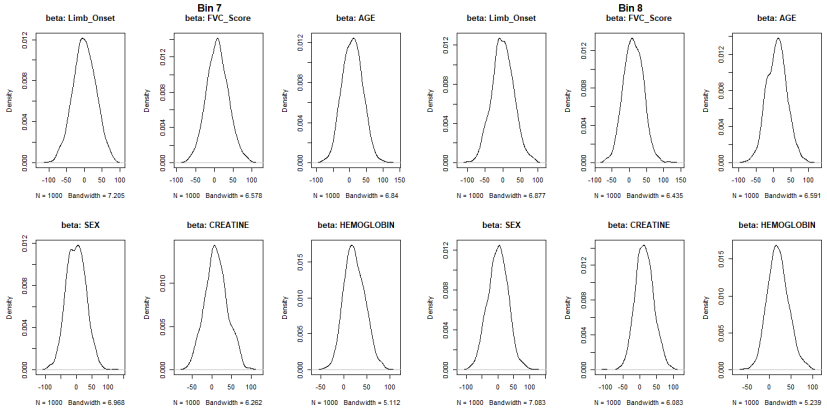


(a) Time step 3.

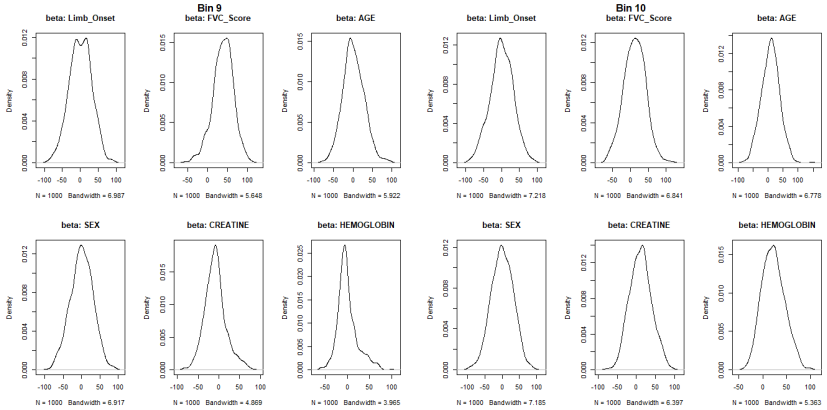
(b) Time step 4.



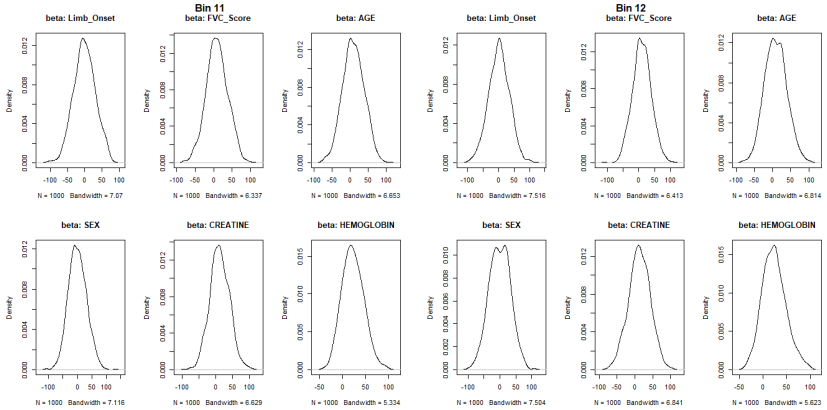
(a) Time step 5. (b) Time step 6.



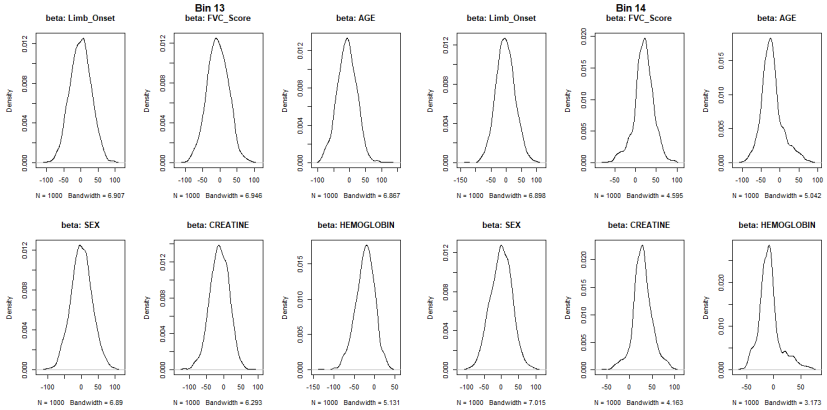
(a) Time step 7. (b) Time step 8.



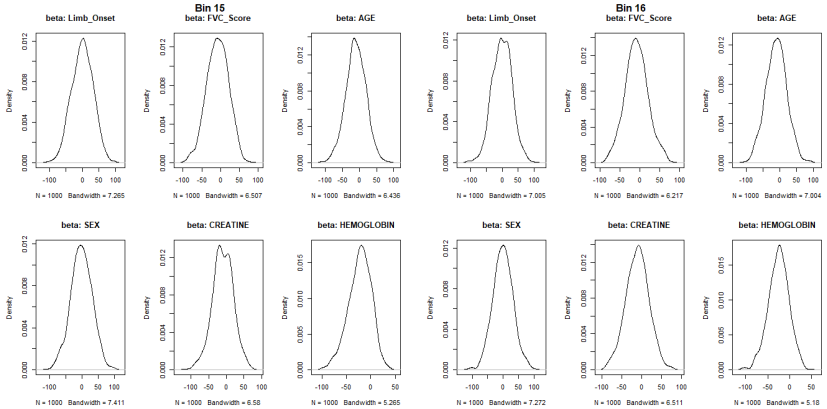
(a) Time step 9. (b) Time step 10.



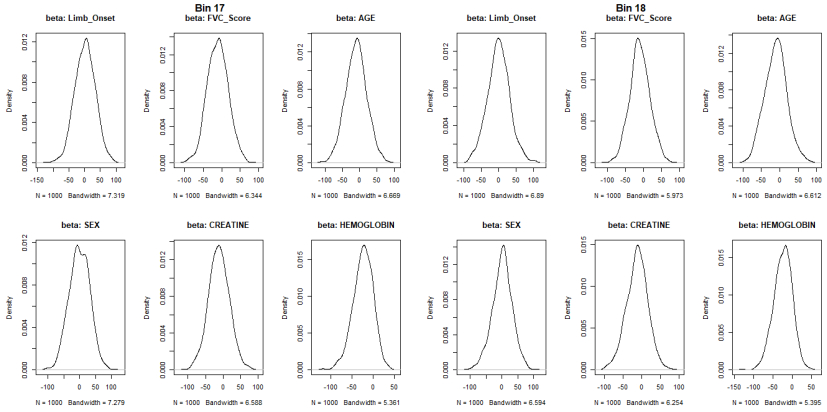
(a) Time step 11. (b) Time step 12.



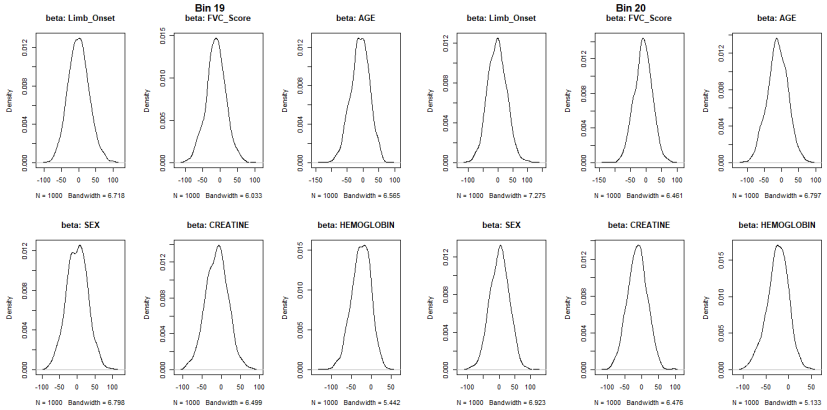
(a) Time step 13. (b) Time step 14.



(a) Time step 15. (b) Time step 16.

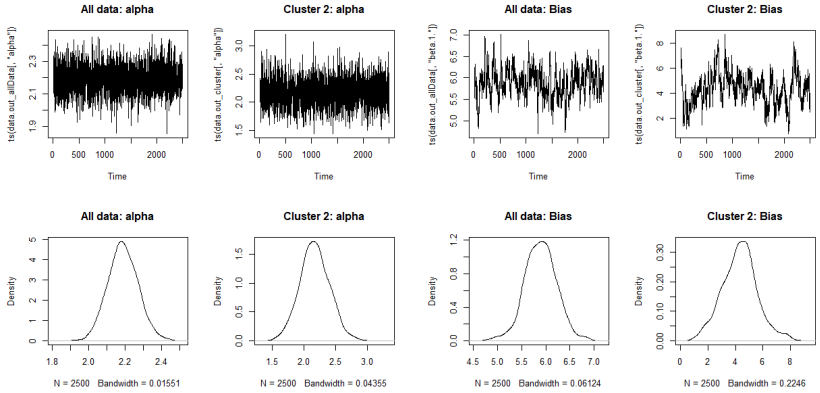


(a) Time step 17. (b) Time step 18.



(a) Time step 19. (b) Time step 20.

B MCMC plots for survival analysis



(a) Mixing of alpha for both types of survival analysis

(b) Mixing of β_1 for both types of survival analysis