Duration of Erasmus+ Mobility Periods

Ulrik Bernhardt Danielsen & Alexander Johansen Ohrt

June 9, 2022

1 Abstract

The duration of Erasmus+ stays abroad is modelled using a Bayesian framework. Finite mixture models are employed—both because of the observed data and because of expectations concerning the durations, which seem to usually be either one or two semesters. Four different models are tested, in order of increasing complexity, with the purpose of finding a model that models the duration adequately. The conclusion is that the suggested models can be implemented successfully, where the success is measured by some ad-hoc statistics. In additon, the exchange duration seems to depend on the gender and the age of each student; females and younger students seem to go for slightly shorter durations when comparing to males and older students. Finally, some more models are proposed without thorough investigation, in hopes of providing inspiration for further improvement of the procedure and the results.

2 Introduction

Every year several hundred thousand students travel abroad to study as exchange students trough the Erasmus+ program. The program is an agreement between many European countries in which students can study abroad without paying tuition at the receiving institution, and guarantees recognition of completed courses when returning home. Additionally, students can apply for a grant which can help cover costs if the cost of living increases during the exchange period. Students are allowed to apply for the Erasmus+ program after at least one year of completed studies at university level. The length of the stay can be both one and two semesters depending on the specific situation—both at bachelor and master level.

In this report we model the duration of Erasmus+ periods using a Bayesian framework. The reason behind this choice of objective is relatively simple; during the time of writing the authors are coming towards the end of a fruitful Erasmus+ exchange in Barcelona. After experiencing relatively different exchange durations in our peers, depending on factors like country of origin and perhaps age, it seems interesting to explore some data in order to get some more insight into the phenomenon.

There are two main questions we want to answer in this work. Firstly, can the duration of exchange periods be appropriately modelled by a finite mixture model framework? Secondly, does the exchange duration differ depending on the gender or the age of each student? In order to answer these questions we define four different models. The first three models we propose model the duration by itself; they try to make a Bayesian model of the duration, without taking into account any other variables. The fourth and final model extends the former models, in order to try to account for covariates like *age*, *gender* and *receiving country*. In the end, we propose another model to investigate the phenomenon further. More specifically, we propose a model whose purpose is to investigate possible covariates, like the ones mentioned, that may affect the probability an individual faces of going on exchange for either one or two semesters, without pursuing any further calculations or discussion regarding the model.

2.1 Data Cleaning

The data was retrieved from the official portal for European data, on May 11, 2021. The Erasmus+ data in the portal contains a lot of information on all mobility periods started during the years from 2014 to

2019. We choose to focus on mobility periods started in 2019 and finalized before May 11, 2021. Prior to cleaning, the data has 736434 rows and 24 columns. Some examples of what these columns give information about are field of education, participant nationality, special needs (binary), participant age at the start of the mobility period, as well as the sending and receiving organization and country of each student. Moreover, the type of mobility is recorded in the data set, which consists of values like planning visits, voluntary service, job shadowing, staff mobility, traineeships and student mobility. This is a very large data set with a lot of information, which means that we have to make some restrictions concerning the data we want to analyze.

First of all, we decide to only take interest in university students which spend time abroad, as opposed to the other Erasmus+ mobility types. We remove most columns, keeping the columns shown in table 1. The table shows some summary statistics of the final data set. Only students aged between 15 and 80, which have a mobility duration between 20 and 365 days, are kept. Notice that this filtering leads to the youngest student in the data set being 17 years old and the oldest student being 73 years old. Other values, which do exist in the data set, do not make sense for the mobility period we seek to model. Additionally, we omit mobility periods with more than one participant, meaning that we only look at students that are recorded singularly. Moreover, because of our regression model, which will be presented last, we omit the undefined genders, which consist of 169 individuals. Finally, as there only are four missing values in the remaining data set, we remove the corresponding rows, instead of imputing the data. Note that the final data set has 211535 rows, with the given columns.

duration	age	gender	nationality	sending.country	receiving.country
Min. : 20	Min. :17.00	Female :126829	DE :31270	DE :31708	ES :31738
1st Qu.:127	1st Qu.:20.00	Male : 84706	ES :30841	ES :31394	IT :20447
Median :142	Median :22.00	Undefined: 0	IT :28321	IT :28563	FR :19466
Mean :168	Mean :22.13		FR :25527	FR :27020	DE :19149
3rd Qu.:174	3rd Qu.:23.00		TR :11662	TR :11919	UK :14394
Max. :365	Max. :73.00		PL : 8162	NL : 9413	PL :13605
			(Other):75752	(Other):71518	(Other):92736

Table 1: Summary statistics for the cleaned data set.

3 Theory on Finite Mixtures

To model the mobility durations we will consider a family of distributions called *mixture distributions*. They arise when the samples are drawn from underlying subgroups, each with different conditions. In the bimodal case with only two groups, this is written as

$$f(y_i|\theta, \mathbf{p}) = p_1 f(y_i|\theta_1) + p_2 f(y_i|\theta_2),$$

where θ_h is the parameter vector for subgroup $h \in \{1, 2\}$. The parameter p_h describes the proportion of the total population belonging to subgroup h. Note that in the bimodal case we can let $p_2 = 1 - p_1$ and only consider one parameter p_1. Notice that this setup can easily be generalized to a situation with more than two groups h, each subgroup having its own parameter p_h and $\sum_h p_h = 1$. However, for this problem, we restrict ourselves to the bimodal case, for reasons that will become clear.

If it is unknown which subgroup the sample is drawn from—as in our situation—we obtain a hierarchical structure by introducing a latent *indicator* variable z_{ih} which is equal to 1 if sample i belongs to subgroup h. The vector z_i, which has components z_{ih} and length h, follows a Multinomial distribution given $\mathbf{p}$. Thus, a hierarchical model may be creating by letting $\mathbf{p}$ follow a Dirichlet distribution—the conjugate prior of the Multinomial [2].

3.1 Modelling Mixtures with Stan

Stan is an open source statistical software for performing Bayesian modelling and inference. The Bayesian inference done in this work will be done using this software, interfaced via R. It uses *Markov chain Monte*

Carlo (MCMC) sampling to generate approximations to the posterior distributions. It does not support finite mixture models directly, since it cannot handle the discrete variables z_{ih} [3]. However it is possible to marginalize out the indicator variables using $\Pr(z_i = h) = p_h$. In the special case of bimodal normal mixtures, the model

$$f(y_i|p,\mu,\sigma) = pN(y_i|\mu_1,\sigma_1^2) + (1-p)N(y_i|\mu_2,\sigma_2^2), \quad p \in [0,1], \tag{1}$$

is fitted, after specifying priors for μ_1, μ_2, σ_1^2 and σ_2^2.

3.2 Identifiability

All finite mixture models are nonidentifiable, meaning that the distribution remains unchanged after permuting the labels of the subgroups [1] [2]. One clear example in the bimodal mixture model is seen when trying to identify which of the two components should be defined as component 1. This is something that will be taken into account when implementing the models in Stan; the important parameters will be defined in increasing order, to avoid this ambiguity. For example, in the bimodal normal mixture, this implies that the means are defined as

$$\mu_1 \in \mathbb{R},\ \mu_1 \leq \mu_2 \in \mathbb{R},$$

such that the means are constrained to following this order. A similar definition is done with the variances of the normal distributions, following

$$\sigma_1^2 \in \mathbb{R}^+,\ \sigma_1^2 \leq \sigma_2^2 \in \mathbb{R}^+.$$

3.3 Regression

It is straightforward to add predictors to the mixture model. Assume inclusion of the predictors $X_1, \ldots, X_m$ and define the observed sample $x^i = (y_i, x_{i1}, \ldots, x_{im})^T$, where the second sub-index indicates the predictor. The super-index is used in order to avoid confusion of a specific point i in the data set with the predictor X_j, $j \in \{1, \ldots, m\}$. A natural starting point is to let the group location parameter be a linear function of $X_1, \ldots, X_m$, i.e.

$$\mu_h|x^i = \beta_{0h} + x_{i1}\beta_{1h} + \ldots + x_{im}\beta_{mh}. \tag{2}$$

Prior distributions for the parameters $\beta_h = (\beta_{0h}, \beta_{1h}, \ldots, \beta_{mh})^T$ need to be specified—possibly group specific. In the bimodal normal mixture case, the model becomes

$$f(y_i|p,\beta,\sigma) = pN(y_i|X^i\beta_1,\sigma_1^2) + (1-p)N(y_i|X^i\beta_2,\sigma_2^2), \quad p \in [0,1], \tag{3}$$

where $X^i\beta_h = \beta_{0h} + x_{i1}\beta_{1h} + \ldots + x_{im}\beta_{mh}$. To increase the complexity of the model one could include interactions, higher order terms, other link functions or other regression techniques.

4 Analysis

4.1 Exploratory Data Analysis

Table 1 shows that the data contains significantly more females than men. A small subset of them are labeled as undefined, which were removed for ease of use in the regression model. The age of the students varies from 17 to 73, with most of them grouped around the median 22. Figure 1 shows the distribution of ages in the data, where the left plot suggests a right-skewed distribution and the right plot suggests that the male students are slightly older than the female students, with slightly less variability around the median.

Figure 2 displays where the students typically travel from and to. The majority of students are from Germany, Spain, Italy and France. At the receiving end we notice that Spain welcomes the largest number of students by quite some margin.

A histogram of the durations grouped by gender is shown in figure 3. The shape looks very similar for both genders, with one spike around 120 days and another around 280 days. This suggests that there might be an underlying grouping in the data, one group with students staying for three to six months and another with students staying for eight to eleven months. This is where the inspiration for the use of mixture models, especially the bimodal case, has mainly come from.

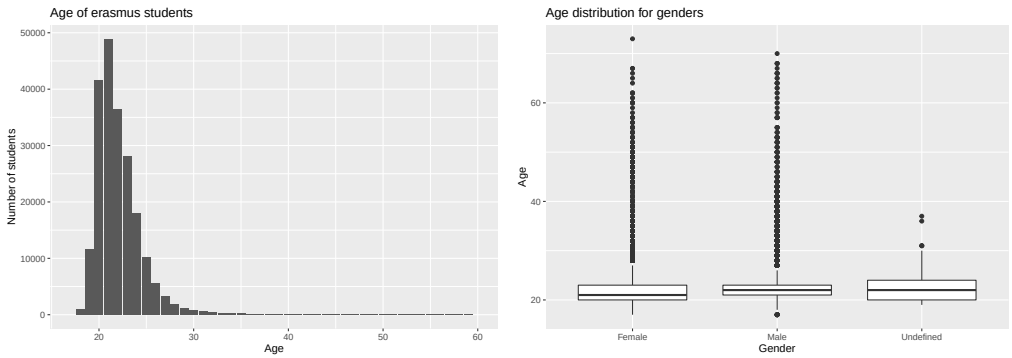


Figure 1: *Left:* Age distribution of students. *Right:* Box plots showing age of students grouped by gender.

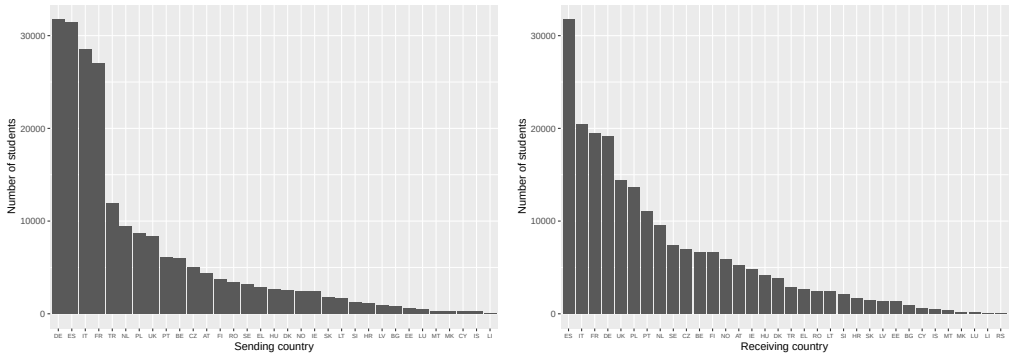


Figure 2: *Left:* Bar plot showing sending countries of Erasmus+ students. *Right:* Bar plot showing receiving countries of Erasmus+ students.

4.2 Models

The response variable that will be modelled is the duration, which is shown in figure 3. We will build our way up from a simple model to a more complex model, doing model checking and validation along the way, with the objective of finding the best possible model for the data. As mentioned several times, the analysis conducted is based on the bimodal normal mixture model, as described in equation (1).

Since the final data set is quite large, a randomly sample set of 15000 points is used when running calculations in Stan, in order for numerical feasibility on consumer laptops. Some summary statistics of the sampled data set are given in table 2. Notice that the quantiles in the continuous variables are similar to the ones seen in table 1, which indicates that the sample is representative of the large data set. Moreover, the orders of the nationalities and countries are intact, indicating the representativeness.

duration	age	gender	nationality	sending.country	receiving.country
Min. : 20.0	Min. :17.00	Female :8945	DE :2236	DE :2271	ES :2260
1st Qu.:127.0	1st Qu.:20.00	Male :6055	ES :2173	ES :2214	IT :1432
Median :142.0	Median :22.00	Undefined: 0	IT :2045	IT :2054	FR :1409
Mean :168.2	Mean :22.11		FR :1790	FR :1910	DE :1372
3rd Qu.:174.0	3rd Qu.:23.00		TR : 799	TR : 830	PL : 970
Max. :364.0	Max. :61.00		PL : 607	NL : 669	UK : 968
			(Other):5350	(Other):5052	(Other):6589

Table 2: Summary statistics for 15000 randomly sampled data points from cleaned data set.

As already noted, in all the following models, an ordering of the means is enforced, such that $\mu_1 < \mu_2$. Moreover, in the three models where the variance is allowed to differ between the groups, the variance

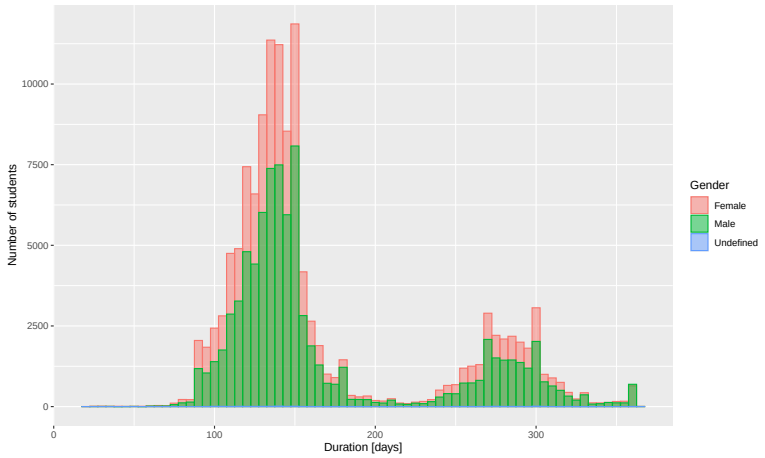


Figure 3: Histogram of number of students per duration of mobility period, grouped by gender.

parameter ordering is also enforced.

4.2.1 Model 1 - Bimodal Gaussian with Equal Variance

The simplest proposed model assumes identical variance σ^2 in both normal components. The Bayesian model consists of a bimodal normal mixture model with $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and with prior distributions

$$\begin{aligned} \mu_1 &\sim N(120, \sigma^2), \\ \mu_2 &\sim N(280, \sigma^2), \\ \sigma &\sim InverseGamma(10, 100). \\ p &\sim U[0, 1]. \end{aligned}$$

The prior for σ is inspired by conjugate priors – the conjugate prior for σ^2 for the normal distribution with known mean is the inverse gamma. However, notice that we choose the inverse gamma as a prior for the standard deviation σ, not the variance σ^2, simply because it was easiest to implement in Stan. If desired, this could easily be changed in order to work with conjugate distributions, but we are not bothered by this, since we use Stan to calculate approximations iteratively. We select a somewhat non-informative version, choosing the parameters (10,100). As for the group means, we use our prior knowledge that they should peak around 120 days and 280 days respectively. In order to visualize the information we capture with these priors, we plot the prior predictive distributions of σ, μ_1, μ_2 and the prior predictive distribution in figure 4. Observe that the prior predictive distribution has two peaks, one at each of the chosen means. Notice that the peaks essentially are symmetric around 200 days—none of them are preferred with the prior information we are defining in the model. The reason behind this is the uniform prior on the mixing parameter, or probability, p. Finally, the prior predictive distribution shows a relatively large variance, which is accordance with our relatively limited prior knowledge of the phenomenon.

Convergence results are shown in table 3. All R-hat values are equal to 1. Notice the variable y_{pred}. This is a generated quantity defined in the Stan model—it generates predictions from the posterior predictive distribution. In this manner, we use Stan directly to generate values from the posterior predictive distribution, instead of simulating them in R using the drawn samples of the posterior distribution. The statistics shown for the simulations from the posterior predictive distribution will be used for model checking. The traceplot is shown on the left in figure 5. Qualitatively, it is apparent that the chains are converging.

The final Gaussian mixture, i.e. the posterior predictive distribution, produced by this model is shown on the right in figure 5. We can see that the model has estimated a shape that resembles the shape of the data seen in figure 3, which is a promising result. The posterior distributions of the other parameters are shown in figure 6.

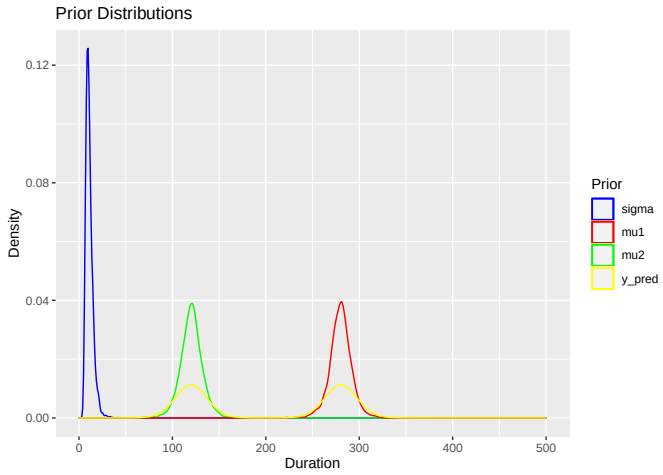


Figure 4: Prior predictive distribution for μ_1 and μ_2, shown in red and green respectively. The prior predictive distribution is shown in yellow.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
mu1	135.19	0.00	0.21	134.77	135.05	135.19	135.32	135.60	3456.47	1.00
mu2	286.05	0.01	0.40	285.27	285.78	286.05	286.32	286.83	4288.90	1.00
sigma	22.39	0.00	0.13	22.13	22.30	22.39	22.48	22.64	4407.77	1.00
p	0.78	0.00	0.00	0.77	0.78	0.78	0.78	0.79	1431.32	1.00
y_pred	168.51	1.05	66.46	94.99	124.96	143.57	174.27	313.26	4016.93	1.00
lp__	-75783.30	0.04	1.45	-75786.95	-75784.01	-75782.98	-75782.23	-75781.48	1610.79	1.00

Table 3: Convergence results for model 1 with 15000 sample data points. The fitting has been run for 2000 iterations per chain.

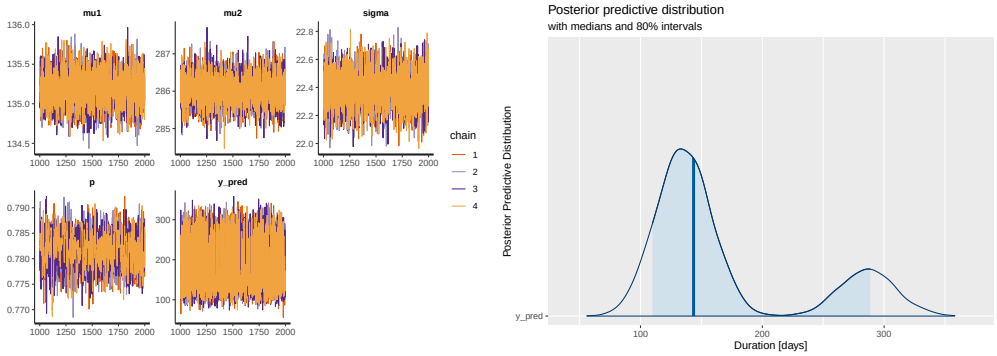
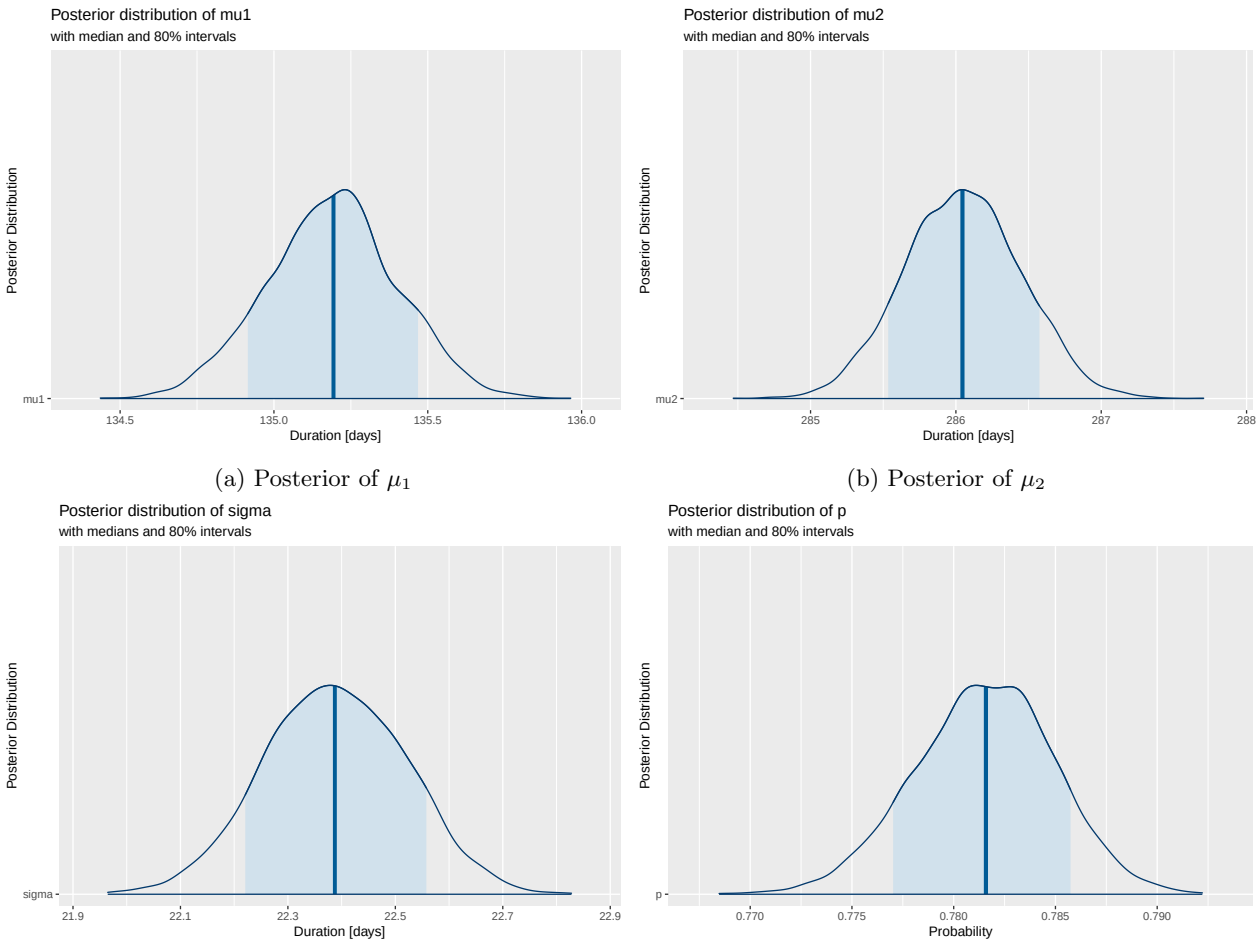


Figure 5: *Left:* Traceplot from fitted model 1 on a sample of 15000 data points, using Stan via R. *Right:* Posterior predictive distribution from model 1. This is a kernel density estimation of the generated quantities y_{pred}. The median is indicated with a vertical line and 80 % probability mass is shaded in blue.



(a) Posterior of μ_1

(b) Posterior of μ_2

(c) Posterior of σ

(d) Posterior of p

Figure 6: Posterior distributions of parameters following estimation of model 1.

Some model checking is essential in order to see if the model is adequately modelling the data and to eventually compare different models to each other. We select some ad-hoc statistics for our problem. The statistics should summarize the data and focus on aspects relevant to our objective with the modelling. We choose the 1st quartile, the median, the mean and the 3rd quartile. Using the posterior predictive distribution, we simulate values of these statistics, in order to approximate their respective reference distributions. Finally, we compare the statistics in our data set, which in this case is the 15000 random point-sample, with the simulated reference distribution given our model.

Before approximating the reference distributions of the statistics, comparing the summary statistics for duration shown in table 2 with the summary statistics for y_{pred} shown in table 3 indicates that the model has modelled the data relatively well—since the first quartile, the mean, the median and the third quartile are quite similar. In order to quantify this further, we perform the calculations described above. The approximated reference distributions of the ad-hoc statistics are shown in figure 7. We can see that some of the vertical lines, which indicate the statistics in the data, are relatively well-placed around the reference distributions, e.g. for the third quartile and the mean, indicating that the model approximates these statistics in the data relatively well. This is not the case for the first quartile and the median, where the reference distributions seem to be centered far away from the respective statistics in the data and have little probability density at these values.

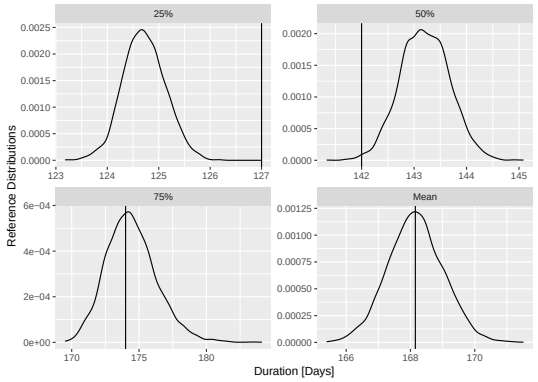


Figure 7: Approximated reference distributions of the selected ad-hoc statistics (first quartile, median, mean and third quartile) from model 1. The vertical lines indicate the respective statistics in the data set used to fit the model.

In addition to showing the reference distributions graphically, we quantify the difference between the statistics simulated from the model and the statistics in the data. This is done by calculating the following quantity for each of the statistics

$$\min\{\Pr(T(\bar{y}) \leq T(y)|y), \Pr(T(\bar{y}) \geq T(y)|y)\}$$

where $\Pr(T(\bar{y})|y)$ denotes the approximate ad-hoc statistic reference distribution and $T(y)$ denotes the respective statistic in the data. Essentially, we are calculating the area of each of the tails, i.e. the area under the distribution to the left of the value in the data and the area to the right of the value, and calculating minimum. A value around 0.5 indicates, as discussed graphically, that the reference distribution is well-placed compared to the statistic in the data. The numerical calculations are showed in table 4. More specifically, it shows the area of each of the tails of each of the reference distributions, denoted by "Left" and "Right" respectively, with the minimum in the third row. As earlier, we conclude that the mean and third quartile seem to be modelled relatively well by the model, while the performance regarding the two other statistics is lacking.

	25%	Mean	50%	75%
Left	1.00000	0.50500	0.00500	0.42975
Right	0.00000	0.49500	0.99500	0.57025
min	0.00000	0.49500	0.00500	0.42975

Table 4: Numerical calculations for model checking of model 1. The areas of the left tail and right tail of the approximated reference distributions, along with their minimum, are given for each ad-hoc statistic.

4.2.2 Model 2 - Bimodal Gaussian with Different Variances

The previous model is extended to a model where the two normal components have distinct variances, denoted by σ_1^2 and σ_2^2. The Bayesian model takes the same general form as earlier, with prior distributions

$$\begin{aligned}
\mu_1 &\sim N(120, \sigma_1^2), \\
\mu_2 &\sim N(280, \sigma_2^2), \\
\sigma_1 &\sim InverseGamma(10, 100), \\
\sigma_2 &\sim InverseGamma(10, 100), \\
p &\sim U[0, 1].
\end{aligned}$$

Convergence results are shown in table 5. The chains have converged according to the R-hat values and there is a significant number of efficient sampled points from each posterior distribution. The traceplot is shown on the left in figure 8, which indicates converging chains.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
mu1	134.82	0.00	0.19	134.45	134.70	134.82	134.94	135.18	4371.11	1.00
mu2	284.60	0.01	0.52	283.56	284.24	284.60	284.95	285.61	4993.11	1.00
sigma1	20.17	0.00	0.14	19.91	20.08	20.17	20.26	20.44	4187.87	1.00
sigma2	29.12	0.01	0.39	28.35	28.85	29.12	29.39	29.89	4623.24	1.00
p	0.78	0.00	0.00	0.77	0.78	0.78	0.78	0.78	4241.79	1.00
y_pred	169.30	1.10	66.88	97.56	125.90	142.69	173.38	320.28	3721.40	1.00
lp__	-75537.81	0.03	1.54	-75541.73	-75538.67	-75537.48	-75536.65	-75535.74	2241.94	1.00

Table 5: Convergence results for model 2 with 15000 sample data points. The model fitting has been run for 2000 iterations per chain.

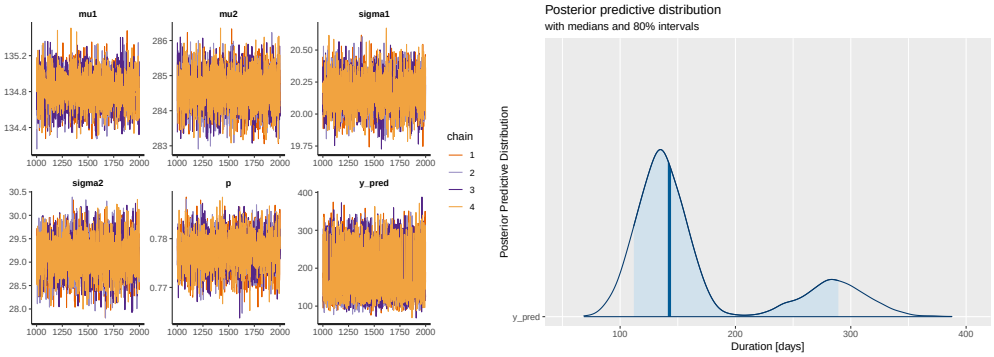
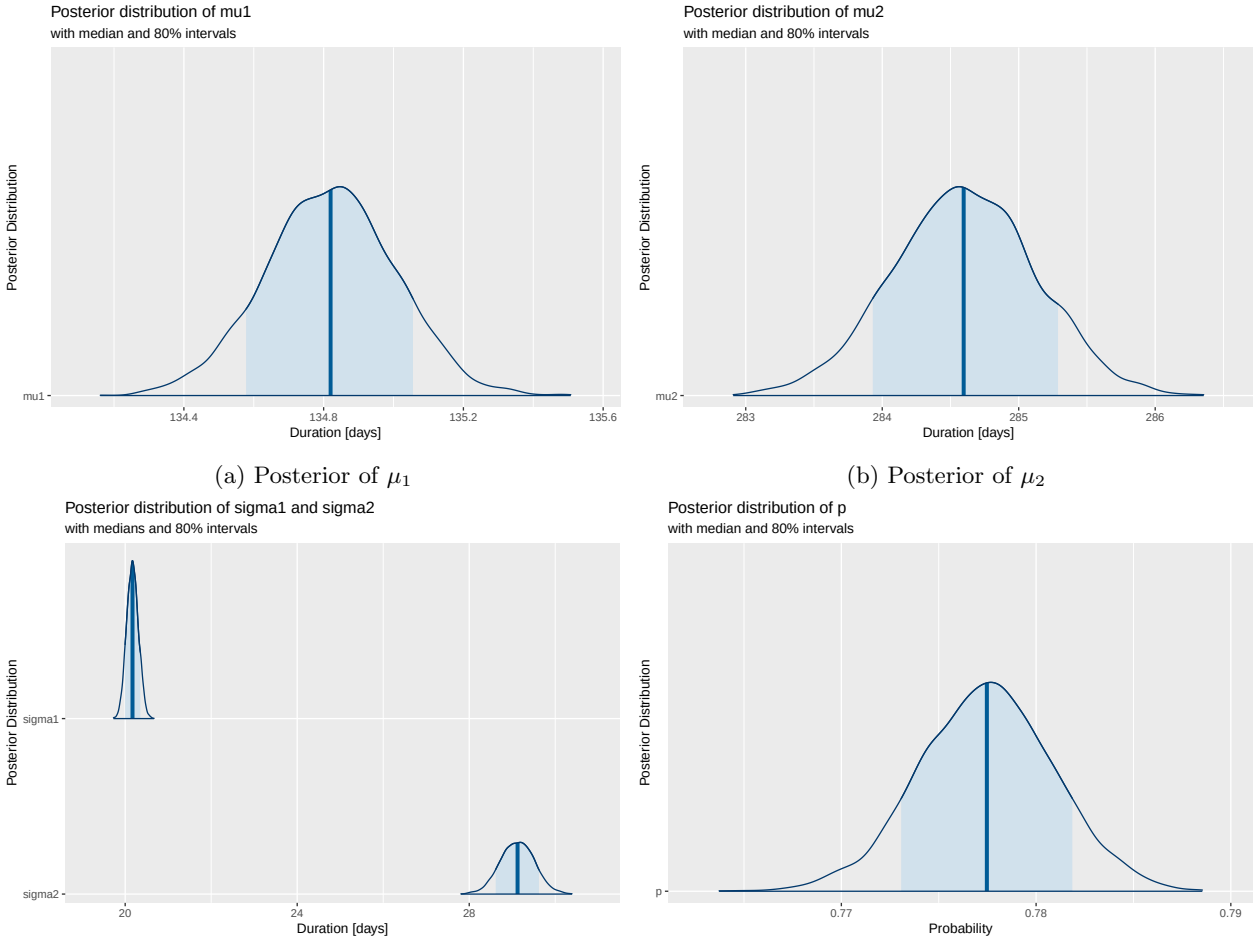


Figure 8: *Left:* Traceplot from fitted model 2 on a sample of 15000 data points, using Stan via R. *Right:* Posterior predictive distribution from model 2. This is a kernel density estimation of the generated quantities y_{pred}. The median is indicated with a vertical line and 80 % probability mass is shaded in blue.

The posterior predictive distribution is shown on the right in figure 8. The shape of it resembles our duration data. Other posterior distributions are shown in figure 9. Notice that the posterior distributions of μ_1, μ_2 and p are very similar to the distributions seen following model 1, with some slight shifts in the median value. Moreover, we notice that the dispersion seems to be slightly larger in the posterior of μ_2 compared to its counterpart in model 1. Similarly, the dispersion seems to be slightly smaller in the posterior of μ_1 compared to in model 1. The reason behind these changes is that the variances of the normal mixture components are given freedom to be different—in the posterior distribution of σ_1 and σ_2 we can see that the values are larger for σ_2 and smaller for σ_1, when comparing to the equal variance σ for the two components in model 1.

A similar model check as in model 1 is repeated for model 2. First of all, we compare the statistics of the generated quantities y_{pred}, as shown in table 5, to the respective statistics in the data set, as shown in table 2. As concluded for model 1, these statistics are very similar, a first indication that the model has done a good job. A thorough comparison between the statistics from each of the models with the data is done later.

The reference distributions of the statistics are calculated and plotted in figure 10. To the naked eye it seems like all statistics are relatively close in the two data sets, except for the first quartile. The reference distribution of the median is clearly closer to the true value, compared to the results we saw in model 1, while the reference distribution of the third quartile has deteriorated compared to in model 1. These conclusions are quantified with numerical calculations shown in table 6. It becomes apparent that the compatibility of the mean and the median between the model and the data is relatively good, while it has deteriorated in the third quantile, when comparing to the performance obtained in model 1.



(a) Posterior of μ_1

(b) Posterior of μ_2

(c) Posterior of σ_1 and σ_2

(d) Posterior of p

Figure 9: Posterior distributions of parameters following estimation of model 2.

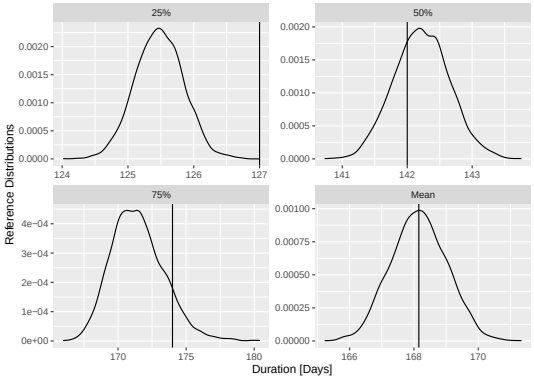


Figure 10: Approximated reference distributions of the selected ad-hoc statistics (first quartile, median, mean and third quartile) from model 2. The vertical lines indicate the respective statistics in the data set used to fit the model.

	25%	Mean	50%	75%
Left	1.00000	0.50150	0.30200	0.90250
Right	0.00000	0.49850	0.69800	0.09750
min	0.00000	0.49850	0.30200	0.09750

Table 6: Numerical calculations for model checking of model 2. The areas of the left tail and right tail of the approximated reference distributions, along with their minimum, are given for each ad-hoc statistic.

4.2.3 Model 3 - Hierarchical Bimodal Gaussian

Hierarchical models in Bayesian statistics can be taken advantage of to model dependencies in the model at different levels. In our case it might be advantageous to include some dependency between the σ's of the two mixture components. In earlier models we have indirectly made assumptions about the relationship between the two parameters, either assuming they are equal or assuming they are entirely separately distributed. By adding an additional hierarchical layer we can allow them to be different while still including a prior belief that they are connected in some way. A connection between the two seems reasonable, because the variance in each normal mixture component might well be controlled by a similar phenomenon. Mathematically this new Bayesian model follows the same form as usual, with prior distributions on the parameters

$$\begin{aligned}
\mu_1 &\sim N(120, \sigma_1^2),\\
\mu_2 &\sim N(280, \sigma_2^2),\\
\sigma_1 &\sim InverseGamma(\nu, \gamma),\\
\sigma_2 &\sim InverseGamma(\nu, \gamma),\\
\nu &\sim N(0, 10),\\
\gamma &\sim InverseGamma(10, 100),\\
p &\sim U[0, 1].
\end{aligned}$$

The hyperpriors for ν and γ are chosen to be somewhat non-informative, with the former following a normal distribution with large variance and the latter following an inverse gamma distribution with large variance, as it seems like a good distribution for modelling standard deviations.

Table 7 gives convergence results and the traceplot is shown on the left in figure 11. Both show that the chains are converging.

The right plot in figure 11 shows the posterior predictive distribution, which has the expected shape; similar to the shape seen in the duration data and in the earlier models. The posterior distributions of parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$ and p are shown in figure 13. These distributions look almost identical to the distributions calculated from model 2. The reader is referred to the discussion under model 2 for more details. The posteriors of the two new parameters ν and γ are shown in figure 12. From these parameter distributions it is possible to learn something about the connection between σ_1 and σ_2, parameterized by an inverse gamma distribution with the two parameters acting as shape and rate parameters respectively.

Model checking is performed similarly to in the previous models. As earlier, comparing the statistics of the generated quantities $\mathrm{y}_{\mathrm{pred}}$, as shown in table 7, to the respective statistics in the data set, as shown in table 2, leads to the conclusion of highly similar statistics. Moreover, the reference distributions of the statistics are calculated. They are plotted in figure 14. In addition, the numerical calculations are given in table 8. Notice that both the graphs and the numerical results are almost identical to those of model 2, which makes sense because of the very similar posterior distributions of the parameters. Thus, the conclusions are the same as for model 2. The advantage of this model is thus that one can learn something about the connection between the two parameters for the standard deviations, with a similar performance as the simpler model 2. The downside of the model is that it is more complicated.

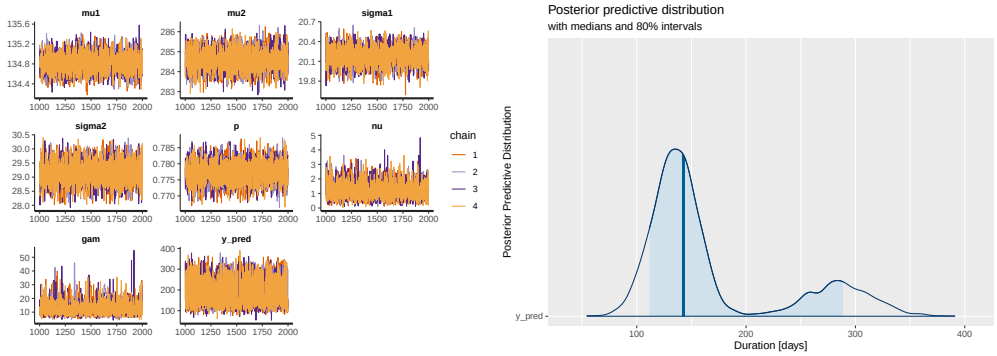


Figure 11: *Left:* Traceplot from fitted model 3 on a sample of 15000 data points, using Stan via R. *Right:* Posterior predictive distribution from model 3. This is a kernel density estimation of the generated quantities y_{pred}. The median is indicated with a vertical line and 80 % probability mass is shaded in blue.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
mu1	134.81	0.00	0.19	134.45	134.69	134.81	134.94	135.19	5053.29	1.00
mu2	284.59	0.01	0.52	283.57	284.24	284.60	284.94	285.61	5402.58	1.00
sigma1	20.18	0.00	0.14	19.91	20.09	20.17	20.27	20.44	4844.19	1.00
sigma2	29.15	0.01	0.39	28.39	28.89	29.14	29.40	29.93	5441.09	1.00
p	0.78	0.00	0.00	0.77	0.78	0.78	0.78	0.78	5045.01	1.00
nu	1.18	0.01	0.60	0.30	0.74	1.08	1.53	2.57	1804.39	1.00
gam	12.91	0.16	5.18	6.27	9.40	11.82	15.10	26.06	1043.16	1.00
y_pred	168.29	1.04	66.79	96.76	125.78	142.64	170.26	322.53	4089.20	1.00
lp__	-75502.24	0.05	1.90	-75506.68	-75503.35	-75501.90	-75500.79	-75499.55	1542.54	1.00

Table 7: Convergence results for model 3 with 15000 sample data points. The model fitting has been run for 2000 iterations per chain.

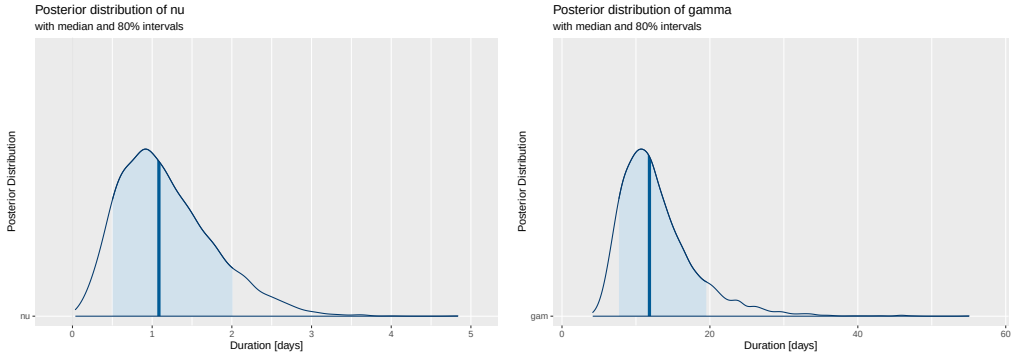
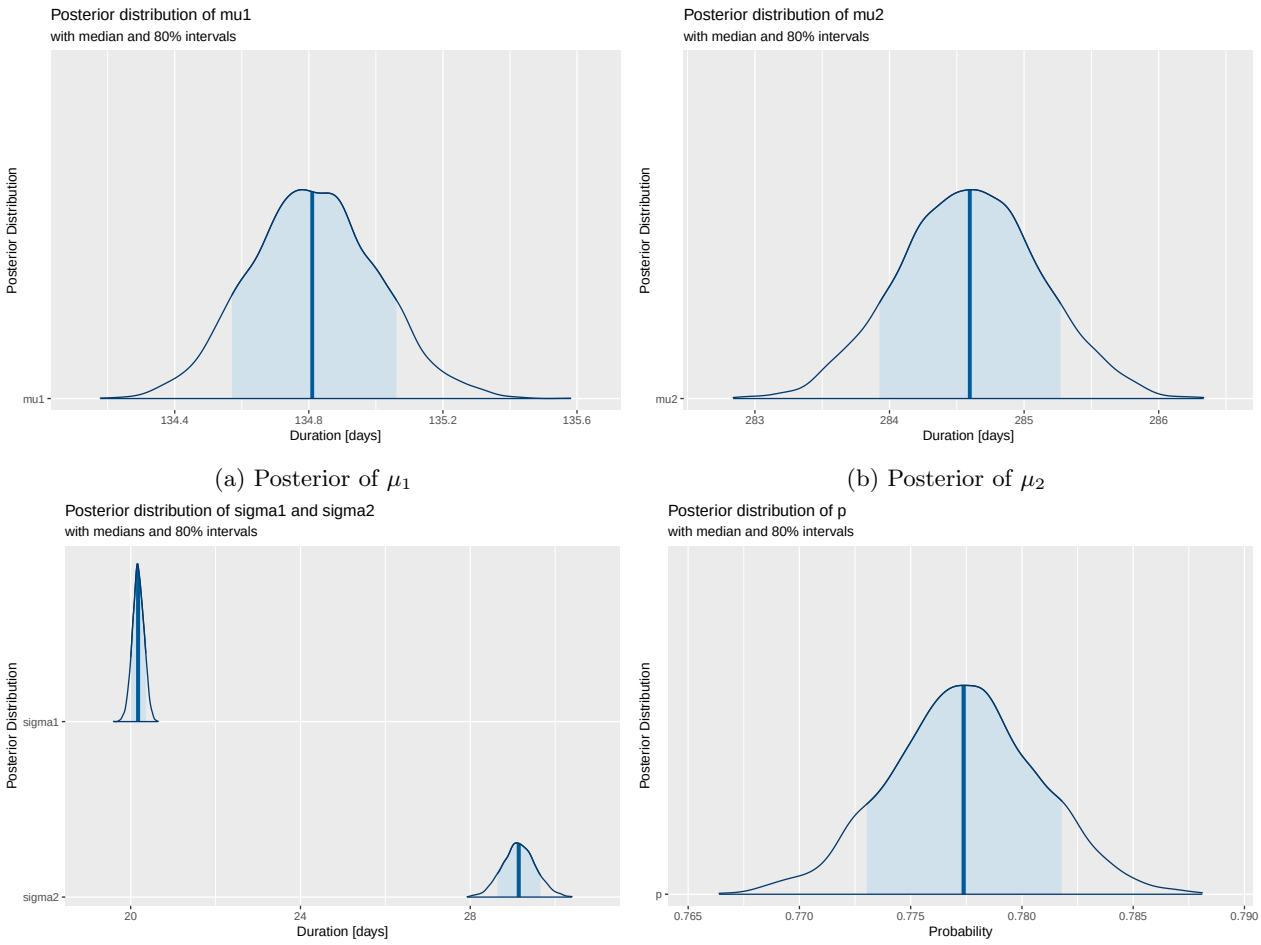


Figure 12: *Left:* Posterior distribution of ν. *Right:* Posterior distribution of γ.



(a) Posterior of μ_1

(b) Posterior of μ_2

(c) Posterior of σ_1 and σ_2

(d) Posterior of p

Figure 13: Posterior distributions of parameters following estimation of model 3.

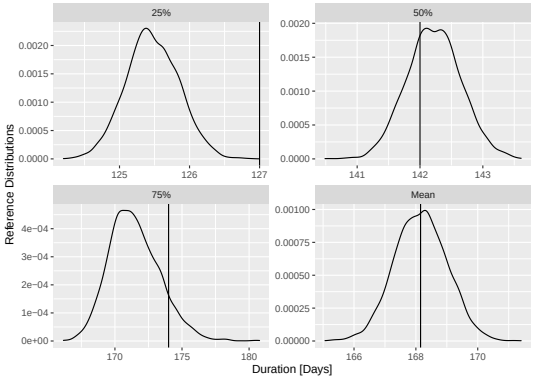


Figure 14: Approximated reference distributions of the selected ad-hoc statistics (first quartile, median, mean and third quartile) from model 3. The vertical lines indicate the respective statistics in the data set used to fit the model.

	25%	Mean	50%	75%
1	1.00000	0.50025	0.30850	0.90775
2	0.00000	0.49975	0.69150	0.09225
3	0.00000	0.49975	0.30850	0.09225

Table 8: Numerical calculations for model checking of model 3. The area of the left tail and right tail of the approximated reference distributions, along with their minimum, is given for each ad-hoc statistic.

4.3 Model 4 - Hierarchical model with Covariates

Model 3 is extended to include predictors. Some possible columns in the data immediately stand out as possible covariates; the *age* and *gender*. Notice that age is a continuous variable, while gender is binary (after removing the undefined genders in the data set).

Let $X_1 \in \mathbb{R}^+$ be the student's age and $X_2 \in \{0, 1\}$ be the gender, where $X_2 = 0$ means the student is female. The conditional means are modelled as in equation (2), which leads to a model on the form shown in equation (3) with $m = 2$. For simplicity we let the the coefficients for age and gender be the same for both groups—the coefficients $\beta_{01}, \beta_{02}, \beta_1$ and β_2 are introduced. The two first parameters are used as intercepts in the linear conditional means in each of the two normal mixture components—these represent the means in the mixture components for age zero and female gender. The latter two parameters are used to describe unit change in the mean with age and change in the mean for males compared to females respectively. Mathematically, the model is formulated as

$$
\begin{aligned}
y|x &\sim pN(y|\beta_{01} + \beta_1 x_1 + \beta_2 x_2, \sigma_1^2) + (1-p)N(y|\beta_{02} + \beta_1 x_1 + \beta_2 x_2, \sigma_2^2),\\
\beta_{01} &\sim N(120, \sigma_1^2),\\
\beta_{02} &\sim N(280, \sigma_2^2),\\
\beta_1 &\sim N(0, 100),\\
\beta_2 &\sim N(0, 100),\\
\sigma_1 &\sim InverseGamma(10, 100),\\
\sigma_2 &\sim InverseGamma(10, 100),\\
p &\sim U[0, 1],
\end{aligned}
$$

where the prior distribution for each parameter is given. Normal distributions with large variance are chosen as prior distributions for the covariates β_1 and β_2, reflecting non-informative priors in practice. The convergence results are shown in table 9. All R-hat values are 1 or very close to 1, showing that the chains are converging. The traceplot is shown to the left in figure 15, which yields the same conclusion about convergence.

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
sigma1	20.17	0.00	0.14	19.88	20.07	20.17	20.27	20.46	1551.28	1.00
sigma2	28.96	0.01	0.39	28.16	28.70	28.97	29.22	29.73	1338.16	1.00
p	0.78	0.00	0.00	0.77	0.78	0.78	0.78	0.78	1202.46	1.00
beta01	125.80	0.06	1.54	122.83	124.75	125.80	126.84	128.85	731.21	1.01
beta02	275.83	0.06	1.58	272.81	274.83	275.80	276.92	278.84	815.69	1.00
beta1	0.39	0.00	0.07	0.25	0.34	0.39	0.44	0.53	733.53	1.01
beta2	0.94	0.01	0.37	0.20	0.69	0.95	1.19	1.66	1305.52	1.00
y_pred	170.10	1.53	66.46	98.04	128.13	145.38	171.87	322.64	1881.11	1.00
lp__	-75518.95	0.07	1.93	-75523.50	-75520.01	-75518.60	-75517.51	-75516.26	735.64	1.01

Table 9: Convergence results for model 4 with 15000 sample data points. The model fitting has been run for 1000 iterations per chain.

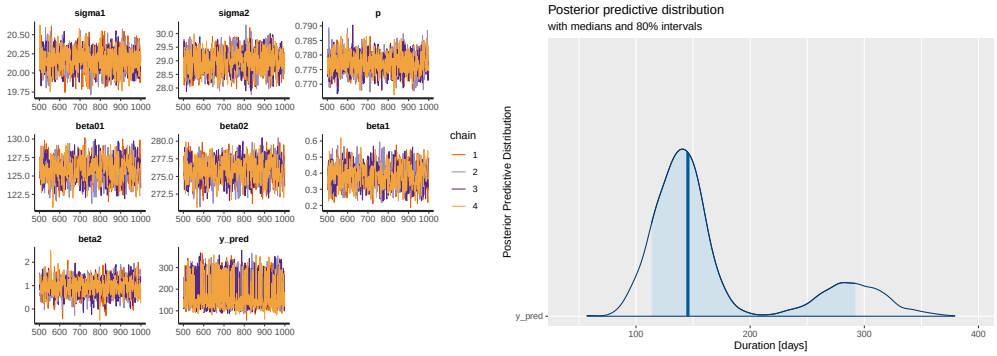


Figure 15: *Left:* Traceplot from fitted model 4 on a sample of 15000 data points, using Stan via R. *Right:* Posterior predictive distribution from model 4. This is a kernel density estimation of the generated quantities y_{pred}. The median is indicated with a vertical line and 80 % probability mass is shaded in blue.

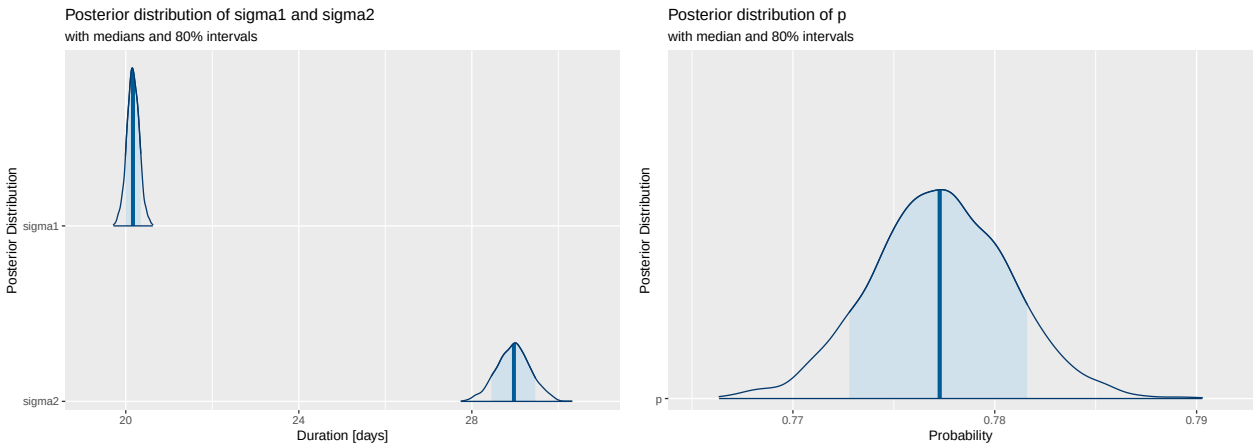
The posterior predictive distribution is shown on the right in figure 15. The shape of the distribution is very similar to the shapes seen in both model 2 and model 3. Posterior distributions of the parameters are shown in figure 16. The distributions of σ_1 and σ_2 look almost identical to the distributions seen in model 2 and model 3, which helps to explain why the shapes of the posterior predictive distributions are similar. The same can be said about the posterior distributions of p. The posteriors for β_{01} and β_{02} show that the intercepts are close to where we would guess them to be a priori. Notice that we do not plot posteriors for μ_1 and μ_2. Because of limitations with vectorization of mixture-models in Stan, we were not able to produce these posteriors. Additionally, because we did not find a work-around for this limitation, we were not able to approximate the reference distributions, as done for earlier models. However, we are still able to compare the statistics of the posterior predictive distribution from model 4 with the statistics in the data. Comparing the statistics of y_{pred} in table 9 with the values in table 2, we again conclude that they are relatively close, showing that the data seems to be modelled well based on these ad-hoc statistics.

What can we learn from the added covariates in the model? Does the age or the gender affect the mean duration of the exchange? Notice that the 95% credible interval for each of the covariates β_1 and β_2, as shown in table 9, does not contain zero for either covariate. This implies that both age and gender has a statistically significant effect on the duration. Additionally, the posterior distributions of the parameters show that the standard deviation of β_2 is large compared to the standard deviation of β_1. Using, for example, the mean values of their respective posterior distributions as estimates for their effect, we learn that the unit increase in the duration of the exchange with age is 0.39 and the increase in the duration of the exchange for males compared to females is 0.94.

5 Comparison of Models

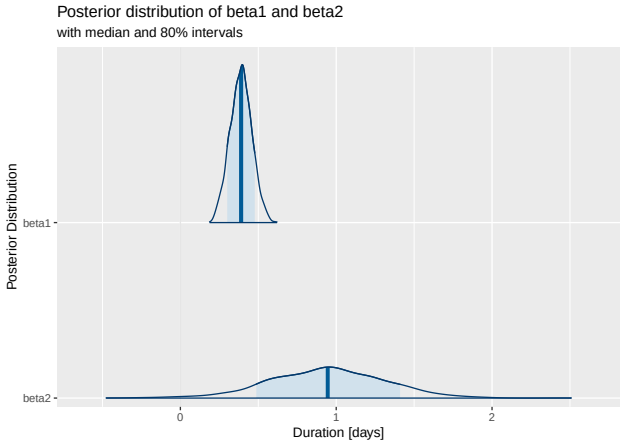
This section does a closer comparison between all the ad-hoc statistics from each of the models with the statistics in the data. The comparisons are done in tables 10 and 11. The former table simply aggregates all the information we have discussed earlier, such that all the values can be seen simultaneously at a glance. The latter table calculates the relative errors between each of the statistics following each of the models and the value of the statistics in the data. The relative errors are presented in percentages. The final row shows which of the models has the lowest absolute relative error.

A general observation that can be made from the tables is that the models are not very far apart, when comparing these statistics. Notice that the largest relative error is approximately 2.4 %, meaning that all the statistics for all the models are at least within 2.4 % of the value in the data. This is a noteworthy result, since this is an error that should be regarded as small enough for many purposes. Despite the fact that all the models are relatively similar when it comes to these values, model 3 gets the majority vote as the best model, as it has the lowest relative error in half of the statistics. However, as always when comparing models, there are several other factors to take into account. The first models are very simple, which gives them a

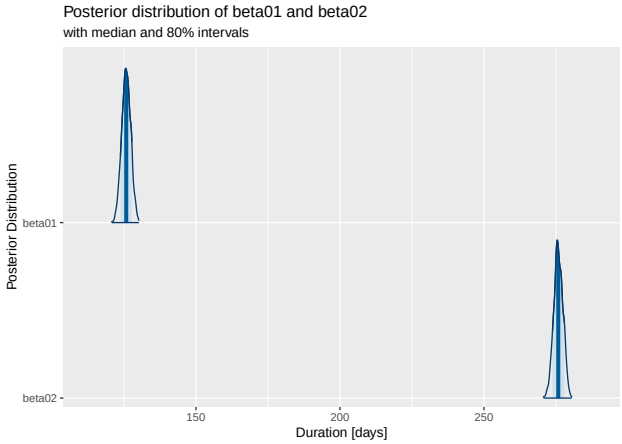


(a) Posterior of σ_1 and σ_2

(b) Posterior of p



(c) Posterior of β_1 and β_2



(d) Posterior of β_{01} and β_{02}

Figure 16: Posterior distributions of parameters following estimation of model 4.

computational advantage over the latter models. Moreover, they can yield easier interpretability, since they are more parsimonious and contain less "moving parts". Model 3 is perhaps more realistic, which perhaps explains why it performs slightly better than model 2 (and the other models) in several of the statistics. Despite the relatively simplistic nature of the first three models, they lack inclusion of covariates to explain the phenomenon. Model 4 gives us the ability to investigate how some covariates affect the exchange duration. In this regard, model 4 has some clear advantages when wanting to investigate the phenomenon more deeply. However, note that the restrictions imposed on model 4 are relatively strict here, which restricts its usability. Thus, the conclusion when it comes to which model is the best is that it depends on the objective of the researcher or the practitioner. If the objective is building a simple predictive model of the data that is quickly calculated, perhaps one of the first two models are good choices. If the objective is building a model for inference, perhaps one of the more complicated models are good starting points.

Notice that the choice of statistics used to compare the posterior predictive distributions with the training data is ad-hoc and somewhat arbitrary. In order to compare the models more thoroughly, one could perhaps have chosen more statistics, like for example comparing standard errors, kurtosis or skewness in the data. The choice of other statistics might lead to different conclusions.

	25%	50%	75%	Mean
Data	127.00000	142.00000	174.00000	168.15153
Model 1	124.95689	143.57221	174.26832	168.50634
Model 2	125.90181	142.68912	173.38282	169.30300
Model 3	125.77898	142.63605	170.26094	168.29177
Model 4	128.13476	145.38458	171.87227	170.09856

Table 10: Statistics in data and calculated from all proposed models.

	25%	50%	75%	Mean
Data	127.00000	142.00000	174.00000	168.15153
Model 1	-1.60875 %	1.10719 %	0.15421 %	0.21100 %
Model 2	-0.86472 %	0.48530 %	-0.35470 %	0.68478 %
Model 3	-0.96143 %	0.44792 %	-2.14888 %	0.08340 %
Model 4	0.89351 %	2.38350 %	-1.22283 %	1.15790 %
min Model	2	3	1	3

Table 11: Relative errors compared to statistics in data (%). The first row shows the statistics in the data and the last row shows the model number that yields the lowest relative (absolute) error.

6 Conclusion

The work shows that the duration of Erasmus+ exchange for students can be appropriately modelled by a finite mixture model framework. Four different models have been implemented, ranging for simple to more complex, which are able to model the data to a sufficient degree. The exchange duration seems to depend on the gender and the age of each student—females and younger students seem to go for slightly shorter durations when comparing to males and older students.

7 Further Work

Several extensions can be done to this work. First of all, model 4 could be modified and extended, by removing some restrictions on the model. The coefficients for age and gender need not be the same for both groups. In addition, model 4 could be extended to include more covariates. We only modelled it using age and gender, but values like receiving country and sending country could also be interesting to investigate in the model. Other changes to the model can be made, like adding interactions between the covariates, adding higher order terms for the continuous variable(s), changing the link function or even kernelizing it.

As already noted, model checking could be done with other statistics and with other methods, like prediction on test data. This would most likely yield more certain conclusions when it comes to what model is best for what purpose.

A question we did not touch upon is if there are any factors that contribute to a smaller or larger probability of going on exchange during either one or two semesters. In the four proposed models we were able to estimate the proportion of people going on exchange for one semester, for example by calculating the mean of the posterior distribution of p (0.78 in all four models). In other words we could quantify the probability of a person belonging to the first cluster of people, which typically go on exchange for one semester, or the second cluster of people, which typically go on exchange for two semesters. However, these models to not enable investigation of what factors contribute to this probability. Perhaps factors like age, receiving country or sending country would significantly contribute to a change in probability. A model for this purpose is proposed below.

We copy the model definition from model 3, to have a starting point.

$$
\begin{aligned}
\mu_1 &\sim N(120, \sigma_1^2), \\
\mu_2 &\sim N(280, \sigma_2^2), \\
\sigma_1 &\sim InverseGamma(\nu, \gamma), \\
\sigma_2 &\sim InverseGamma(\nu, \gamma), \\
\nu &\sim N(0, 10), \\
\gamma &\sim InverseGamma(10, 100), \\
\text{logit}(p) &\sim \Gamma X, \\
\Gamma &\sim N(0, 100I).
\end{aligned}
$$

Instead of setting a uniform prior for p, we link it to a linear predictor using the logit function, similarly to how it was done for the conditional mean of each of the normal mixture components in model 4. This can be viewed as a logistic regression on the parameter p. In this case, Γ represents a vector of parameters that should be estimated by the model. They should be given prior distributions as well, which in this case is a multivariate normal distribution of the correct dimensions, with a large variance. Again, notice the similarity with model 4. The covariate matrix X contains the values of the covariates used. Note that this addition can be made to all the models we have proposed in this work or added to new models that have not been investigated here. Lastly, since we had some slight identifiability issues for previous models, we assume that this model might lead to similar problems.

References

[1] Michael Betancourt. *Identifying Bayesian Mixture Models*. 2017. URL: https://mc-stan.org/users/documentation/case-studies/identifying_mixture_models.html.

[2] A. Gelman et al. *Bayesian Data Analysis (2nd ed.)* Chapman and Hall/CRC, 2003. ISBN: 9780429258480.

[3] Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual, 2.29*. 2019. URL: https://mc-stan.org.