School of Mathematical and Statistical Sciences
Arizona State University

Student: Benjamin Muhlmann

Other students: Wilmer O. Martinez, Jennifer Rastegar, Emily Gilmore

Course: STP 540

Program: Statistics MS

Instructor: Robert McCulloch

Date: Fall 2019

STP 540 FINAL PROJECT
MIXTURE DISTRIBUTIONS USING EM AND GIBBS SAMPLING ALGORITHMS

Wilmer O. Martinez
Department of Mathematics and Statistics
Arizona State University, Tempe
womartin@asu.edu

Jennifer Rastegar
Department of Mathematics and Statistics
Arizona State University, Tempe
jrastega@asu.edu

Benjamin Muhlmann
Department of Mathematics and Statistics
Arizona State University, Tempe
bmuhlman@asu.edu

Emily Gilmore
Ira A. Fulton Schools of Engineering
Arizona State University, Tempe
eagilmo1@asu.edu

December 10, 2019

ABSTRACT

Mixture distributions are measurements of random variables. Using mixture distributions, we are able to take a variety of available data and draw conclusions by conditioning these measures to construct such probability models[1]. The data we sampled concerns houses in California and the various parameters defining those houses; i.e. room number, latitude, longitude, etc. We discovered that many of the parameters were expectantly normally distributed. We decided to tackle this data by sampling from the latitude portion of the data. We will also sample a real world example regarding reaction times of schizophrenics versus non-schizophrenics.

In this project we will outline two separate methods to sample given data. The first method is Expectation-Maximization, and Gibbs Sampling. We will outline the specific process used to employ both of these algorithms, the methods governing each approach, real and simulated data, and final our conclusions as to the performance of these methods and analyzing both approaches.

1 Introduction

Expectation-Maximization (EM) is an iterative, frequentist process for computing the maximum likelihood estimations (MLE). This aids us in determining the parameters that then give us the distribution of the model produced by the data that were observed. Gibbs sampling is a Bayesian, Monte-Carlo Markov Chain technique; it generates posterior samples from the distribution conditioned on prior components sampled so far, and then iterates that step. The primary difference between the two processes is that EM returns a single point from the distribution, and Gibbs sampling is the posterior distribution conditioned on the given data.

As part of our project, we were asked to look at utilizing the two different techniques for mixture modeling and comparing both methods with real and simulated data. The data that was provided to us is for calhouse in the datasets. Specifically, we were tasked with modeling latitude data and reaction times of schizophrenics.

2 Description of the problem

By coding EM and Gibbs Sampler algorithms for mixture modelling we evaluate and compare the behaviour of two methods. For the EM Algorithm we modify the code available in EM.here. This algorithm although considered only the

case of two mixture sample, it allows easily extend to the case of more than two subgroups and with small changes we could modify for the mixture distributions different from normal.

For the Gibbs Sampler algorithm we modify the code available in Gibbs.here by following the recommendations in chapter 22 [1] and we include the estimation of the variance since in the original one consider that constant. The adaptation of this code meets the requirements of the suggested project.

2.1 Mixture distributions

A mixture model represents and allows us to infer upon a complex structure of a random variable. This model corresponds to a mixture distribution, which, simply put, is a mixture of two or more probability distributions. The probability distributions in the mixture can be univariate or multivariate and can be the same or different types of discrete/continuous distributions with varying parameters. The simulated and California data sets in this report deal strictly with Gaussian Mixture Models.

Mixture models are commonly estimated by attributing weights (i.e., probabilities that sum to 1) to the sub-distributions. The EM Algorithm is used in this "clustering" method and estimates the model using latent variables. On the other hand, the Gibbs Sampler (or MCMC) method estimates the model through conditioning the prior or posterior distribution.

2.2 EM algorithm

The EM Algorithm is split into two steps; expectation step, and maximization step.

1. E step: Given the current values (θ', p') compute the expected value of

$$log(p(y, \Delta|\theta, p)) \tag{1}$$

   where the expectation is over $\Delta|y, \theta, p$
2. M step: Get new values of (θ, p) by optimizing the expected log likelihood computed in the E step. As usual, iterate until convergence.

2.3 Gibbs sampling

In Gibbs sampling, we observe the posterior, i.e. the data, starting with the parameters given; i.e. let θ=(θ_1,θ_2,...,θ_k) where is θ is a set of parameters. We then iterate each step by conditioning the next step, i.e. θ_j | θ_1, θ_2,...,θ_{j-1}, θ_{j+1},..., θ_k,... specifically the more recent steps. This allows us to draw from a function π, where $\pi(\theta_k)$ and θ is a draw from the marginal under π, the joint distribution.

3 Simulations examples

3.1 Univariate Mixture of Two Normal Distributions

To observe the behavior of the EM and MCMC methods, we start by simulating univariate mixture random variables composed of two Gaussian, i.e. normal, distributions with a total of 1000 instances. The first univariate mixture random variable, X, is composed of two distributions with means μ_1 = -2, μ_2 = 2, and equal variances σ_1^2=σ_2^2=1 (see Figure 1a). The second univariate mixture random variable, Y, is also composed of two distributions with means μ_1 = -2, μ_2 = 2, but now with unequal variances of σ_1^2=1 and σ_2^2=2 (see Figure 1b).

Then both EM and MCMC methods are used to estimate the parameters of these distributions. For the mixture with equal variances (x), the EM Algorithm took 19 iterations to converge the log-likelihood with mixture proportions of 0.540 and 0.537 (see Figure 2 for iteration convergence and Figure 3a for densities). The Gibbs Sampler with 5000 iterations took a processing time of 30.2 seconds to generate mixing proportions of 0.537 and 0.463, roughly the same as the output from EM (see Figure 3b for densities).

For the mixture with unequal variances (y), the EM Algorithm took 50 iterations to converge the log-likelihood with mixture proportions of 0.538 and 0.461. (see Figure 4 for iteration convergence and Figure 5a for densities). The Gibbs Sampler with 5000 iterations took a processing time of 54.08 seconds to generate mixing proportions of 0.649 and 0.351. (see Figure 5b for densities).

Table 1 shows the results of the estimated parameters for the two mixtures. For the mixture with equal variances, both EM and MCMC offer nearly identical results. For the mixture with unequal variances, however, the EM Algorithm gives

a much larger spread to the second group (almost double the original variance). The Gibbs Sampler ends up separating the groups much further from each other than EM but with similar variances. In both cases, the EM Algorithm converges much faster than the Gibbs Sampler method.

		Mean		Variance		Size (%)	
Variable	Method	1	2	1	2	1	2
x	EM	-1.978	2.021	1.038	0.966	0.540	0.459
	Gibbs	-1.989	1.999	1.039	1.046	0.537	0.463
y	EM	-2.020	2.029	0.997	3.877	0.538	0.461
	Gibbs	-1.751	2.806	1.190	1.519	0.649	0.351

Table 1: Estimate parameters

3.2 Univariate Mixture of Two Normal and One Gamma Distributions

In this case we simulate a mixture distributions from two normal distributions with means 5 and 7, and variances 1 and 0.5, respectively. The gamma distribution has shape $\alpha = 2$ and rate $\beta = 2$ parameters. The proportions for the three subgroups are 0.35, 0.3 and 0.35, respectively. To evaluate the EM algorithm in this example we use the R package mclust which identify 5 components. By restricting to 3 components we get results (estimate parameters) close to the real ones Table 2. Regarding to the Gibbs sampling algorithm the estimate parameters seem similar to the real parameters. In the Figure 6a we show the densities approximations from EM and Gibbs, in particular the results from EM tends to be more accurate. Although with both EM and Gibbs we use mixture of normal distributions the approximation is amazing. This result tries to mimic the point mention in [1] Chapter 22 that *any density can be accurately approximate using a mixture of sufficiently many normals*.

	Mean			Variance			Size (%)		
Method	1	2	3	1	2	3	1	2	3
EM	0797	5.050	7.047	0.243	1.473	0.144	0.292	0.426	0.282
Gibbs	0.989	5.095	6.989	0.820	1.057	1.063	0.312	0.347	0.312

Table 2: Estimate parameters

4 Real applications

4.1 Latitude variable

This example comes from a data set located on the STP 540 course webpage (access data at calhouse.csv). Although this data set includes several parameters, the parameter that offered one of the more interesting, multi-modal, distributions was the Latitude parameter of 20,000+ houses in California (see Figure 7 for histogram).

Figure 8a is the EM Algorithm used to model the sample data. As it shows, these two are nearly identical. Figure 8b, the Gibbs sampling, equally, is practically identical in shape to Figure 8a, but in the Gibbs model, we have a distribution instead of a single point. It is likely this would lead to greater accuracy in the Gibbs model. Mathematically, by conditioning on the posterior there is a greater use of the sample data available, whereas the EM selects via the indicator function, Gibbs utilizes the marginal distribution based on all previous steps. This does not appear to be a problem for this data, as both samples mimic the data adequately, and such as we can see in Figure 9 where there is not evidence of dependence for both mean without burning and variance after burning 1000. Further, we can see the similarity in the mixture model parameters, generated in Table 3, where we outline mean, variance, and size of each subgroup.

Other notable differences came in processing times. EM Algorithm had a time difference of 0.75 seconds from beginning to end. Gibbs sampling algorithm had a time difference of 52.95 secs from beginning to end (this does not include loading the data set). 70 times the computing need to run Gibbs vs EM. So, there could be a potential accuracy vs. computing required equation to balance depending on what one is working on.

4.2 Reaction times of schizophrenics

This example comes from Chapter 22 of [1]. The data is available at schizophrenics.data. The reaction times for 11 non-schizophrenics and 6 schizophrenics are recorded 30 times each. The boxplots (shown in Figure 10) are a

preliminary indicator that the reaction times of the schizophrenic participants have both higher mean and variance. The objective of this example is to estimate the difference (if there is one) the distributions of reaction times for schizophrenic vs. non-schizophrenic participants. This is done by finding the mixture distributions and comparing modes.

One detail before discussing the results of the Gibbs and EM algorithms is that reactions times were logged prior to analysis (because they were all positive).

As we can see from the Figure 11a) and the table of results in Table 3, the EM Algorithm did a great job of separating the two distributions. The Gibbs Sampler on the other hand did not perform so well 11. We believe this is due to the small amount of data available in the sample. However, with the two methods the estimation of the difference between the two modes (means subgroup 1 and subgroup 2) is similar to the estimation of the parameter τ in the Chapter 22. That can be checked easily from Table 3.

		Mean		Variance		Size (%)	
Variable	Method	1	2	1	2	1	2
Latitude	EM	33.857	37.835	0.251	1.151	0.554	0.446
	Gibbs	33.930	37.940	0.633	0.972	0.578	0.422
Shizop	EM	5.743	6.295	0.036	0.184	0.778	0.221
	Gibbs	5.721	6.134	0.414	0.637	0.647	0.353

Table 3: Estimate parameters. Latitude and Shizop are the results of the application exercises.

5 Conclusions and final comments

We want to applaud both algorithms. By using real and simulated data we can conclude that the approximation from both methods is a great start point of analysis particularly with messier looking data. Simulated exercises allow us to check how each method works and the accuracy of their estimations. However, the estimation of the variance mainly is hard as we evidenced in the simulation y and the case of three mixture distributions.

About the real examples for the latitude variable was relatively easy the estimate process for both methods, although initially the R package mclust identified 9 components. For schizophrenics case the estimation is more complicated perhaps since either the small sample size or the small difference between the two groups.

Most notably, both algorithms were easy to grasp and have several real-world applications. Besides the great showdown between frequentist vs Bayesian, any statistical modeler would keep both tools in their proverbial arsenal.

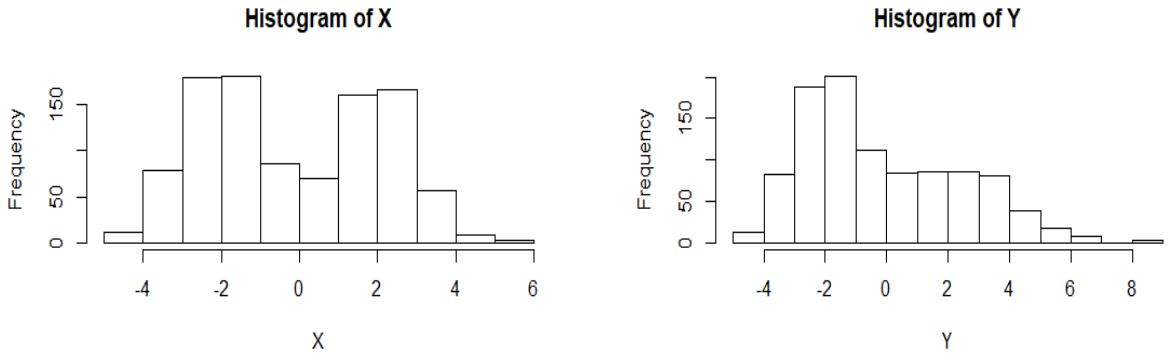
References

[1] Andrew Gelman, John Carlin, Hal Stern, David Dunson, Aki Vehtari, and Donald Rubin. Bayesian Data Analysis. Chapman and Hall Book, Third edition, 2014.

[2] George Kour and Raid Saabne. Fast classification of handwritten on-line arabic characters. In *Soft Computing and Pattern Recognition (SoCPaR), 2014 6th International Conference of*, pages 312–318. IEEE, 2014.

[3] Guy Hadash, Einat Kermany, Boaz Carmeli, Ofer Lavi, George Kour, and Alon Jacovi. Estimate and replace: A novel approach to integrating deep neural networks with existing applications. *arXiv preprint arXiv:1804.09028*, 2018.

6 Figures



(a) (b)

Figure 1: (a) Univariate Mixture Random Variable with Equal Variances. (b) Univariate Mixture Random Variable with Unequal Variances

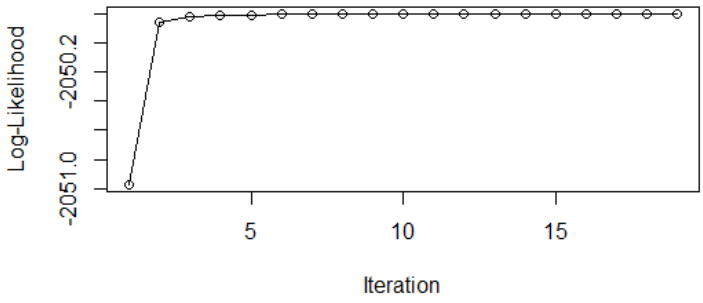
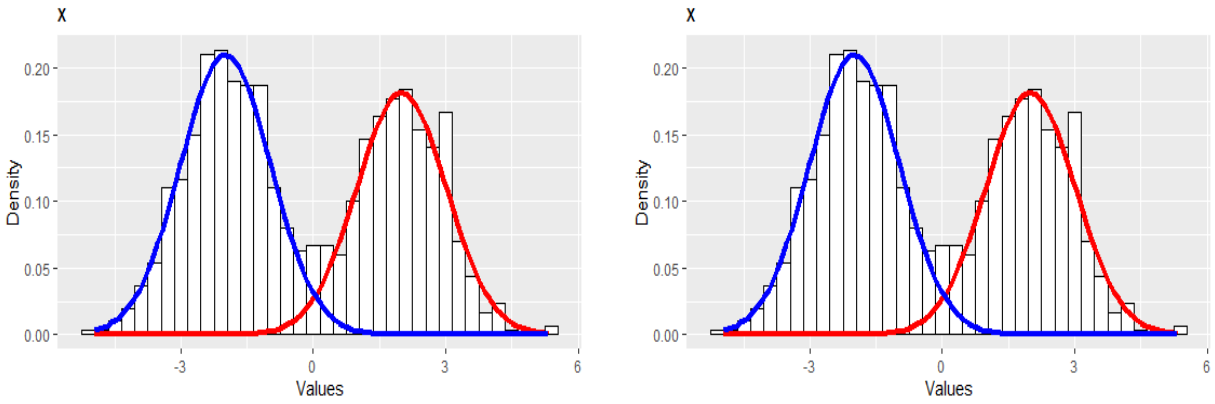


Figure 2: EM Algorithm on X: Simulated data log-likelihood as a function of iteration number



(a) (b)

Figure 3: (a) EM Algorithm on X: Mixture Distributions on Group 1 (Blue) and Group 2 (Red). (b) Gibbs Sampler/MCMC on X: Mixture Distributions on Group 1 (Blue) and Group 2 (Red)

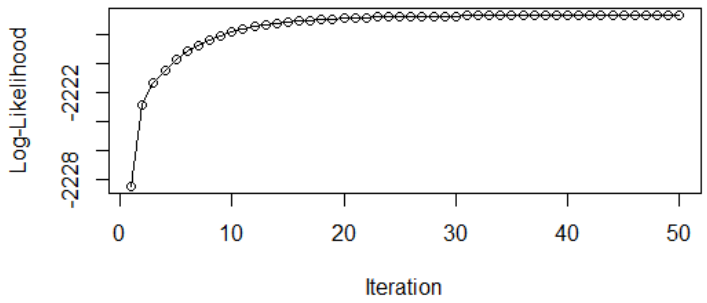
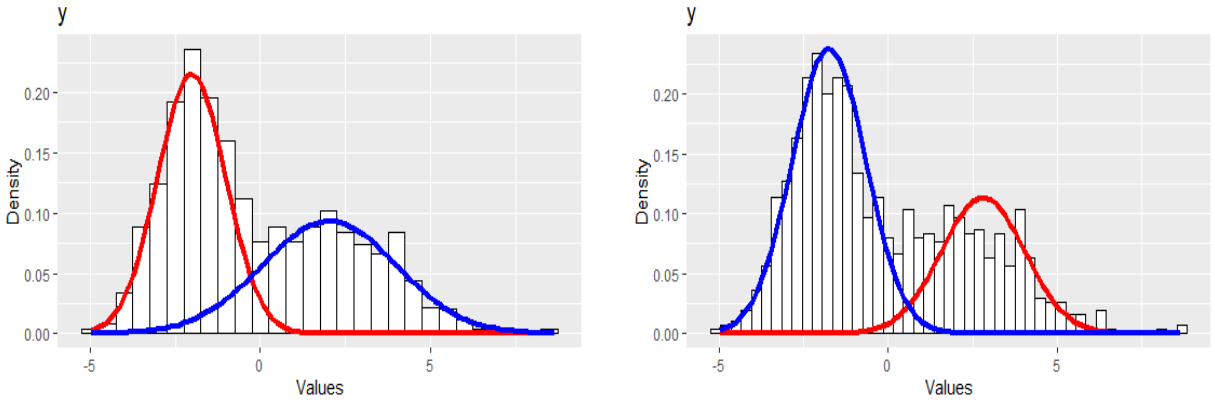
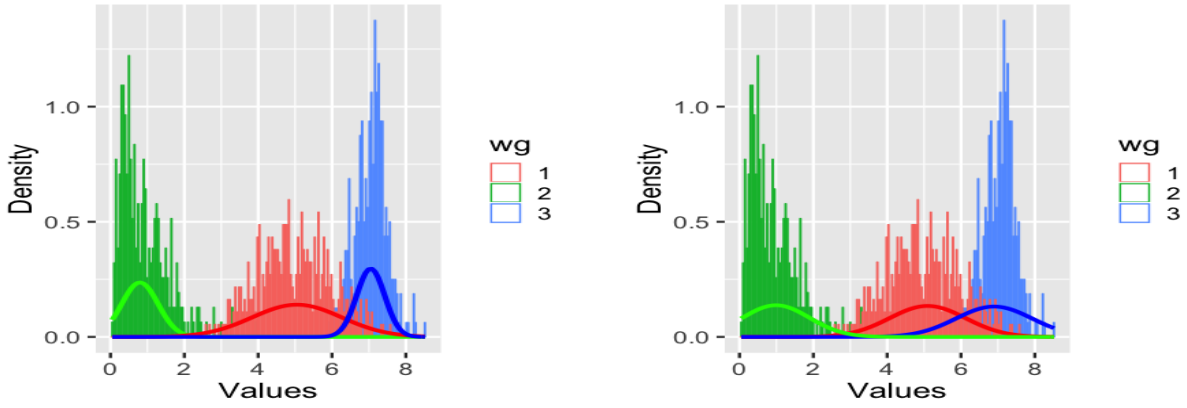


Figure 4: EM Algorithm on Y: Simulated data log-likelihood as a function of iteration number



(a) (b)

Figure 5: (a) EM Algorithm on Y: Mixture Distributions on Group 1 (Blue) and Group 2 (Red). (b) Gibbs Sampler/MCMC on Y: Mixture Distributions on Group 1 (Blue) and Group 2 (Red)



(a) (b)

Figure 6: (a) Density from EM Algorithm (b) Density from Gibbs Sampler/MCMC

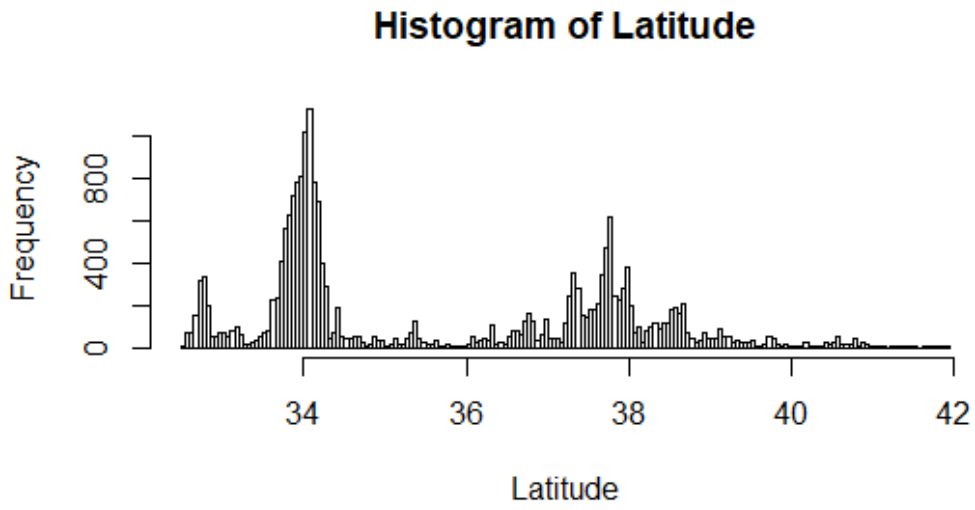
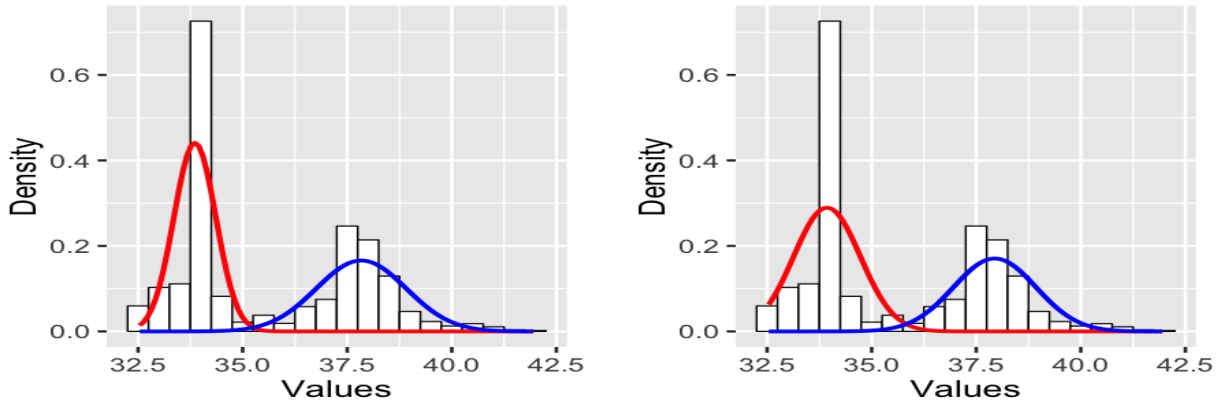


Figure 7: Latitude Histogram from calhouse.csv data



(a) (b)

Figure 8: (a) Latitude distribution from EM algorithm. (b) Latitude distribution from Gibbs Sampling

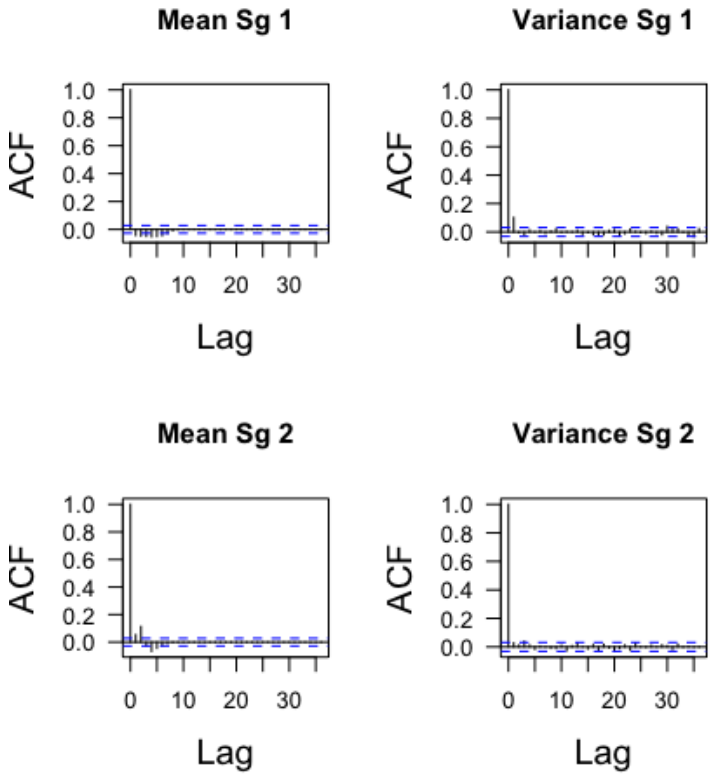


Figure 9: Autocorrelation function for mean and variance in the two subgroups. For the mean we do not burn, and for the variance we burn the first 1000

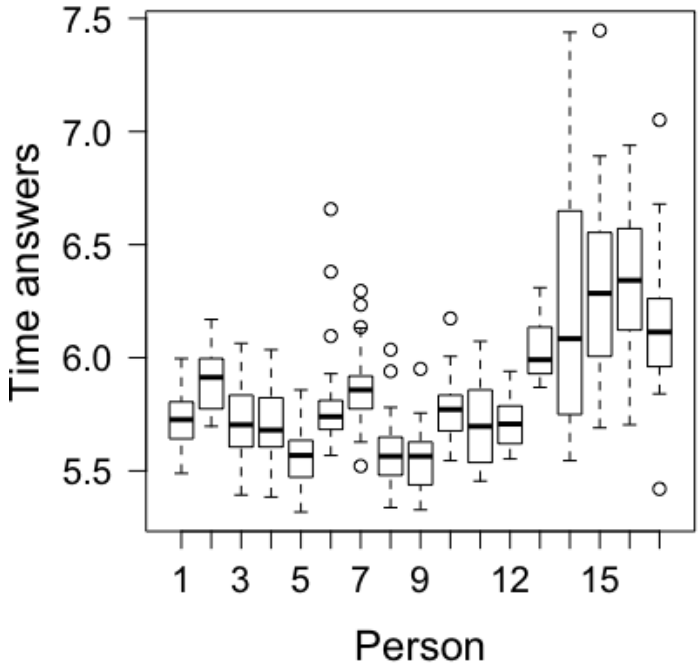
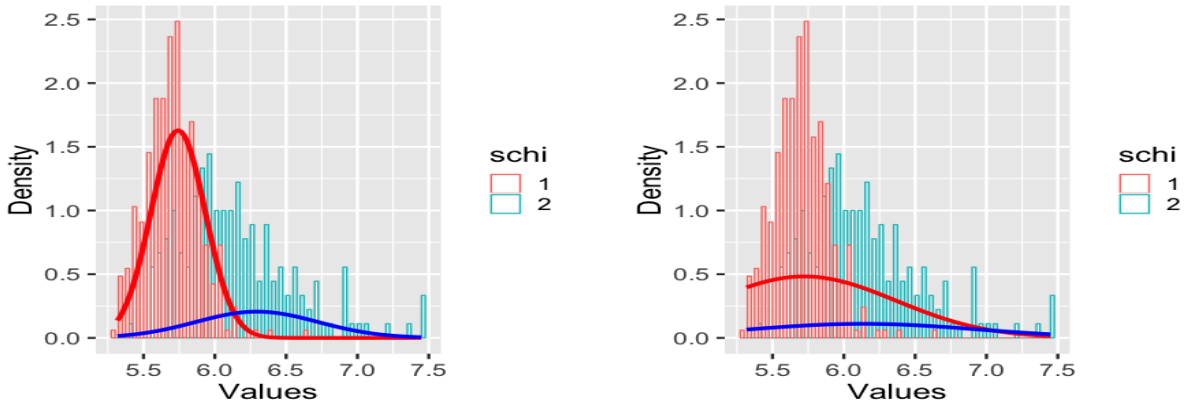


Figure 10: Reaction times of 11 no-schizophrenics and 6 schizophrenics people



(a) (b)

Figure 11: (a) Schizophrenics distribution from EM algorithm. (b) Schizophrenics distribution from Gibbs Sampling. For both (a) and (b) 1 means no-schizophrenics and 2 schizophrenics