2012

Bayesian Regression Inference Using a Normal Mixture Model

Hernan Maldonado

BAYESIAN REGRESSION INFERENCE USING A NORMAL MIXTURE MODEL

A Thesis

Submitted to McAnulty College

and Graduate School of Liberal Arts

Duquesne University

In partial fulfillment of the requirements for

the degree of Masters of Science

By

Hernan Maldonado

August 2011

BAYESIAN REGRESSION INFERENCE USING A NORMAL MIXTURE MODEL

By

Hernan Maldonado

Approved July 19, 2011

John Kern, Ph.D.
Associate Professor of Statistics
(Dissertation Director)

Donald Simon, Ph.D.
Associate Professor of Computer Science
(Director of Graduate Study)

Eric Ruggieri, Ph.D.
Assistant Professor of Statistics
(Committee Member)

Jeffrey Jackson, Ph.D.
Professor of Computer Science
Department Chair

James C. Swindal, Ph.D.
Acting Dean
McAnulty College and Graduate School
of Liberal Arts

ABSTRACT

BAYESIAN REGRESSION INFERENCE USING A NORMAL MIXTURE MODEL

By

Hernan Maldonado

August 2011

In this thesis we develop a two component mixture model to perform a Bayesian regression. We implement our model computationally using the Gibbs sampler algorithm and apply it to a dataset of differences in time measurement between two clocks. The dataset has "good" time measurements and "bad" time measurements that were associated with the two components of our mixture model. From our theoretical work we show that latent variables are a useful tool to implement our Bayesian normal mixture model with two components. After applying our model to the data we found that the model reasonably assigned probabilities of occurrence to the two states of the phenomenon of study; it also identified two processes with the same slope, different intercepts and different variances.

ACKNOWLEDGMENT

I would like to thank my advisor, Dr. John Kern II, for being an outstanding advisor, mentor and person. His ideas and guidance in the process of writing this thesis were essential to its completion. John's enthusiasm and encouragement were an important part of my continued motivation to work and to complete this project. His personal warmth and helpfulness are invaluable. John has my wholehearted gratitude.

Special thanks to my committee and the entire Mathematics and Computer Science Department faculty, staff, and my fellow students. Yinz made my studies at Duquesne a wonderful experience. I want to thank my officemate of two great years Monir Sharker for his wise advice, for his patience and his never-ending understanding, a great man he is.

I would also like to thank Nicole Pernischova for providing me with the data for this thesis.

I want to mention that my studies at Duquesne were possible thanks to the financial support of the Mathematics and Computer Science Department. I'm very grateful to the institution and to the people that represent it and work on it. May God continue blessing this school and its people.

Finally, I would like to thank my wife Laura and my son Benjo for all their love and support in the process of completing this thesis and the Masters in Computational Mathematics degree. It has been a great and enriching experience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Box and Tiao (1968) introduced a Bayesian procedure to analyze a phenomenon in which a given set of observations were considered to be generated by more than one specific stochastic model. In their paper it is assumed that a linear model generates data from two different processes of differing variance. It is assumed that one process has a variance of σ^2 while the other has a variance of $k^2\sigma^2$ and both processes share the same mean. This seminal paper was written to tackle the need to expand the common assumption that:

> "Most statistical procedures are arrived under the assumption that *each one* of a given set of observations is generated by a specific stochastic model containing a modest number of adjustable parameters" Box and Tiao (1968).

Since the publication of the paper by Box and Tiao (1968) research on this topic has flourished and today mixture models are commonly used to solve problems of classification in heterogeneous populations. In this context heterogeneous population classification can be understood as the challenge to associate an event to its causing phenomenon. As Gelman et al. (2004) describe, the distribution of heights in a population of adults reflects the mixture of males and females in the population. Mixture models are a tool to model these heterogeneous populations (males and females) by using separate univariate distributions, rather than a single bimodal distribution.

An application of mixture models is done by Ding (2006) who uses regression mixture models to analyze students' performance in mathematics across races and gender. Ding's paper studies whether the effects of independent variables on a dependent variable differ across groups, either in terms of intercept or slope. The author found evidence that is consistent with studies based on conventional regression analysis showing that child's math self-concept would be a strong predictor of actual math performance, of social competence, and of approach to learning; however the author's findings revealed that self-concept does not predict well average math performance for children. The differences between the results obtained with regression mixture models and the classical regression shows that using mixture models can present a different perspective on the results of an analysis.

In this thesis a mixture model with latent variables is developed and implemented. As pointed out by Muthén (2001), the goal of using latent variables is to identify items that indicate classes well, estimate class probabilities, relate class probabilities to covariates, and classify individuals into classes.

The data used in this thesis was obtained from timing differences between two clocks: a Windows clock and an Rbox clock. A Windows clock is the clock that comes with the Windows OS. An Rbox clock is another timing device that can be installed into a computer to increase the timing accuracy (over that of the Windows clock).

A Windows clock was the device in charge of recording time during the experiments. However, it sometimes records its time and waits several (crucial) milliseconds to ask and record Rbox's time. This leads to large discrepancies between the two clocks when measuring the time of a frequently repeated experiment.

The challenge then becomes to identify which timing measurements are “good” measurements and which are “bad” measurements in repeated experiments. This differentiation between “good” and “bad” allows scientists to discard those measurements that are not coming from an accurate report of the time by the PC device.

We apply our Normal Mixture model to the data by programming a Gibbs Sampler algorithm developed in Geman and Geman (1984) using the R statistical package[1].

In this thesis we found that latent variables are a good tool to implement a Bayesian Normal Mixture Model with two components. In an experiment with two states, “bad” and “good”, a Bayesian Normal Mixture Model appropriately assigned probabilities of occurrence to the two states of the phenomenon clearly identifying two processes with the same slope, different intercepts and different variances.

The next chapter of this thesis is a presentation of the statistical model and the computational algorithm. The third chapter presents an application of the model to a data-set; the fourth chapter presents the results and the final chapter concludes.

[1] R Development Core Team (2010)

Chapter 2

The Statistical Model

The model presented here is a two component normal mixture model. The density function for a random variable Y believed to come from one of two simple linear regression equations is [1]:

$$
\begin{aligned}
f(y \mid \boldsymbol{x}) =& p \left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \right) \exp \left\{ -\frac{[y - (\beta_{01} + \beta_{11}x)]^2}{2\sigma_1^2} \right\} + \\
& (1-p) \left(\frac{1}{\sqrt{2\pi\sigma_2^2}} \right) \exp \left\{ -\frac{[y - (\beta_{02} + \beta_{12}x)]^2}{2\sigma_2^2} \right\}
\end{aligned}
\tag{2.1}
$$

Equation 2.1 represents a belief that a phenomenon with two outcomes can be described using two mutually exclusive linear processes that differ from each other on intercept ($\beta_{01} \neq \beta_{02}$), slope ($\beta_{11} \neq \beta_{12}$) or variance ($\sigma_1^2 \neq \sigma_2^2$). In this model one of the linear processes takes place with probability p and the other with probability $(1-p)$.

Using latent indicators z_i, the joint distribution of the dependent variables

[1] The notation for this chapter can be better followed if you refer to Appendix C

$\{y_1, y_2, \cdots, y_n\}$ and indicators $\{z_1, z_2, \cdots, z_n\}$ is[2]:

$$\pi(y \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \left\{ p \left(\frac{1}{\sqrt{2\pi\sigma_1^2}} \right) \exp\left\{ -\frac{[y_i - (\beta_{01} + \beta_{11}x_i)]^2}{2\sigma_1^2} \right\} \right\}^{1-z_i} \cdot \left\{ (1-p) \left(\frac{1}{\sqrt{2\pi\sigma_2^2}} \right) \exp\left\{ -\frac{[y_i - (\beta_{02} + \beta_{12}x_i)]^2}{2\sigma_2^2} \right\} \right\}^{z_i} \tag{2.2}$$

To estimate the parameters of the model we use a Bayesian approach. This approach recognizes the posterior distribution for the parameters as proportional to the likelihood times the joint prior for those parameters. In this thesis we specify Jeffreys priors (Jeffreys (1961))[3]
for σ_1^2, σ_2^2 and constant non-informative priors for $p, \beta_{11}, \beta_{12}$ and z, and use empirical Bayes[4] to specify the hyper-parameters (μ_0, μ_1, v_0, v_1) for Gaussian priors on β_{01}, β_{02}. When we applied our model to our dataset, we faced a common problem on Bayesian mixture models known as the "label switching problem". This problem is caused by the symmetry in the likelihood of the model parameters. To address this issue we used empirical Bayes to set the values of hyper-parameters of the informative prior distributions for our β_{01} and β_{02} parameters [5]. This approach gives a joint prior of [6]

$$g(\boldsymbol{\theta}) \propto \frac{1}{\sqrt{2\pi v_0}} \exp\left\{ -\frac{1}{2v_0} (\beta_{01} - \mu_0)^2 \right\} \cdot \frac{1}{\sqrt{2\pi v_1}} \exp\left\{ -\frac{1}{2v_1} (\beta_{02} - \mu_1)^2 \right\} \cdot \frac{1}{\sigma_1^2} \cdot \frac{1}{\sigma_2^2} \tag{2.3}$$

Where μ_0 ia a hyper-parameter that represents the mean intercept for the

[2]Using $\boldsymbol{\theta} = \{\mathbf{z}, p, \beta_{01}, \beta_{11}, \sigma_1^2, \beta_{02}, \beta_{12}, \sigma_2^2\}$ and $\boldsymbol{\theta}_{-\boldsymbol{\sigma_1^2}} = \{\mathbf{z}, p, \beta_{01}, \beta_{11}, \beta_{02}, \beta_{12}, \sigma_2^2\}$ to shorten the notation

[3]Jeffreys prior satisfies the local uniformity property: a prior that does not change much over the region in which the likelihood is significant and does not assume large values outside that range. It is based on the Fisher information matrix.

Jeffreys prior is locally uniform and hence noninformative. It provides an automated scheme for finding a noninformative prior for any parametric model $p(\boldsymbol{y} \mid \boldsymbol{\theta})$. Another appealing property of Jeffreys prior is that it is invariant with respect to one-to-one transformations. The invariance property means that if you have a locally uniform prior on $\boldsymbol{\theta}$ and $\phi(\boldsymbol{\theta})$ is a one-to-one function of $\boldsymbol{\theta}$, then $p(\phi(\boldsymbol{\theta})) = \pi(\boldsymbol{\theta}) \cdot |\phi'(\boldsymbol{\theta})|^{-1}$ is a locally uniform prior for $\phi(\boldsymbol{\theta})$.

[4]Carlin and Louis (2008)

[5]For a description of the "label switching problem" see Stephens (2000)

[6]The full derivation of the priors can be found on Appendix A.0.2

bottom cloud given by a ordinary linear regression ran on the data. Also this is the starting value for the β_{01} parameter. μ_1 is a hyper-parameter that represents the mean intercept for the top cloud given by a ordinary linear regression ran on the data. Also this is the starting value for the β_{02} parameter. v_0 is a hyper-parameter that represents the expected variance of the intercept for the bottom cloud given by a ordinary linear regression ran on the data. Also this is the starting value for the σ_1 parameter. v_1 is a hyper-parameter that represents the expected variance of the intercept for the top cloud given by a ordinary linear regression ran on the data. Also this is the starting value for the σ_2 parameter.

Combining 2.2 and 2.3 via multiplication gives the posterior distribution for $\boldsymbol{\theta}$:

$$\pi(\boldsymbol{\theta}) \propto \pi(\boldsymbol{y} \mid \boldsymbol{\theta}) \cdot g(\boldsymbol{\theta}) \tag{2.4}$$

Given a multivariate posterior distribution for $\boldsymbol{\theta}$, it is easier to sample from a conditional distribution than to marginalize by integrating over a joint distribution. Then the full conditional distributions for the parameters $\boldsymbol{\theta}$ of our model are [7] [8]:

Let

$$A = p\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)\exp\left\{-\frac{[y_i - (\beta_{01} + \beta_{11}x_i)]^2}{2\sigma_1^2}\right\}$$

$$B = (1-p)\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)\exp\left\{-\frac{[y_i - (\beta_{02} + \beta_{12}x_i)]^2}{2\sigma_2^2}\right\}$$

$$\boxed{z_i \mid \boldsymbol{\theta}_{-\mathbf{z}}, \mathbf{x}, \mathbf{y} \sim \text{Bern}\left(\tfrac{A}{A+B}\right)}$$

$$\boxed{\beta_{01} \mid \boldsymbol{\theta}_{-\boldsymbol{\beta}_{01}}, \mathbf{x}, \mathbf{y} \sim N\left(\overline{y}_0 - \beta_{11}\overline{x}_0, \tfrac{\sigma_1^2}{n_0}\right)}$$

$$\boxed{\beta_{02} \mid \boldsymbol{\theta}_{-\boldsymbol{\beta}_{02}}, \mathbf{x}, \mathbf{y} \sim N\left(\overline{y}_1 - \beta_{12}\overline{x}_1, \tfrac{\sigma_2^2}{n_1}\right)}$$

$$\boxed{\beta_{11} \mid \boldsymbol{\theta}_{-\boldsymbol{\beta}_{11}}, \mathbf{x}, \mathbf{y} \sim N\left(\frac{\sum\limits_{i\ni z_i=0} y_i x_i}{\sum\limits_{i\ni z_i=0} x_i^2} - \beta_{01}\frac{\sum\limits_{i\ni z_i=0} x_i}{\sum\limits_{i\ni z_i=0} x_i^2}, \frac{\sigma_1^2}{\sum\limits_{i\ni z_i=0} x_i^2}\right)}$$

[7] Derivation of the full conditional densities can be found on Appendix A

[8] $n_j = \sum\limits_{i\ni z_i=j} z_i$ for $j = 0, 1$

$$\beta_{12} \mid \boldsymbol{\theta_{-\beta_{12}}}, \mathbf{x}, \mathbf{y} \sim N\left(\frac{\sum\limits_{i \ni z_i=1} y_i x_i}{\sum\limits_{i \ni z_i=1} x_i^2} - \beta_{02}\frac{\sum\limits_{i \ni z_i=1} x_i}{\sum\limits_{i \ni z_i=1} x_i^2}, \frac{\sigma_2^2}{\sum\limits_{i \ni z_i=1} x_i^2}\right)$$

$$\sigma_1^2 \mid \boldsymbol{\theta_{-\sigma_1^2}}, \mathbf{x}, \mathbf{y} \sim \text{Inv-Gamma}\left(\frac{n_0}{2}, \sum_{i \ni z_i=0} \frac{[y_i-(\beta_{01}+\beta_{11}x_i)]^2}{2}\right)$$

$$\sigma_2^2 \mid \boldsymbol{\theta_{-\sigma_2^2}}, \mathbf{x}, \mathbf{y} \sim \text{Inv-Gamma}\left(\frac{n_1}{2}, \sum_{i \ni z_i=1} \frac{[y_i-(\beta_{02}+\beta_{12}x_i)]^2}{2}\right)$$

$$p \mid \boldsymbol{\theta_{-\mathbf{p}}}, \mathbf{x}, \mathbf{y} \sim \text{Beta}\,(1+n_1, 1+n_0)$$

Having recognized the full conditional distributions for the parameters in our model we apply the Gibbs Sampler algorithm to sample from the marginal distribution of the parameters in the model. The Gibbs Sampler is presented in the following Section in the context of our model.

2.1 The Gibbs Sampler Algorithm

We used the Gibbs sampler framework developed in Geman and Geman (1984) to sample from the conditional distributions presented in Section 2. The idea of the Gibbs sampler algorithm in this case is that:

Given a multivariate distribution it is simpler to sample from a conditional distribution than to marginalize by integrating over a joint distribution. Suppose we want to obtain k samples of $\boldsymbol{\theta} = \{\mathbf{z}, p, \beta_{01}, \beta_{11}, \sigma_1, \beta_{02}, \beta_{12}, \sigma_2\}$ from a joint distribution $p(\mathbf{z}, p, \beta_{01}, \beta_{11}, \sigma_1, \beta_{02}, \beta_{12}, \sigma_2)$. Denote the i^{th} sample by $\boldsymbol{\theta}^{(i)} = \{\mathbf{z}^{(\mathbf{i})}, p^{(i)}, \beta_{01}^{(i)}, \beta_{11}^{(i)}, \sigma_1^{(i)}, \beta_{02}^{(i)}, \beta_{12}^{(i)}, \sigma_2^{(i)}\}$. We proceed as follows:

1. We begin with some initial value $\boldsymbol{\theta}^{(0)}$ for each variable parameter [9].

2. For each sample $i = \{1 \cdots k\}$, sample each variable parameter $\theta_j^{(i)}$ from the conditional distribution $p(\boldsymbol{\theta_j} \mid \boldsymbol{\theta_{-\theta_j}})$ [10] . That is, sample each of the

[9] The parameters were originally initialized to several different values. However upon running into the label switching problem we started the parameters close to values found after applying the k-means clustering technique to our data. This topic is expanded on Chapter 3

[10] In general an example of this process is $p(\boldsymbol{\theta_j^{(i)}} \mid \boldsymbol{\theta_1^{(i)}}, \cdots, \boldsymbol{\theta_{j-1}^{(i)}}, \boldsymbol{\theta_{j+1}^{(i-1)}}, \cdots, \boldsymbol{\theta_n^{(i-1)}})$

variable parameters from the distribution of that parameter conditioned on all other parameters, making use of the most recent values and updating the variable with its new value as soon as it has been sampled. For example: $p(\sigma_1^{(i)} \mid \mathbf{z^{(i)}}, p^{(i)}, \beta_{01}^{(i)}, \beta_{11}^{(i)}, \beta_{02}^{(i-1)}, \beta_{12}^{(i-1)}, \sigma_2^{(i-1)})$.

The samples then approximate the joint distribution of all model parameters. Furthermore the marginal distribution of any subset of parameters can be approximated by simply examining the samples for that subset of parameters, ignoring parameters that are not of interest. In addition, the posterior expected value of any parameter can be approximated by averaging over all the samples [11].

2.1.1 Computing

The computational aspect of the modelling was done in R[12], the full code is presented in Appendix B; this section presents the code for simulating from the joint posterior distribution for the parameters in the R programming language.

```
for(i in 1:(lagg*TOTAL+burn))
{
n0<-sum(1-z) ##Recalculating n0
n1<-n-n0     ##Recalculating n1

##***************** Calculations for b01 *************
R<-(sum((1-z)*y)/n0)-(b11*sum((1-z)*x)/n0)
b01<-rnorm(1,(R*n0*v01+m01*sigma1^2)/(n0*v01+sigma1^2)
,((n0/sigma1^2)+(1/v01))^(-1))

##***************** Calculations for b02 **************
K<-(sum(z*y)/n1)-(b12*sum(z*x)/n1)
b02<-rnorm(1,(K*n1*v02+m02*sigma2^2)/(n1*v02+sigma2^2)
,((n1/sigma2^2)+(1/v02))^(-1))


#****************** Calculations for b11 ***************
b11<-rnorm(1, (sum((1-z)*x*y)/sum((1-z)*x^2))
-b01*(sum((1-z)*x)/sum((1-z)*x^2)),sqrt(sigma1/sum((1-z)*x^2)))
```

[11] Gelman et al. (2004)
[12] R Development Core Team (2010)

```
##****************** Calculations for b12 **************
b12<-rnorm(1, (sum((z)*x*y)/sum((z)*x^2))
-b02*(sum((z)*x)/sum((z)*x^2)),sqrt(sigma2/sum((z)*x^2)))

##****************** Calculations for Sigma1 ************
sigma1 <- 1/rgamma(1,n0/2,(sum((1-z)*(y-(b01+b11*x))^2))/2)


##****************** Calculations for Sigma2 ************
sigma2 <- 1/rgamma(1,n1/2,(sum((z)*(y-(b02+b12*x))^2))/2)


##******************* Calculations for p **************
p <- rbeta(1,1+n1,1+n0)
while(p < 0.09 && p>0.93){
p <- rbeta(1,1+n1,1+n0)
}

##******************* Calculations for z ***************
exp1<-exp(-(0.5/sigma1)*(y-(b01+b11*x))^2)
exp2<-exp(-(0.5/sigma2)*(y-(b02+b12*x))^2)
zprob<-(p*(1/sqrt(sigma1))*exp1)/(p*(1/sqrt(sigma1))*exp1 +
(1-p)*(1/sqrt(sigma2))*exp2)
z<-rbinom(1000,1,zprob)


#**** Accumulation for parameters Values in vectors ****
if(i%%lagg==0&&i>burn)
{
b01vec<-c(b01vec,b01)
b02vec<-c(b02vec,b02)
b11vec<-c(b11vec,b11)
b12vec<-c(b12vec,b12)
svec1<-c(svec1,sigma1)
svec2<-c(svec2,sigma2)
pvec<-c(pvec,p)
zs<-rbind(zs,z)
}}
```

Chapter 3

The Experiment

Having established our normal mixture model, we now turn to present the dataset to which we apply our theoretical model. This dataset was provided by a private company to be used for research with the goal of receiving help on a timing issue that is described below.

3.1 The challenge

Researchers want to perform experiments with highly accurate timing using computers. Windows is a common operating system but it is not a Real Time Operating System (RTOS), and therefore is not very good at timing events accurately. A company that needed to time an experiment created the Rbox, a device which has a local hardware clock used to ensure accurate timing to 1 μs. Combined, the PC and the Rbox can provide a way to measure the expected delay from the PC clock reporting time.

The set-up of the time reporting system has the following steps:

1. PC records Rbox starting time
2. PC records its starting time
3. The experiment runs and ends

4. PC records its end time and

5. PC asks Rbox for its current time (end time)

The problem with this algorithm is the lazy nature of the Windows OS. Sometimes Windows records its time and waits several (crucial) milliseconds to ask the Rbox for its time. This leads to large deviations and inconsistent timings between the two time measuring devices.

The challenge then becomes to identify which experiment trials are "good" measurements and which ones are "bad" measurements. A mechanism to differentiate data coming from "good" measurements versus data coming from "bad" measurements would allow the experimenters to discard those runs that are not coming from an accurate report of the time by the PC device.

"Good" measurements are considered those that simultaneously meet two conditions :

- The starting times for the Rbox and the PC clocks are close in time [1]

- The ending times for the Rbox and the PC clocks are close in time

"Bad" measurements are those that don't meet the previous two conditions simultaneously. Example 1 on Figure 3.1 shows a case in which the experiment is measured correctly. Examples 2 through 4 on Figure 3.1 show cases in which the experiment is measured badly.

[1] There is no specific boundary to differentiate close and far in time on this experiment because the only times reported were the absolute times that each machine took measuring the event.

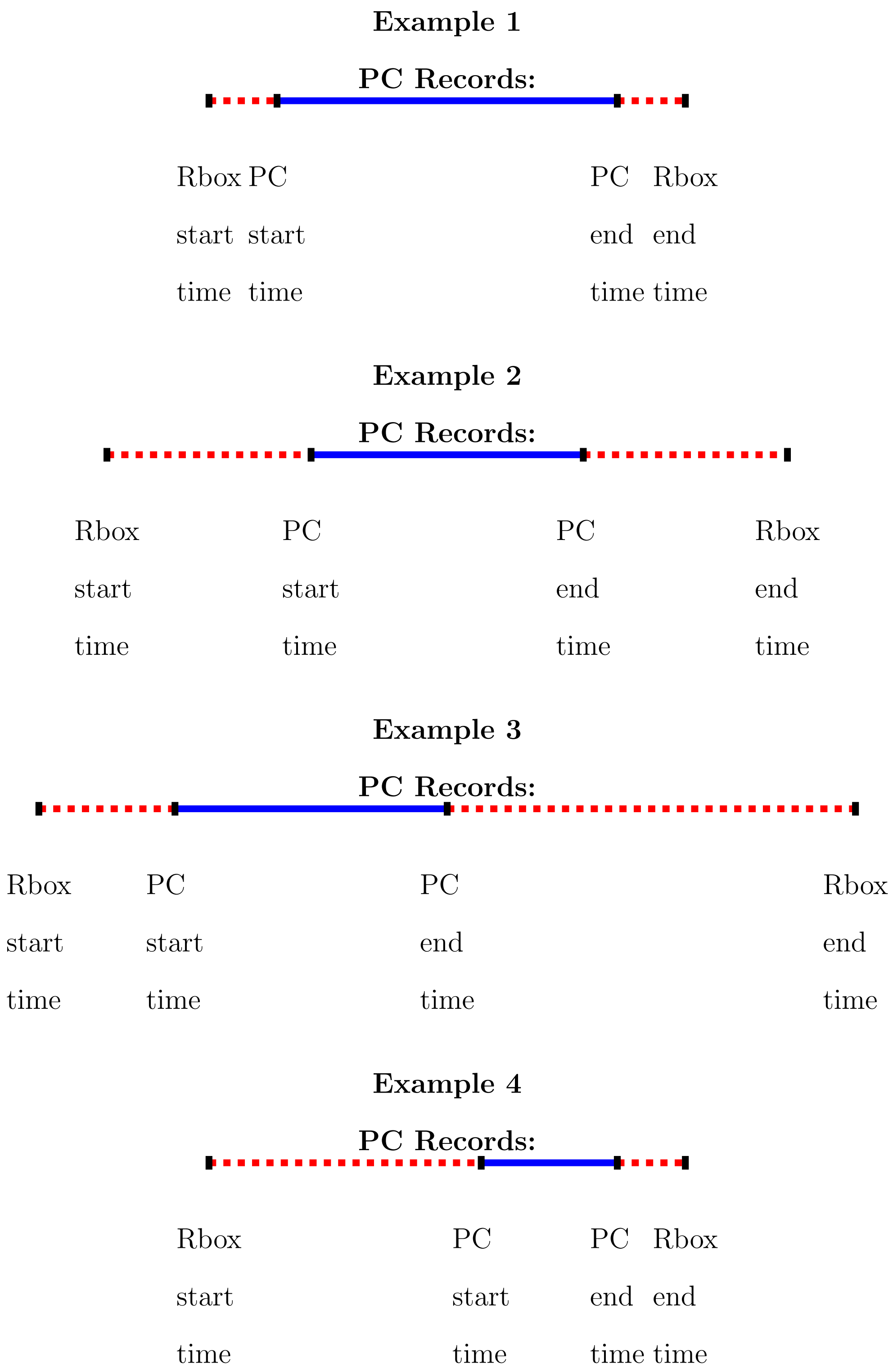

Figure 3.1: Time recording process

3.2 Data Description: The PC and RBox timing

The data obtained came from the time reported from 1000 repetitions of an experiment done on a PC computer with an Rbox timer. The measurement lasted 6 hours and 53 minutes and yielded 1000 observations of each time measuring device (Rbox, PC clock).

The data we are interested in analyzing is the difference in time measurement between the two time measuring devices. The data is plotted in Figure 3.2[2][3].

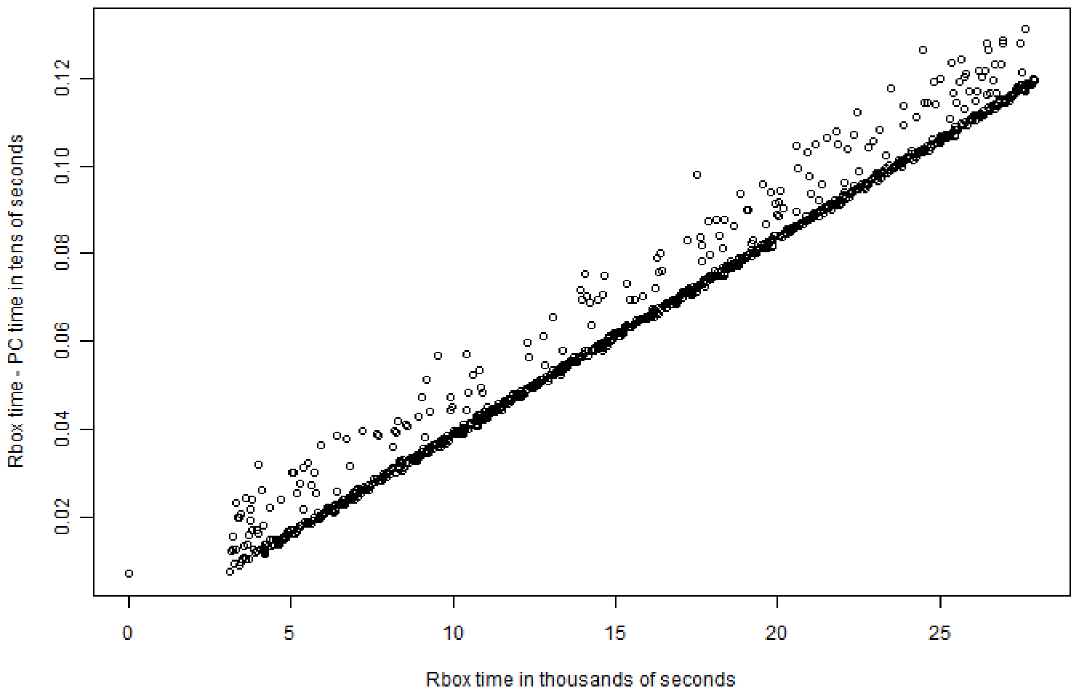

Figure 3.2: Differences between Rbox time and PC time

[2]The data axis were used to be consistent through the thesis. The scales were useful since they helped to the stability of the numerical algorithm.

[3]The first data point on the left is an outlier. It appeared on the data and we did not have much information on its causes therefore we decided to keep it in the dataset as originally obtained.

Chapter 4

Experiment Results

We applied the statistical model described in Chapter 2 to the data presented in Section 3.2 and used the following conditions for the Gibbs sampler:

- 5000 simulated values of each parameter were saved.
- A lag of 30 iterations were used to avoid autocorrelation.
- A burn of 1000 iterations were used to allow for convergence.
- A total of 151,000 iterations were ran.
- Jeffreys priors for σ_1^2

After running the model with non-informative uniform priors for β_{01} and β_{02} and running into the label switching problem, normal informative prior distributions were used for β_{01} and β_{02} (the parameters that represent the intercepts)[1]. These procedure was carried out by dividing the data in two groups using the k-means clustering method [2]. The two resulting groups are presented in figure 4.1.

[1]As we explained earlier, when we applied our model to our dataset, we faced a common problem on Bayesian mixture models known as the "label switching problem". This problem is caused by the symmetry in the likelihood of the model parameters. To address this issue we used empirical Bayes to set the values of informative prior distributions for our β_{01} and β_{02} parameters. For a description of the "label switching problem" see Stephens (2000)

[2]Given a set of observations $(x_1, x_2, \cdots, x_n)$ where each observation is a d-dimensional real vector, k-means aims to partitions the n observations into k sets $(k \leq n)$ $S = \{S_1, S_2, \cdots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS) $arg \min \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$ where μ_i is the mean of points in S_i Hair et al. (2005)

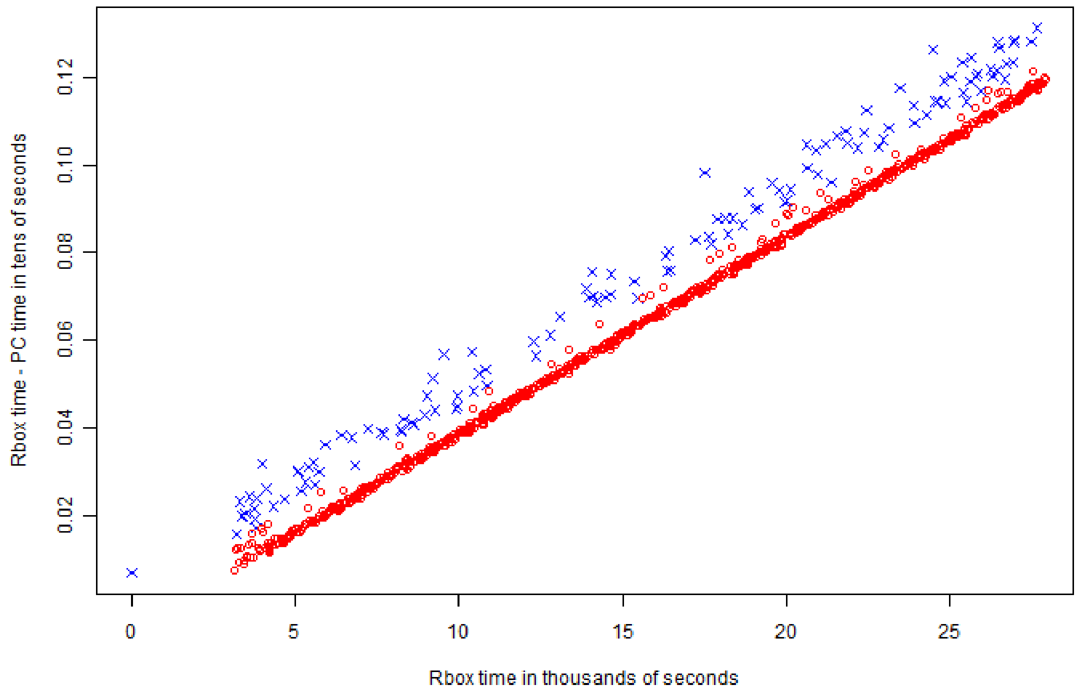

Figure 4.1: Cluster separation for time differences between Rbox and PC

On each one of these two groups of data we fit a linear model. We used the intercept as the mean for our prior distributions of β_{01} and β_{02} and the standard error of the intercept as our variance for the prior distributions.

We let the mean of the top cloud be μ_0 and the standard error for top cloud be v_0. Similarly for the bottom cloud the mean is defined to be μ_1 and the standard error v_1. The values for this parameters are presented on Table 4.1 [3]. Under these conditions the resulting parameter estimates are presented in Table 4.2:

The initial values of β_{01} and β_{02} were set to be equal to μ_1 and μ_0 respectively. The other starting values were chosen from repeated experiments to be close to their convergence values.

[3] The full mathematical derivation of the posterior distributions for β_{01} and β_{02} can be found on Appendix A.0.2

Table 4.1: Prior values

Bottom cloud with less variance	
Prior Parameter	Starting Value
μ_1	-0.006058537
v_1	9.335157e-05
Top cloud with more variance	
Prior Parameter	Starting Value
μ_0	0.005059207
v_0	0.0006941547

Table 4.2: Parameters Convergence

Bottom cloud with less variance		
	Starting Value	Average Value
β_{01}	-0.006058537	-0.006315958
β_{11}	0.004492471	0.004493892
σ_1^2	$2.368350e-07$	$2.990065e-07$
Top cloud with more variance		
	Starting Value	Average Value
β_{02}	0.005059207	-0.001409068
β_{12}	0.004509729	0.004494087
σ_2^2	$3.306623e-05$	$3.393336e-05$

The model shows that the median[4] p value is 65%. This tells us that about 65% of the data will come from a "good" run and 35% from a "bad" run. However, p can be as low as 54% or as high as 73% as shown on Table 4.3.

Table 4.3: Quantiles for p

Quantiles for p				
0%	25%	50%	75%	100%
0.5413966	0.6411983	0.6558685	0.6712449	0.7331182

Figure 4.2 plots the original data coloring each observation according to the probability obtained on each position of the z vector. An observation that is colored closer to blue is associated with a higher probability of coming from a "good" run ($p = 1$). An observation that is colored closer to white is associated with a higher probability of coming from a "bad" run ($p = 0$).

Figure 4.2 also includes the lines resulting from the average of all the estimated y's that resulted from all pairs β_{01}, β_{11} and β_{02}, β_{12} using the equations $\hat{y} = b_{01} + xb_{11}$ and $\hat{y} = b_{02} + xb_{12}$ respectively.

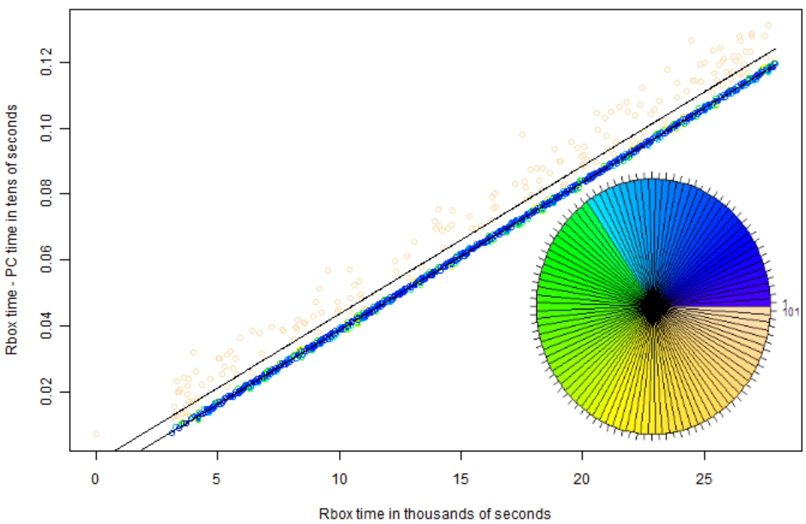

Figure 4.2: Differences between Rbox time and PC time.

[4]the median was chosen to show the quartiles for the distribution of p which would also give a sense of the distribution of the parameter.

The histograms for the realizations of β_{01} and β_{02} are shown on Figure 4.3. The white noise aspect shown on the time series plots presented on Figure 4.4 illustrates that convergence of the parameters was reached. The autocorrelation function plots presented on Figure 4.5 show that our choice of lag eliminated the autocorrelation among different parameter realizations (no lines stick out of the blue bands).

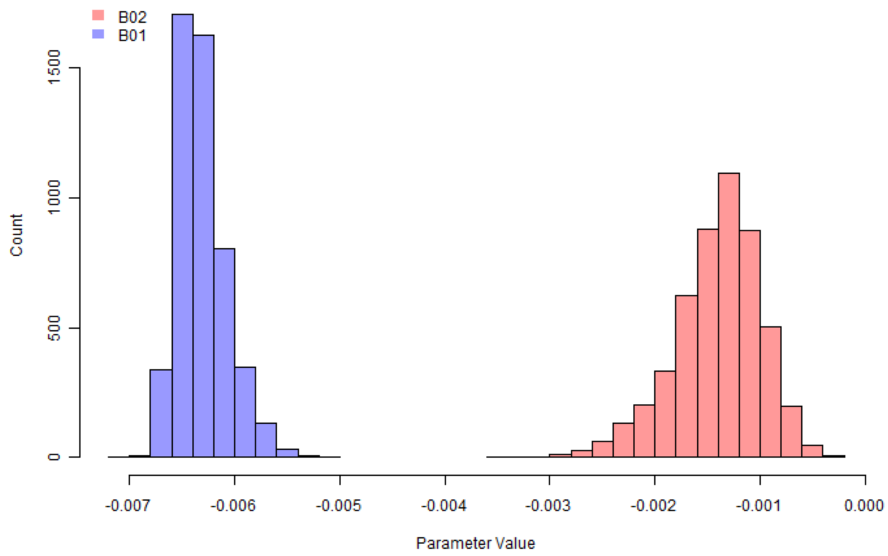

Figure 4.3: Histograms of the marginal posterior realizations for β_{01} and β_{02}.

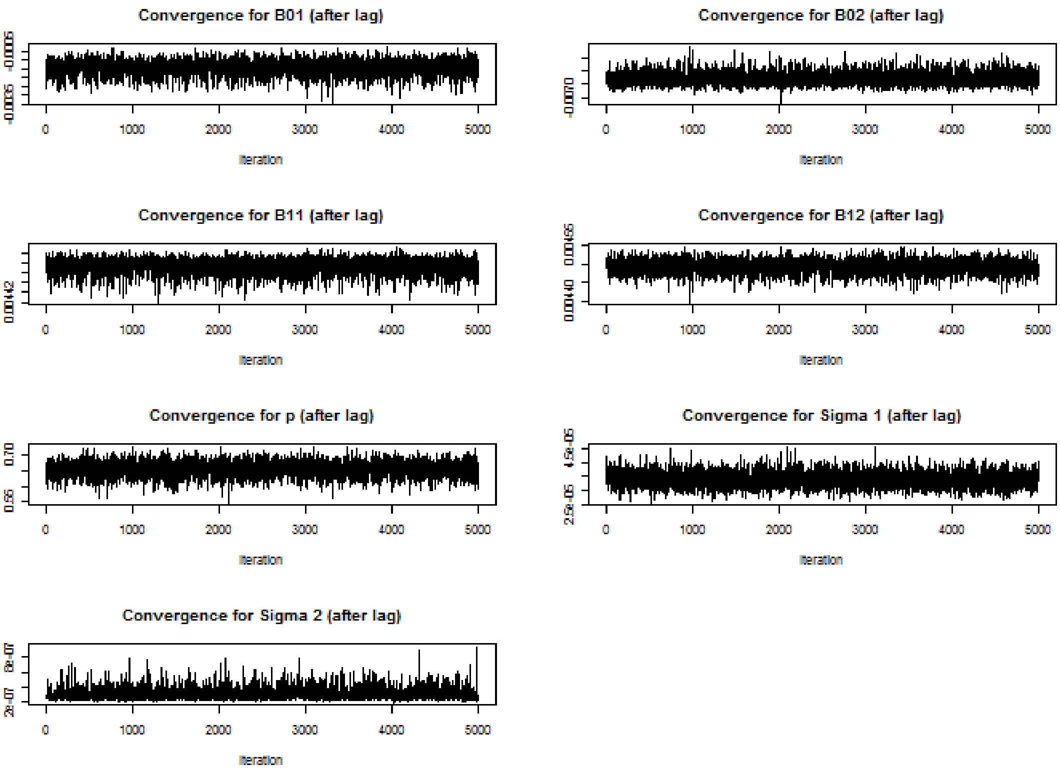

Figure 4.4: Convergence of the Parameters.

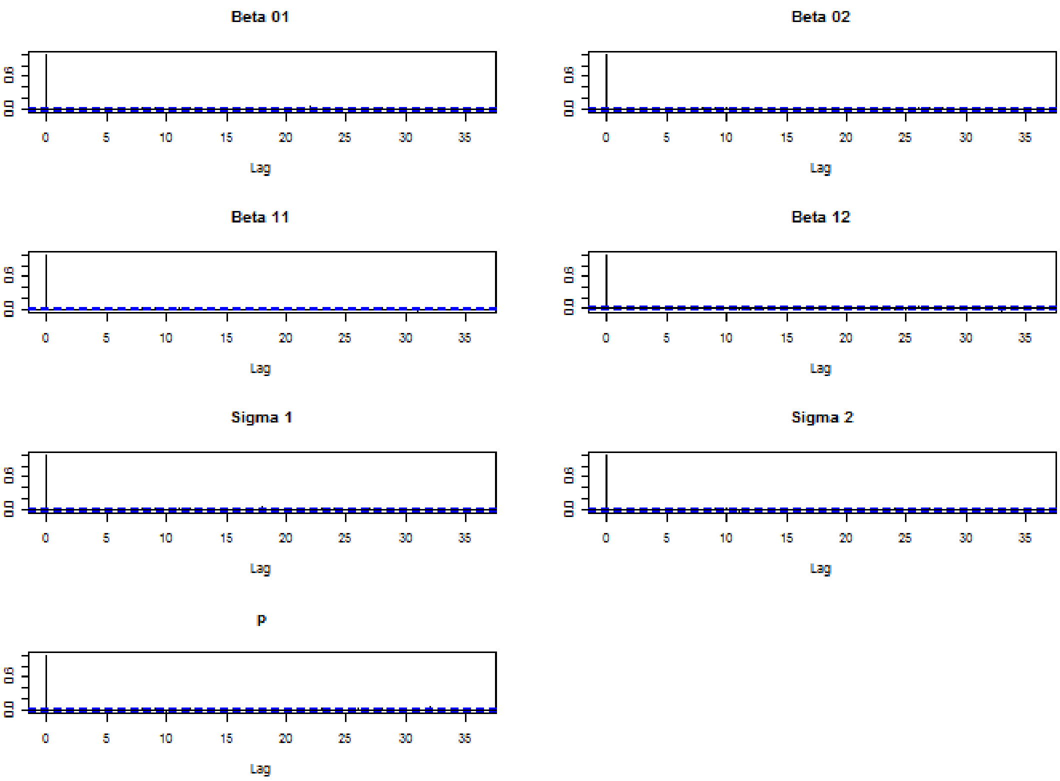

Figure 4.5: ACF for the Parameters.

Chapter 5

Conclusion and Future Work

Using latent variables, we designed a theoretical Bayesian Normal Mixture Model with two components. Then we tested the Bayesian Normal Mixture Model experimentally on a process with two states "good" and "bad" that resulted in the assignation of a probability of around 65% of occurrence to the "good" state of the phenomenon. The model performed better computationally by including prior distributions for the intercepts that were based on the data. This prevents the mixture model from degenerating into a single (one-component) regression model.

The model suggests that for the Rbox and PC time measurement dataset there are two processes: One process has larger variance than the other ($2.990065e-07$, $3.393336e-05$), the two processes have similar slopes (0.004493892, 0.004494087) and the two processes have different intercepts (-0.006315958 ,-0.001409068).

The expected difference between PC time and Rbox time in our application is a linear function of Rbox time. With an estimated 65% chance, we have that an observation can be associated with the equation: (PC time - Rbox Time) $= -0.006315958+$ (Rbox time) 0.004493892 an occurrence of a "good" measurement. With an estimated 35% we have that an observation can be associated with the equation (PC time -Rbox Time) $= -0.001409068+$ (Rbox time) 0.004494087 an occurrence of a "bad" measurement.

For future work an alternative approach to our model is to address the "label switching problem" by using a different route than the informative prior distributions given by the empirical Bayes analysis.

A generalization of this model that will include more than two components would also improve the applicability of the model.

Bibliography

G.E.P. Box and G.C. Tiao. A bayesian approach to some outlier problems. *Biometrika*, 55(1), 1968.

Bradley P. Carlin and Thomas A. Louis. *Bayesian Methods for Data Analysis.* Chapman & Hall/CRC, third edition, 2008.

Cody S. Ding. Practical assessment. *Research and Evaluation*, 11(11), 2006.

Andrew C. Gelman, John B. Stern, Hal S. Rubin, and Donald B. *Bayesian Data Analysis.* Chapman & Hall/CRC, second edition, 2004.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distribution, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741, November 1984.

Joseph F. Hair, Bill Black, Barry Babin, Rolph E. Anderson, and Ronald L. Tatham. *Multivariate Data Analysis.* Prentice hall, sixth edition, 2005.

H. Jeffreys. *Theory of Probability.* Oxford University Press, third edition, 1961.

Bengt O. Muthén. *New Developments and Techniques in Structural Equation Modeling*, chapter Latent Variable Mixture Modeling, pages 1–33. Lawrence Erlbaum Associates, Inc, first edition, 2001.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2010. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Matthew Stephens. Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society*, 62(4):795–809, 2000.

Wikipedia, The Free Encyclopedia. K-means clustering, 2011. URL `http://en.wikipedia.org/wiki/K-means_clustering`. [Online; accessed 24-April-2011].

Appendix A

Two component Normal mixture distribution

The following is a two component normal mixture distribution:

$$\begin{aligned} f(y \mid \boldsymbol{x}) =& p\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)\exp\left\{-\frac{[y-(\beta_{01}+\beta_{11}x)]^2}{2\sigma_1^2}\right\}+ \\ &(1-p)\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)\exp\left\{-\frac{[y-(\beta_{02}+\beta_{12}x)]^2}{2\sigma_2^2}\right\} \end{aligned} \tag{A.1}$$

Using latent indicators z_i, the joint distribution of the dependent variables $\{y_1, y_2, \cdots, y_n\}$ and indicators $\{z_1, z_2, \cdots, z_n\}$ is[1]:

$$\begin{aligned} \pi(y \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} &\left\{p\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)\exp\left\{-\frac{[y_i-(\beta_{01}+\beta_{11}x_i)]^2}{2\sigma_1^2}\right\}\right\}^{1-z_i} \cdot \\ &\left\{(1-p)\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)\exp\left\{-\frac{[y_i-(\beta_{02}+\beta_{12}x_i)]^2}{2\sigma_2^2}\right\}\right\}^{z_i} \end{aligned} \tag{A.2}$$

The joint priors are:

[1] Using $\boldsymbol{\theta} = \{\mathbf{z}, p, \beta_{01}, \beta_{11}, \sigma_1^2, \beta_{02}, \beta_{12}, \sigma_2^2\}$ and $\boldsymbol{\theta}_{-\boldsymbol{\sigma_1^2}} = \{\mathbf{z}, p, \beta_{01}, \beta_{11}, \beta_{02}, \beta_{12}, \sigma_2^2\}$ to shorten the notation

$$g(\boldsymbol{\theta}) \propto \frac{1}{\sqrt{2\pi v_0}} \exp\left\{-\frac{1}{2v_0}\left(\beta_{01}-\mu_0\right)^2\right\} \cdot \frac{1}{\sqrt{2\pi v_1}} \exp\left\{-\frac{1}{2v_1}\left(\beta_{02}-\mu_1\right)^2\right\} \cdot \frac{1}{\sigma_1^2} \cdot \frac{1}{\sigma_2^2} \tag{A.3}$$

Combining A.2 and A.3 via multiplication gives the posterior distribution for $\boldsymbol{\theta}$:

$$\pi(\boldsymbol{\theta}) \propto \pi(y \mid \boldsymbol{\theta}) \cdot g(\boldsymbol{\theta}) \tag{A.4}$$

A.0.1 Derivation of the Full Conditional Distributions for z_i

From equation (A.2) we can identify the marginal distribution for the z_i parameters to be:

Let

$$A = p\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right) \exp\left\{-\frac{[y_i-(\beta_{01}+\beta_{11}x_i)]^2}{2\sigma_1^2}\right\}$$

$$B = (1-p)\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right) \exp\left\{-\frac{[y_i-(\beta_{02}+\beta_{12}x_i)]^2}{2\sigma_2^2}\right\}$$

$$\boxed{z_i \mid \boldsymbol{\theta}_{-\mathbf{z}}, \mathbf{x}, \mathbf{y} \sim \text{Bern}\left(\tfrac{A}{A+B}\right)}$$

Now re-writing (A.2) we obtain:

$$\begin{aligned} =&p^{n-\sum\limits_{i\ni z_i=1} z_i}\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^{n-\sum\limits_{i\ni z_i=1} z_i} \exp\left\{-\sum_{i\ni z_i=1}\frac{[y_i-(\beta_{01}+\beta_{11}x_i)]^2}{2\sigma_1^2}\right\} \cdot \\ &(1-p)^{\sum\limits_{i\ni z_i=1} z_i}\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^{\sum\limits_{i\ni z_i=1} z_i} \exp\left\{-\sum_{i\ni z_i=1}\frac{[y_i-(\beta_{02}+\beta_{12}x_i)]^2}{2\sigma_2^2}\right\} \end{aligned} \tag{A.5}$$

Calling A to

$$\exp\left\{-\sum_{i\ni z_i=0}\frac{[y_i-(\beta_{01}+\beta_{11}x_i)]^2}{2\sigma_1^2}\right\}$$

and B to

$$\exp\left\{-\sum_{i\ni z_i=1}\frac{[y_i-(\beta_{02}+\beta_{12}x_i)]^2}{2\sigma_2^2}\right\}$$

Expanding A we get:

$$=\exp\left\{-\sum_{i\ni z_i=0}\frac{y_i^2-2\beta_{01}y_i-2\beta_{11}y_ix_i+\beta_{01}^2+2\beta_{01}\beta_{11}x_i+\beta_{11}^2x_i^2}{2\sigma_1^2}\right\}$$

A.0.2 Derivation of the Full Conditional Distributions for β_{01} and β_{02}

Now the marginal density function for β_{01} can be found by re-expressing A as:

$$\begin{aligned}
&=\exp\left\{-\sum_{i\ni z_i=0}\frac{-2\beta_{01}y_i+\beta_{01}^2+2\beta_{01}\beta_{11}x_i+y_i^2-2\beta_{11}y_ix_i+\beta_{11}^2x_i^2}{2\sigma_1^2}\right\}\\
&\propto\exp\left\{-\frac{\left[-2\beta_{01}\sum_{i\ni z_i=0}y_i+n_0\beta_{01}^2+2\beta_{01}\beta_{11}\sum_{i\ni z_i=0}x_i\right]}{2\sigma_1^2}\right\}\\
&=\exp\left\{-\frac{\left[-2\beta_{01}\frac{\sum_{i\ni z_i=0}y_i}{n_0}+\beta_{01}^2+2\beta_{01}\beta_{11}\frac{\sum_{i\ni z_i=0}x_i}{n_0}\right]}{\frac{2\sigma_1^2}{n_0}}\right\}
\end{aligned}\tag{A.6}$$

Calling $\overline{y}_0=\frac{\sum_{i\ni z_i=0}y_i}{n_0}$ and $\overline{x}_0=\frac{\sum_{i\ni z_i=0}x_i}{n_0}$ we can write the previous equation as:

$$\begin{aligned}
&=\exp\left\{-\frac{\left[\beta_{01}^2-2\beta_{01}\left(\overline{y}_0-\beta_{11}\overline{x}_0\right)-\left(\overline{y}_0-\beta_{11}\overline{x}_0\right)^2+\left(\overline{y}_0-\beta_{11}\overline{x}_0\right)^2\right]}{\frac{2\sigma_1^2}{n_0}}\right\}\\
&=\exp\left\{-\frac{\left[\beta_{01}-\left(\overline{y}_0-\beta_{11}\overline{x}_0\right)\right]^2}{\frac{2\sigma_1^2}{n_0}}\right\}
\end{aligned}\tag{A.7}$$

Now we can replace (A.7) in equation in (A.5).

$$
\begin{aligned}
& p^{n-\sum\limits_{i \ni z_i=1} z_i} \left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^{n-\sum\limits_{i \ni z_i=1} z_i} \exp\left\{-\sum_{i \ni z_i=0} \frac{[y_i - (\beta_{01} + \beta_{11}x_i)]^2}{2\sigma_1^2}\right\}. \\
& (1-p)^{\sum\limits_{i \ni z_i=1} z_i} \left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^{\sum\limits_{i \ni z_i=1} z_i} \exp\left\{-\sum_{i \ni z_i=1} \frac{[y_i - (\beta_{02} + \beta_{12}x_i)]^2}{2\sigma_2^2}\right\} \\
& \propto p^{n_0} \left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^{n_0} \exp\left\{-\frac{[\beta_{01} - (\overline{y}_0 - \beta_{11}\overline{x}_0)]^2}{\frac{2\sigma_1^2}{n_0}}\right\} \qquad \text{(A.8)}
\end{aligned}
$$

Finally we conclude that the full conditional distribution for β_{01} is:

$$
\boxed{\beta_{01} \mid \boldsymbol{\theta_{-\beta_{01}}}, \mathbf{x}, \mathbf{y} \sim N\left(\overline{y}_0 - \beta_{11}\overline{x}_0, \frac{\sigma_1^2}{n_0}\right)}
$$

Similarly for β_{02} we obtained that

$$
\begin{aligned}
& p^{n-\sum\limits_{i \ni z_i=1} z_i} \left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^{n-\sum\limits_{i \ni z_i=1} z_i} \exp\left\{-\sum_{i \ni z_i=0} \frac{[y_i - (\beta_{01} + \beta_{11}x_i)]^2}{2\sigma_1^2}\right\}. \\
& (1-p)^{\sum\limits_{i \ni z_i=1} z_i} \left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^{\sum\limits_{i \ni z_i=1} z_i} \exp\left\{-\sum_{i \ni z_i=1} \frac{[y_i - (\beta_{02} + \beta_{12}x_i)]^2}{2\sigma_2^2}\right\} \\
& \propto (1-p)^{n_1} \left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^{n_1} \exp\left\{-\frac{[\beta_{02} - (\overline{y}_1 - \beta_{12}\overline{x}_1)]^2}{\frac{2\sigma_2^2}{n_1}}\right\} \qquad \text{(A.9)}
\end{aligned}
$$

and we conclude that the full conditional distribution for β_{02} is::

$$
\boxed{\beta_{02} \mid \boldsymbol{\theta_{-\beta_{02}}}, \mathbf{x}, \mathbf{y} \sim N\left(\overline{y}_1 - \beta_{12}\overline{x}_1, \frac{\sigma_2^2}{n_1}\right)}
$$

Informative Prior Distributions

After running the model with uniform prior distributions for β_{01} and β_{02} normal informative prior distributions were used for β_{01} and β_{02} (the parameters that represent the intercepts) [2]. These procedure was done by dividing the data in

[2]When we applied our model to our dataset, we faced a common problem on Bayesian mixture models known as the "label switching problem". This problem is caused by the symmetry in the likelihood of the model parameters. To address this issue we used empirical Bayes to set the values of informative prior distributions for our β_{01} and β_{02} parameters. For a description of the "label switching problem" see Stephens (2000)

two groups using the k-means clustering method [3]. The two resulting groups are presented in figure A.1.

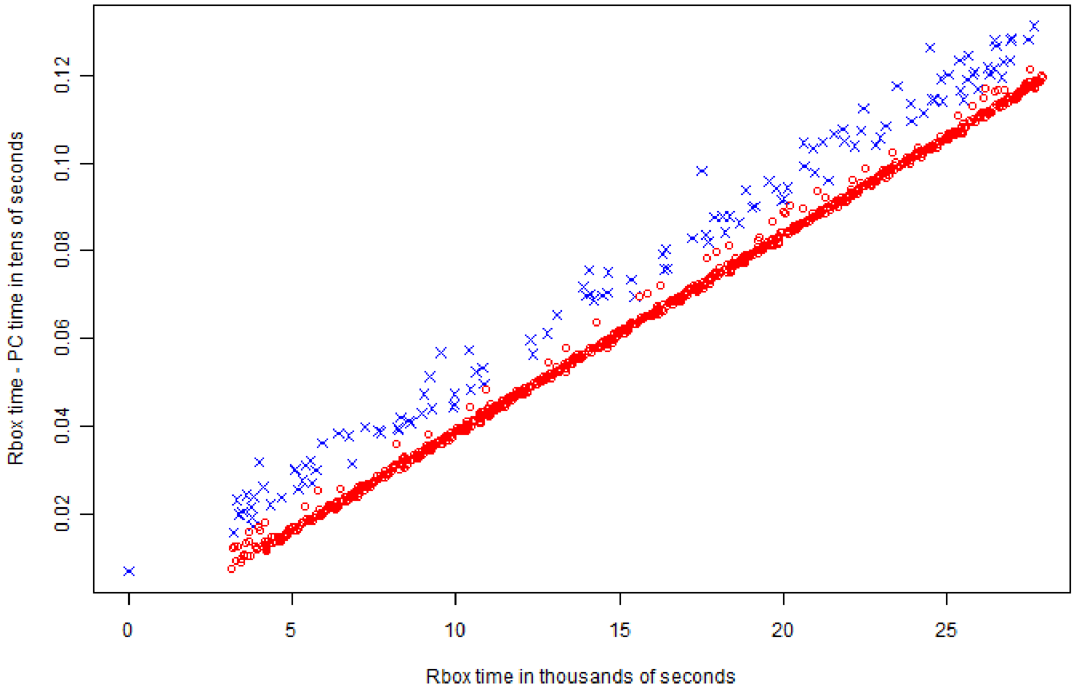

Figure A.1: Cluster separation for time differences between Rbox and PC

On each one of these two groups of data we fit a simple linear regression model. We used the intercept as the mean for our prior distributions and the standard error of the intercept as our variance for the prior distributions of our β_{01} and β_{02} parameters. Mathematically the derivation is as follows:

Letting

$$R_i = \overline{y}_i - \beta_{1j}\overline{x}_i \text{ for } i = 0,\ j = 1 \text{ } and \text{ } , i = 1,\ j = 2$$

and letting the mean of the top cloud be μ_0 and the standard error for top cloud be v_0. Similarly for the bottom cloud the mean is defined to be μ_1 and the

[3] Given a set of observations $(x_1, x_2, \cdots, x_n)$ where each observation is a d-dimensional real vector, k-means aims to partitions the n observations into k sets $(k \leq n)$ $S = \{S_1, S_2, \cdots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS) $arg \min \sum_{i=1}^{k} \sum_{x_j \in S_i} \|x_j - \mu_i\|^2$ where μ_i is the mean of points in S_i Hair et al. (2005)

standard error v_1

then:

$$\underbrace{\frac{1}{\sqrt{2\pi\frac{\sigma_j^2}{n_i}}}\exp\left\{-\frac{n_i}{2\sigma_j^2}\left(\beta_{0j}-R_i\right)^2\right\}}_{\text{Posterior}}\cdot\underbrace{\frac{1}{\sqrt{2\pi v_i}}\exp\left\{-\frac{1}{2v_i}\left(\beta_{0j}-\mu_i\right)^2\right\}}_{\text{Prior}}$$

$$\propto\exp\left\{-\frac{n_i}{2\sigma_j^2}\left(\beta_{0j}^2-2\beta_{0j}R+R^2\right)-\frac{1}{v_i}\left(\beta_{0j}^2-2\beta_{0j}\mu_i+\mu_i^2\right)\right\}$$

$$\propto\exp\left\{-\frac{1}{2}\left[\beta_{0j}^2\left(\frac{n_i}{\sigma_j^2}\frac{1}{v_i}\right)-2\beta_{0j}\left(\frac{Rn_i}{\sigma_j^2}+\frac{\mu_i}{v_i}\right)\right]\right\}$$

$$\propto\exp\left\{-\frac{1}{2}\left(\frac{n_i}{\sigma_j^2}\frac{1}{v_i}\right)\left[\beta_{0j}^2-2\beta_{0j}\frac{\left(\frac{Rn_i}{\sigma_j^2}+\frac{\mu_i}{v_i}\right)}{\left(\frac{n_i}{\sigma_j^2}\frac{1}{v_i}\right)}\right]\right\}$$

$$\propto\exp\left\{-\frac{1}{2}\left(\frac{n_i}{\sigma_j^2}\frac{1}{v_i}\right)\left[\beta_{0j}^2-2\beta_{0j}\left(\frac{Rn_iv_i+\mu_i\sigma_j^2}{n_iv_i+\sigma_j^2}\right)\right]\right\}$$

which implies that:

$$\boxed{\beta_{0j}\mid\boldsymbol{\theta}_{-\boldsymbol{\beta_{0j}}},\mathbf{x},\mathbf{y}\sim N\left(\left(\frac{R_in_iv_i+\mu_i\sigma_j^2}{n_iv_i+\sigma_j^2}\right),\left(\frac{n_i}{\sigma_j^2}\frac{1}{v_i}\right)^{-1}\right)}$$

A.0.3 Derivation of the Full Conditional Distributions for β_{11} and β_{12}

The marginal density function for β_{11} can be found by re-expressing A and letting

$$C=\left[\frac{\sum_{i\ni z_i=0}y_ix_i-\left(\beta_{01}\sum_{i\ni z_i=0}x_i\right)}{\sum_{i\ni z_i=0}x_i^2}\right]$$

as shown bellow:

$$
\begin{aligned}
A &= \exp\left\{-\sum_{i\ni z_i=0}\frac{y_i^2-2\beta_{01}y_i-2\beta_{11}y_ix_i+\beta_{01}^2+2\beta_{01}\beta_{11}x_i+\beta_{11}^2x_i^2}{2\sigma_1^2}\right\}\\
&= \exp\left\{-\frac{-2\beta_{11}\sum\limits_{i\ni z_i=0}y_ix_i+2\beta_{01}\beta_{11}\sum\limits_{i\ni z_i=0}x_i+\beta_{11}^2\sum\limits_{i\ni z_i=0}x_i^2}{2\sigma_1^2}\right\}\cdot\\
&\quad \exp\left\{\frac{\sum\limits_{i\ni z_i=0}y_i^2-2\beta_{01}\sum\limits_{i\ni z_i=0}y_i+n\beta_{01}^2}{2\sigma_1^2}\right\}\\
&\propto \exp\left\{-\frac{-2\beta_{11}\sum\limits_{i\ni z_i=0}y_ix_i+2\beta_{01}\beta_{11}\sum\limits_{i\ni z_i=0}x_i+\beta_{11}^2\sum\limits_{i\ni z_i=0}x_i^2}{2\sigma_1^2}\right\}\\
&= \exp\left\{-\frac{\beta_{11}^2-2\beta_{11}C-C^2+C^2}{\frac{2\sigma_1^2}{\sum\limits_{i\ni z_i=0}x_i^2}}\right\}\\
&= \exp\left\{-\frac{\left[\beta_{11}-\left(\frac{\sum\limits_{i\ni z_i=0}y_ix_i}{\sum\limits_{i\ni z_i=0}x_i^2}-\beta_{01}\frac{\sum\limits_{i\ni z_i=0}x_i}{\sum\limits_{i\ni z_i=0}x_i^2}\right)\right]^2}{\frac{2\sigma_1^2}{\sum\limits_{i\ni z_i=0}x_i^2}}\right\}
\end{aligned}
\tag{A.10}
$$

Finally we can replace (A.10) in equation in (A.5).

$$
\begin{aligned}
&p^{n-\sum\limits_{i\ni z_i=0}z_i}\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^{n-\sum\limits_{i\ni z_i=0}z_i}\exp\left\{-\sum_{i\ni z_i=0}\frac{[y_i-(\beta_{01}+\beta_{11}x_i)]^2}{2\sigma_1^2}\right\}\cdot\\
&(1-p)^{\sum\limits_{i\ni z_i=1}z_i}\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^{\sum\limits_{i\ni z_i=1}z_i}\exp\left\{-\sum_{i\ni z_i=1}\frac{[y_i-(\beta_{02}+\beta_{12}x_i)]^2}{2\sigma_2^2}\right\}\\
&\propto (p)^{n_0}\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^{n_0}\exp\left\{-\frac{\left[\beta_{11}-\left(\frac{\sum\limits_{i\ni z_i=0}y_ix_i}{\sum\limits_{i\ni z_i=0}x_i^2}-\beta_{01}\frac{\sum\limits_{i\ni z_i=0}x_i}{\sum\limits_{i\ni z_i=0}x_i^2}\right)\right]^2}{\frac{2\sigma_1^2}{\sum\limits_{i\ni z_i=0}x_i^2}}\right\}
\end{aligned}
$$

Then we conclude that the marginal distribution for the coefficient β_{11} is distributed as:

$$
\boxed{\beta_{11}\mid\boldsymbol{\theta}_{-\boldsymbol{\beta_{11}}},\mathbf{x},\mathbf{y}\sim N\left(\frac{\sum\limits_{i\ni z_i=0}y_ix_i}{\sum\limits_{i\ni z_i=0}x_i^2}-\beta_{01}\frac{\sum\limits_{i\ni z_i=0}x_i}{\sum\limits_{i\ni z_i=0}x_i^2},\frac{\sigma_1^2}{\sum\limits_{i\ni z_i=0}x_i^2}\right)}
$$

Similarly we found for β_{12}:

$$p^{n-\sum\limits_{i\ni z_i=0} z_i}\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^{n-\sum\limits_{i\ni z_i=0} z_i}\exp\left\{-\sum_{i\ni z_i=0}\frac{[y_i-(\beta_{01}+\beta_{11}x_i)]^2}{2\sigma_1^2}\right\}\cdot$$
$$(1-p)^{\sum\limits_{i\ni z_i=1} z_i}\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^{\sum\limits_{i\ni z_i=1} z_i}\exp\left\{-\sum_{i\ni z_i=1}\frac{[y_i-(\beta_{02}+\beta_{12}x_i)]^2}{2\sigma_2^2}\right\}$$
$$\propto(1-p)^{n_1}\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^{n_1}\exp\left\{-\frac{\left[\beta_{12}-\left(\frac{\sum\limits_{i\ni z_i=1} y_ix_i}{\sum\limits_{i\ni z_i=1} x_i^2}-\beta_{02}\frac{\sum\limits_{i\ni z_i=1} x_i}{\sum\limits_{i\ni z_i=1} x_i^2}\right)\right]^2}{\frac{2\sigma_2^2}{\sum\limits_{i\ni z_i=1} x_i^2}}\right\} \quad \text{(A.11)}$$

Then we conclude that the marginal distribution for the coefficient β_{12} is distributed as:

$$\boxed{\beta_{12}\mid\boldsymbol{\theta}_{-\boldsymbol{\beta_{12}}},\mathbf{x},\mathbf{y}\sim N\left(\frac{\sum\limits_{i\ni z_i=1} y_ix_i}{\sum\limits_{i\ni z_i=1} x_i^2}-\beta_{02}\frac{\sum\limits_{i\ni z_i=1} x_i}{\sum\limits_{i\ni z_i=1} x_i^2},\frac{\sigma_2^2}{\sum\limits_{i\ni z_i=1} x_i^2}\right)}$$

A.0.4 Derivation of the Full Conditional Distributions for σ_1 and σ_2

The marginal distribution for σ_1 can be found in the following way:

$$p^{n-\sum\limits_{i\ni z_i=1} z_i}\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^{n-\sum\limits_{i\ni z_i=1} z_i}\exp\left\{-\sum_{i\ni z_i=0}\frac{[y_i-(\beta_{01}+\beta_{11}x_i)]^2}{2\sigma_1^2}\right\}\cdot$$
$$(1-p)^{\sum\limits_{i\ni z_i=1} z_i}\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^{\sum\limits_{i\ni z_i=1} z_i}\exp\left\{-\sum_{i\ni z_i=1}\frac{[y_i-(\beta_{02}+\beta_{12}x_i)]^2}{2\sigma_2^2}\right\}$$
$$\propto p^{n_0}\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^{n_0}\exp\left\{-\sum_{i\ni z_i=0}\frac{[y_i-(\beta_{01}+\beta_{11}x_i)]^2}{2\sigma_1^2}\right\}$$
$$\propto(\sigma_1)^{-n_0}\exp\left\{-\sum_{i\ni z_i=0}\frac{[y_i-(\beta_{01}+\beta_{11}x_i)]^2}{2\sigma_1^2}\right\} \quad \text{(A.12)}$$

with a prior distribution of $1/\sigma_1^2$ then we have that (A.12) is proportional to:

$$\propto(\sigma_1^2)^{-(\frac{n_0}{2}+1)}\exp\left\{-\sum_{i\ni z_i=0}\frac{[y_i-(\beta_{01}+\beta_{11}x_i)]^2}{2\sigma_1^2}\right\} \quad \text{(A.13)}$$

from (A.13) we find σ_1 distribution to be:

$$\boxed{\sigma_1^2 \mid \boldsymbol{\theta}_{-\boldsymbol{\sigma_1^2}}, \mathbf{x}, \mathbf{y} \sim \text{Inv-Gamma}\left(\tfrac{n_0}{2}, \sum_{i \ni z_i=0} \tfrac{[y_i-(\beta_{01}+\beta_{11}x_i)]^2}{2}\right)}$$

Similarly for σ_2 we obtained its marginal distribution in the following way:

$$\begin{aligned}
&p^{n-\sum_{i \ni z_i=0} z_i}\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^{n-\sum_{i \ni z_i=0} z_i} \exp\left\{-\sum_{i \ni z_i=0}\frac{[y_i-(\beta_{01}+\beta_{11}x_i)]^2}{2\sigma_1^2}\right\}\cdot \\
&(1-p)^{\sum_{i \ni z_i=1} z_i}\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^{\sum_{i \ni z_i=1} z_i} \exp\left\{-\sum_{i \ni z_i=1}\frac{[y_i-(\beta_{02}+\beta_{12}x_i)]^2}{2\sigma_2^2}\right\} \\
&\propto (1-p)^{n_1}\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^{n_1} \exp\left\{-\sum_{i \ni z_i=1}\frac{[y_i-(\beta_{02}+\beta_{12}x_i)]^2}{2\sigma_2^2}\right\} \\
&\propto (\sigma_2)^{-n_1} \exp\left\{-\sum_{i \ni z_i=1}\frac{[y_i-(\beta_{02}+\beta_{12}x_i)]^2}{2\sigma_2^2}\right\} \qquad \text{(A.14)}
\end{aligned}$$

with a prior distribution of $1/\sigma_2^2$ then we have that (A.14) is proportional to:

$$\propto \left(\sigma_2^2\right)^{-(\frac{n_1}{2}+1)} \exp\left\{-\sum_{i \ni z_i=1}\frac{[y_i-(\beta_{02}+\beta_{12}x_i)]^2}{2\sigma_2^2}\right\} \qquad \text{(A.15)}$$

from (A.15) we find σ_2 distribution to be:

$$\boxed{\sigma_2^2 \mid \boldsymbol{\theta}_{-\boldsymbol{\sigma_2^2}}, \mathbf{x}, \mathbf{y} \sim \text{Inv-Gamma}\left(\tfrac{n_1}{2}, \sum_{i \ni z_i=1} \tfrac{[y_i-(\beta_{02}+\beta_{12}x_i)]^2}{2}\right)}$$

A.0.5 Derivation of the Full Conditional Distributions for p

The marginal posterior distribution for the parameter p can be obtained directly recalling equation (A.5)

$$\begin{aligned}
&p^{n-\sum_{i \ni z_i=1} z_i}\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)^{n-\sum_{i \ni z_i=1} z_i} \exp\left\{-\sum_{i \ni z_i=0}\frac{[y_i-(\beta_{01}+\beta_{11}x_i)]^2}{2\sigma_1^2}\right\}\cdot \\
&(1-p)^{\sum_{i \ni z_i=1} z_i}\left(\frac{1}{\sqrt{2\pi\sigma_2^2}}\right)^{\sum_{i \ni z_i=1} z_i} \exp\left\{-\sum_{i \ni z_i=1}\frac{[y_i-(\beta_{02}+\beta_{12}x_i)]^2}{2\sigma_2^2}\right\}
\end{aligned} \qquad \text{(A.16)}$$

From (A.5) we conclude that:

$$p \sim \text{Beta}\left(1 + \sum_{i \ni z_i = 1} z_i, 1 + n - \sum_{i \ni z_i = 1} z_i\right)$$

$$\boxed{p \mid \boldsymbol{\theta}_{\mathbf{-p}}, \mathbf{x}, \mathbf{y} \sim \text{Beta}\left(1 + n_1, 1 + n_0\right)}$$

Appendix B

R-Code: Two Component Normal Mixture Distribution

```
library(zoo)
rm(list=ls(all=TRUE))
detach(Mydata)
## Paths for the folders to retrieve data and store outputs
PATH='...'
PATH1='...'
NAME='date.png' #Common Name of the files

Mydata<-read.table(paste(PATH1,
"timing.txt",sep=""), header=T)
attach(Mydata)

y<-(Difference/10000)
x<-(Rboxtime/1000000)

## Detrending the cluster
m<-lm(y~x)
detr<-zoo(resid(m),x)

d<-data.frame(x,ydet=detr)

### K-means clustering
fit <- kmeans(d$y, 2)

## Saving the file
png(paste(PATH,"Clusters",NAME,sep=""),
width=700, height=500)
```

```
plot(x[fit$cluster==2],y[fit$cluster==2],
col='blue', ann=F, pch=4)
points(x[fit$cluster==1],y[fit$cluster==1],
col='red', ann=F, pch=1)
title(main = "", ylab = "Rbox time -
PC time in tens of seconds",
xlab="Rbox time in thousands of seconds")
dev.off()

## New values for the clusters
fit$cluster[fit$cluster==1] <- 0
fit$cluster[fit$cluster==2] <- 1
z<-fit$cluster

lm.c0<-lm(y[z==0]~x[z==0])
lm.c1<-lm(y[z==1]~x[z==1])

### Parameter initialization
## Applying the coefficients of the simple linear regression
## to the parameters of the model.
#bottom cloud with less variance
b01<- summary(lm.c0)$coeff[1,1] #-0.006
#Top cloud with more variance
b02<- summary(lm.c1)$coeff[1,1]#0.005

b11<-summary(lm.c0)$coeff[2,1] #0.0045
b12<-summary(lm.c1)$coeff[2,1] #0.0045

#bottom cloud with less variance
m01<-summary(lm.c0)$coeff[1,1]
#bottom cloud with less variance
v01<-summary(lm.c0)$coeff[1,2]
#Top cloud with more variance
m02<-summary(lm.c1)$coeff[1,1]
#Top cloud with more variance
v02<-summary(lm.c1)$coeff[1,2]

n<-length(x)
be<-0

n0 <- sum(1-z)
n1 <- n-n0
Z01<-0
Z02<-0



##it is already sigma1 squared
```

```
#bottom cloud with less variance
sigma1<- (0.000486657)^2

##it is already sigma2 squared
#top cloud with more variance
sigma2<- (0.005750324)^2

p<-0

##Vector definition
b01vec<-NULL
b02vec<-NULL
b11vec<-NULL
b12vec<-NULL
svec1<-NULL
svec2<-NULL
pvec<-NULL
zs<-NULL
zprob<-NULL

##Values of the Lag, burn and total number of iterations
lagg<-30
burn<-1000
TOTAL<-5000



p <- rbeta(1,1+n1,1+n0)


for(i in 1:(lagg*TOTAL+burn))
{

n0<-sum(1-z)
n1<-n-n0

##calculations for b01
R<-(sum((1-z)*y)/n0)-(b11*sum((1-z)*x)/n0)
b01<-rnorm(1,(R*n0*v01+m01*sigma1^2)/
(n0*v01+sigma1^2),((n0/sigma1^2)+(1/v01))^(-1))

##calculations for b02
K<-(sum(z*y)/n1)-(b12*sum(z*x)/n1)
b02<-rnorm(1,(K*n1*v02+m02*sigma2^2)/
(n1*v02+sigma2^2),((n1/sigma2^2)+(1/v02))^(-1))
```

```
##calculations for b11
b11<-rnorm(1, (sum((1-z)*x*y)/sum((1-z)*x^2))-
b01*(sum((1-z)*x)/sum((1-z)*x^2)),sqrt(sigma1/sum((1-z)*x^2)))

##calculations for b12
b12<-rnorm(1, (sum((z)*x*y)/sum((z)*x^2))-
b02*(sum((z)*x)/sum((z)*x^2)),sqrt(sigma2/sum((z)*x^2)))

##calculations for sigma1
sigma1 <- 1/rgamma(1,n0/2,(sum((1-z)*(y-(b01+b11*x))^2))/2)


##calculations for sigma2
sigma2 <- 1/rgamma(1,n1/2,(sum((z)*(y-(b02+b12*x))^2))/2)


##calculations for p
p <- rbeta(1,1+n1,1+n0)
while(p < 0.09 && p>0.93){
p <- rbeta(1,1+n1,1+n0)
}

##calculations for z
exp1<-exp(-(0.5/sigma1)*(y-(b01+b11*x))^2)
exp2<-exp(-(0.5/sigma2)*(y-(b02+b12*x))^2)
zprob<-(p*(1/sqrt(sigma1))*exp1)/
(p*(1/sqrt(sigma1))*exp1 + (1-p)*(1/sqrt(sigma2))*exp2)
z<-rbinom(1000,1,zprob)

if(i%%lagg==0&&i>burn)
{
b01vec<-c(b01vec,b01)
b02vec<-c(b02vec,b02)
b11vec<-c(b11vec,b11)
b12vec<-c(b12vec,b12)
svec1<-c(svec1,sigma1)
svec2<-c(svec2,sigma2)
pvec<-c(pvec,p)
zs<-rbind(zs,z)
}
}

####optional features
tsplot<-function(vec,w){
n<-length(vec)
x<-c(1:n)
plot(x,vec, type="l",xlab="Iteration", main=w, ylab="")
title(font.main=4)
```

```
}


require(graphics)

## Creating bins for the coloring of the results
zprob<-NULL
bins<-NULL
bins<-seq(0,1, by=0.01)
for(i in 1:length(zs[1,]))
{
zprob<-c(zprob,mean(zs[,i]))
}
bins<-round(zprob*100)+1



##Scatter plot

png(paste(PATH,"Scatter-",NAME,sep=""),  width=700, height=500)
plot(x,y, col=rainbow(101)[bins], ann=F)
abline(mean(b01vec),mean(b11vec))
abline(mean(b02vec),mean(b12vec))
title(main = "", ylab = "Rbox time - PC time in tens of seconds",
xlab="Rbox time in thousands of seconds")
dev.off()

## TSplots
png(paste(PATH,"Convergence-",NAME,sep=""),  width=700, height=500)
par(mfrow=c(4,2))
tsplot(b01vec,"Convergence for B01 (after lag)")
tsplot(b02vec,"Convergence for B02 (after lag)")
tsplot(b11vec,"Convergence for B11 (after lag)")
tsplot(b12vec,"Convergence for B12 (after lag)")
tsplot(pvec,"Convergence for p (after lag)")
tsplot(svec1,"Convergence for Sigma 1 (after lag)")
tsplot(svec2,"Convergence for Sigma 2 (after lag)")
dev.off()

## ACF plots
png(paste(PATH,"ACF-",NAME,sep=""),  width=700, height=500)
par(mfrow=c(4,2))
acf(b01vec, main='Beta 01', ylab='')
acf(b02vec, main='Beta 02', ylab='')
acf(b11vec, main='Beta 11', ylab='')
acf(b12vec, main='Beta 12', ylab='')
```

```
acf(svec1, main='Sigma 1', ylab='')
acf(svec2, main='Sigma 2', ylab='')
acf(pvec, main='p', ylab='')
dev.off()

teta<-1:length(b01vec)
plot(teta, b01vec, xlim=range(min(teta), max(teta)),
ylim=range(min(b01vec), max(b02vec)))
points(teta,b02vec, col="red")


y1vec<-NULL
y2vec<-NULL

for(i in 1:TOTAL)
{
y1vec<-c(y1vec,b01vec[i]+b11vec[i]*x)
y2vec<-c(y2vec,b02vec[i]+b12vec[i]*x)
}
y1mat<-NULL
y1mat<-matrix(y1vec,5000,1000,byrow=TRUE)
y1mean<-apply(y1mat,2,mean)
y2mat<-NULL
y2mat<-matrix(y2vec,5000,1000,byrow=TRUE)
y2mean<-apply(y2mat,2,mean)


## Scatter Kern style
png(paste(PATH,"Kern-Scatter-",NAME,sep=""),
width=700, height=500)
plot(x,y, col=rainbow(101)[bins], ann=F)
lines(x,y1mean)
lines(x,y2mean)
title(main = "", ylab = "Rbox time - PC time in tens of seconds",
xlab="Rbox time in thousands of seconds")
dev.off()

# Plot of the values of the parameters B0i
png(paste(PATH,"B0i-Diffs-",NAME,sep=""),  width=700, height=500)
plot(c(1:length(b01vec)),b01vec, ann=F, ylim=c(-0.008,0.003))
points(c(1:length(b02vec)),b02vec, col='red' )
title(main = "B01 and B02", ylab = "Value", xlab="Rbox Time ")
dev.off()

# Data plot
png(paste(PATH,"data",NAME,sep=""),  width=700, height=500)
plot(x,y, ann=F)
title(main = "", ylab = "Rbox time - PC time in tens of seconds",
```

```
xlab="Rbox time in thousands of seconds")
dev.off()
```

Appendix C

Notation

Parameter	Meaning
y	Dependent variable (Pc time - Rbox time)
x	Independent variable (Rbox time)
β_{01}	Intercept of one cloud of data (bottom cloud)
β_{02}	Intercept of one cloud of data (top cloud)
β_{11}	Trend of one cloud of data (bottom cloud)
β_{12}	Trend of one cloud of data (top cloud)
σ_1^2	Variance of one cloud of data (bottom cloud)
σ_2^2	Variance of one cloud of data (top cloud)
p	Parameter that captures the proportion of observations on one cloud
z_i	Parameter that signals to which cloud a particular belongs.
n	Total number of observations
n_0	Sum of the observations for which $z_i = 0$
n_1	Sum of the observations for which $z_i = 1$
μ_0	Hyper-parameter that represents the mean intercept for the bottom cloud given by a ordinary linear regression ran on the data. Also this is the starting value for the β_{01} parameter

Parameter	Meaning
μ_1	Hyper-parameter that represents the mean intercept for the top cloud given by a ordinary linear regression ran on the data. Also this is the starting value for the β_{02} parameter
v_0	Hyper-parameter that represents the expected variance of the intercept for the bottom cloud given by a ordinary linear regression ran on the data. Also this is the starting value for the σ_1 parameter
v_1	Hyper-parameter that represents the expected variance of the intercept for the top cloud given by a ordinary linear regression ran on the data. Also this is the starting value for the σ_2 parameter
$\boldsymbol{\theta}$	A notation summary symbol that represents the parameters $\mathbf{z}, p, \beta_{01}, \beta_{11}, \sigma_1^2, \beta_{02}, \beta_{12}, \sigma_2^2\}$