Dataset 1 must include 500 trajectories of 6 time points generated from two distributions. In distribution 1, the outcome is distributed with mean that is a linear regression of time, with intercept term 2 and slope 5, and errors that follow a standard normal distribution. In distribution 2, the outcome is distributed with mean that is also a linear function of time with intercept term 5 and slope 7 and errors that follow a standard normal distribution. Generate the mixture of 500 trajectories assuming an average of 70% of data generated from distribution 1 and 30% generated from distribution 2. Plot the 500 trajectories and color the two groups differently.

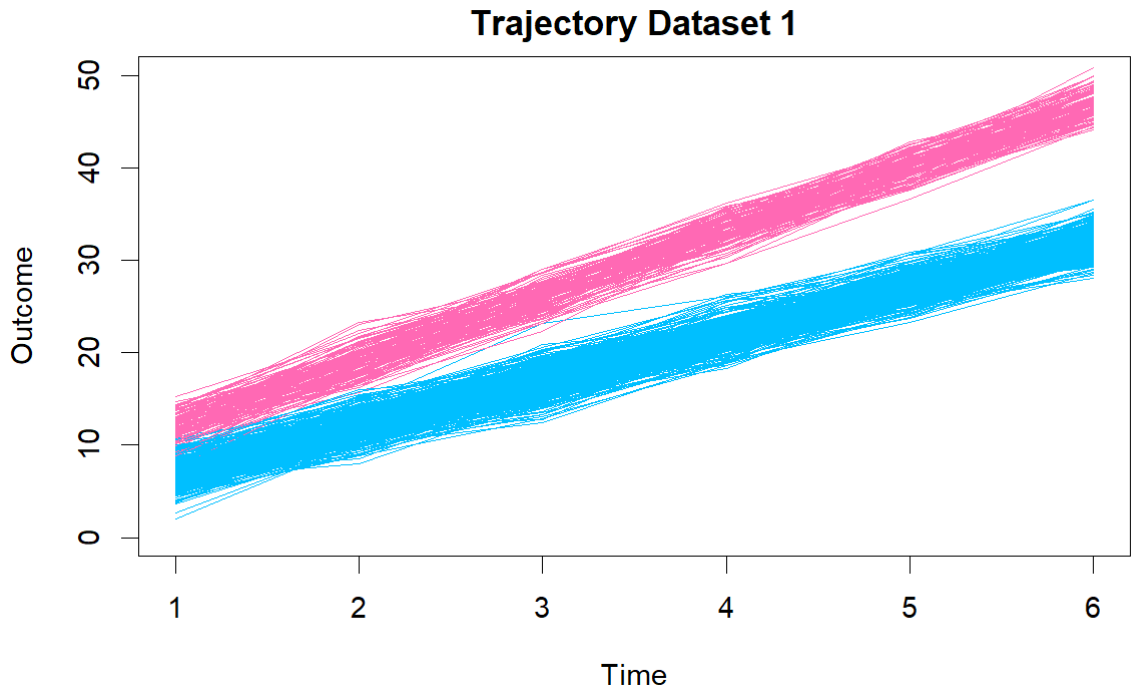
```
# Longitudinal Trajectories Dataset 1
```{r}
data1_beta0 <- c(2, 5)
data1_beta1 <- c(5, 7)
data1_alpha <- rnorm(500, 0, 1)
data1_distribution = 1 + rbinom(n=500, size=1, p=0.3)
table(data1_distribution)

data1_x <- c(1:6)
data1_y <- c()
data1_id <- c()

for(i in 1:500){
  data1_id <- rbind(data1_id, rep(i, 6))
  data1_y <- rbind(data1_y, c(data1_beta0[data1_distribution[i]] +
                                data1_beta1[data1_distribution[i]]*data1_x +
                                data1_alpha[i] + rnorm(6,0,1)
                              )
                   )
}

trajectory_data1 <- list(n_subjects=500, id=data1_id, x=data1_x, y=data1_y)
```
```

```
data1_distribution
  1   2
350 150
```



Dataset 2 includes 500 trajectories of 6 time points generated from three distributions. In distribution 1, the outcome is distributed with mean that is a linear regression of time, with intercept term 2 and slope 5, and errors that follow a standard normal distribution. In distribution 2, the outcome is distributed with mean that is also a linear function of time with intercept term 5 and slope 7 and errors that follow a standard normal distribution. Similarly, in distribution 3, the outcome is distributed with mean that is also a linear function of time with intercept term 6 and slope 2 and errors that follow a standard normal distribution. Generate the mixture of 500 trajectories assuming an average of 50% of data generated from distribution 1, 30% generated from distribution 2 and 20% generated from distribution 3.

```
# Longitudinal Trajectories Dataset 2
```{r}
data2_beta0 <- c(2, 5, 6)
data2_beta1 <- c(5, 7, 2)
data2_alpha <- rnorm(500, 0, 1)

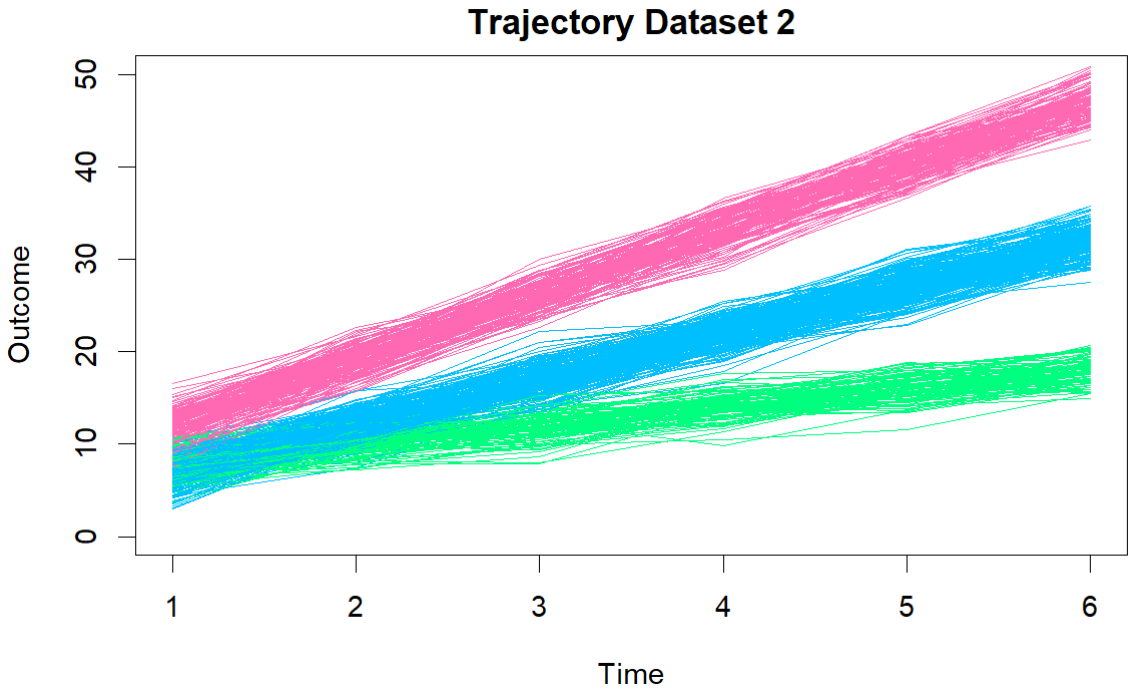
data2_generator = rmultinom(n=500, size=1, prob=c(0.5, 0.3, 0.2))
data2_distribution <- which(data2_generator==1, arr.ind=TRUE)[,1]
table(data2_distribution)

data2_x <- c(1:6)
data2_y <- c()
data2_id <- c()

for(i in 1:500){
  data2_id <- rbind(data2_id, rep(i, 6))
  data2_y <- rbind(data2_y, c(data2_beta0[data2_distribution[i]] +
                                data2_beta1[data2_distribution[i]]*data2_x +
                                data2_alpha[i] + rnorm(6,0,1)
                              )
                   )
}

trajectory_data2 <- list(n_subjects=500, id=data2_id, x=data2_x, y=data2_y)
```
```

```
data2_distribution
  1   2   3
249 151 100
```



Analyze both data sets using Bayesian model-based clustering, assuming only 2 clusters. Include some diagnostic plots, summary statistics and slope and intercept terms estimated for the two clusters. Compare the clusters assignment generated with MCMC and the true cluster label and discuss the accuracy of cluster membership estimated with the MCMC. Include a plot of the trajectories and color them based on the cluster label estimated with the MCMC.

Dataset 1

$$Cluster\ 1\ y = 2.0075 + 5.0530$$
$$Cluster\ 2\ y = 4.9933 + 6.9739$$

```
Iterations = 2001:3000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 1000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

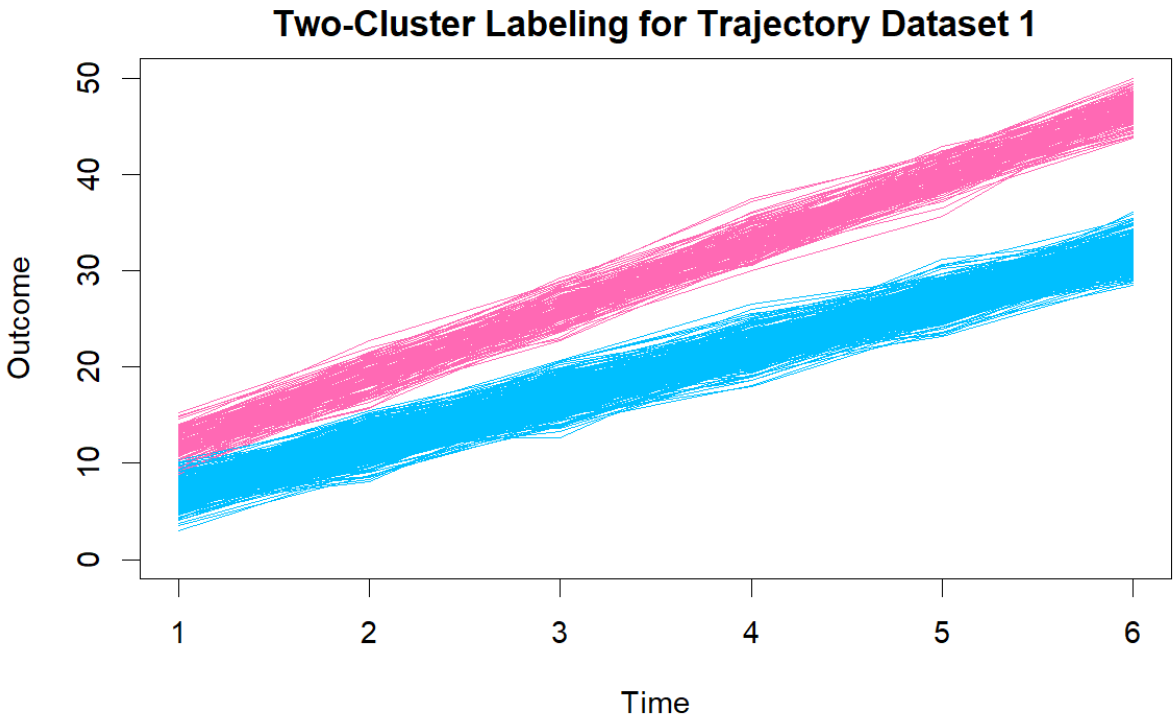
        Mean      SD  Naive SE Time-series SE
b0[1] 2.0051 0.07475 0.0023637      0.0094121
b0[2] 5.0501 0.09629 0.0030448      0.0092161
b1[1] 4.9942 0.01887 0.0005968      0.0027107
b1[2] 6.9736 0.02492 0.0007879      0.0024949
theta 0.3006 0.01951 0.0006171      0.0008086

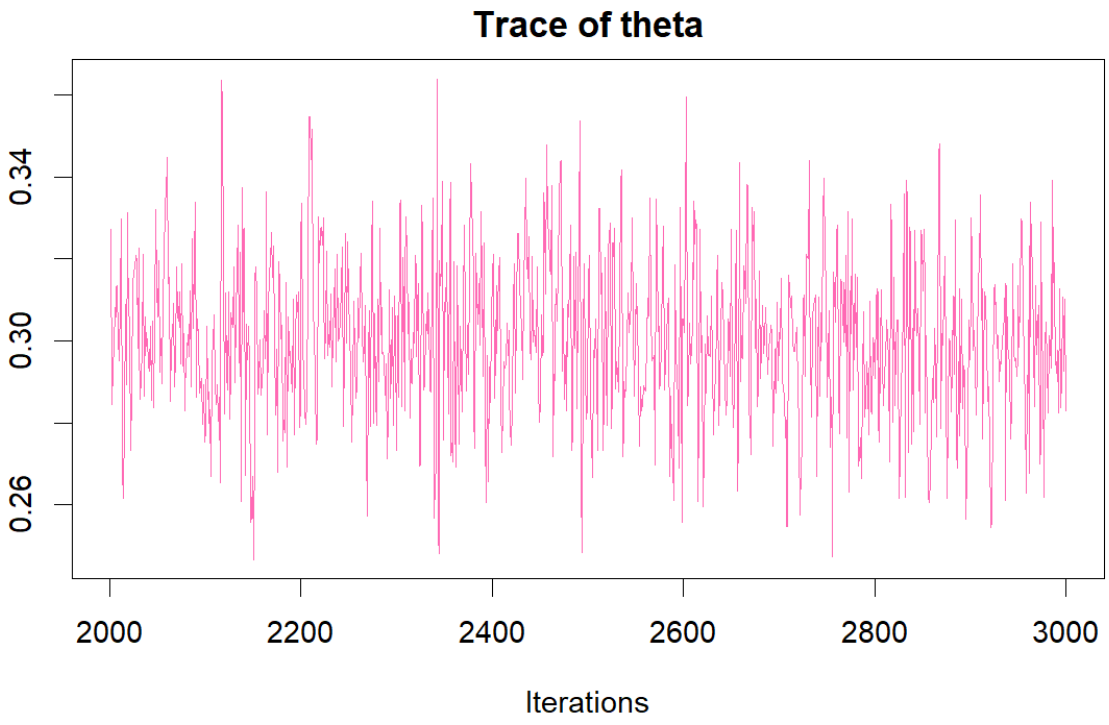
2. Quantiles for each variable:

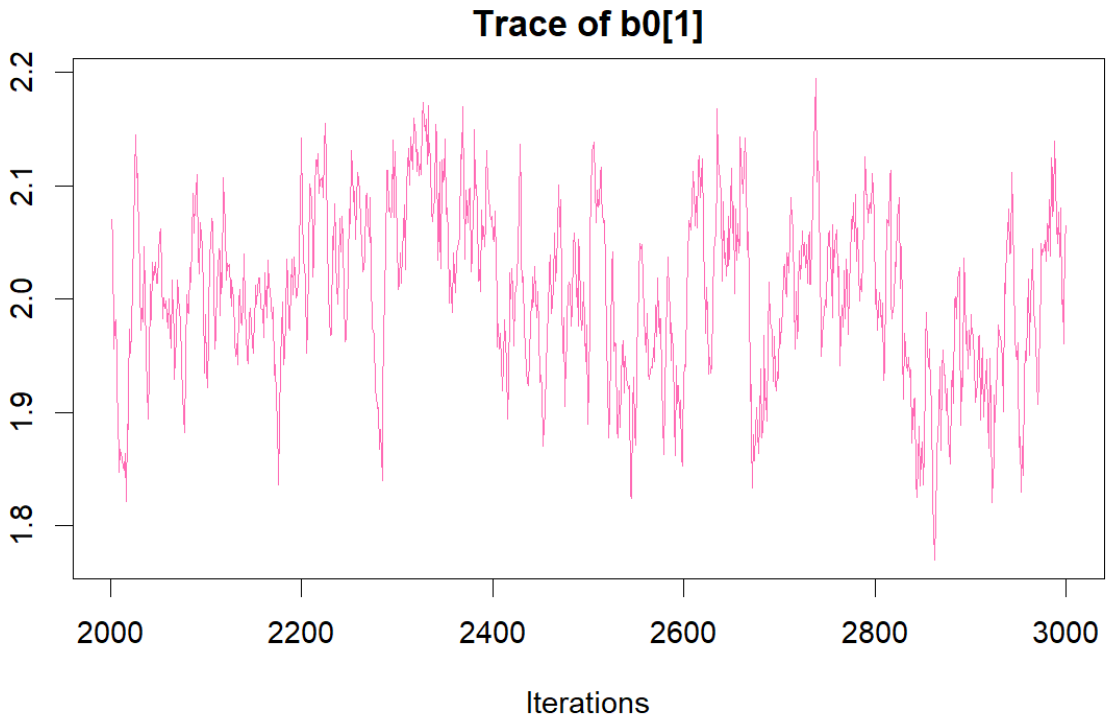
        2.5%    25%    50%    75%  97.5%
b0[1] 1.8556 1.9544 2.0075 2.0595 2.1389
b0[2] 4.8668 4.9848 5.0540 5.1141 5.2388
b1[1] 4.9600 4.9804 4.9933 5.0068 5.0337
b1[2] 6.9249 6.9565 6.9739 6.9907 7.0216
theta 0.2617 0.2875 0.3002 0.3129 0.3388
```

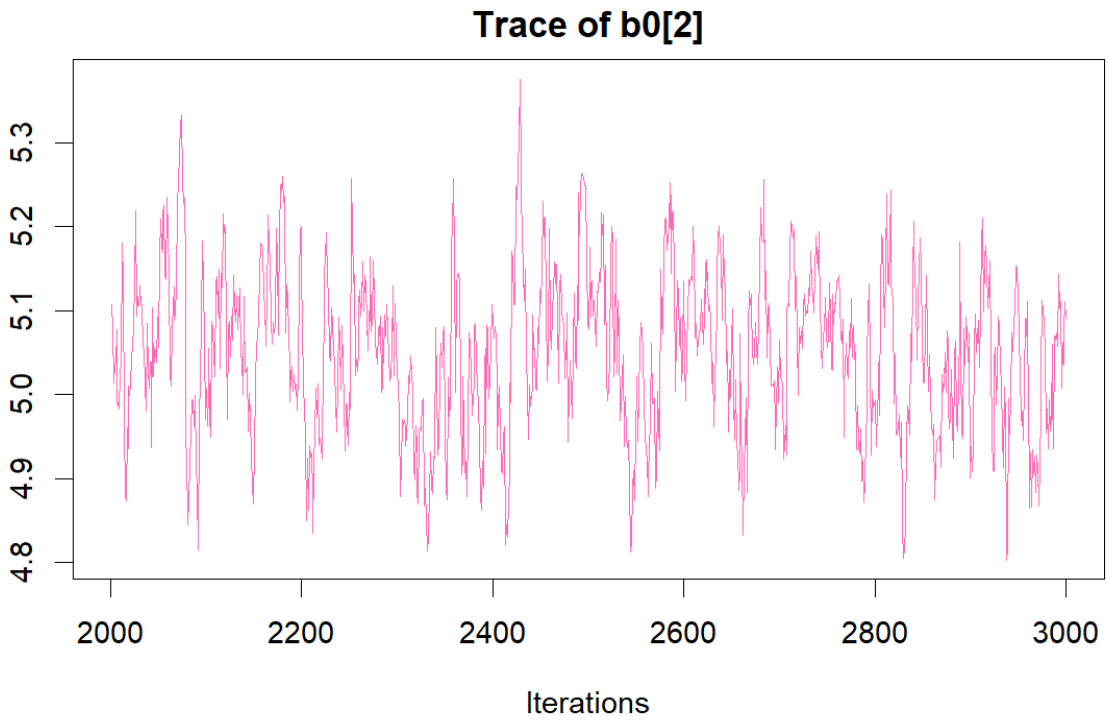
		Classification	
		1	2
True Cluster	1	350	0
	2	0	150

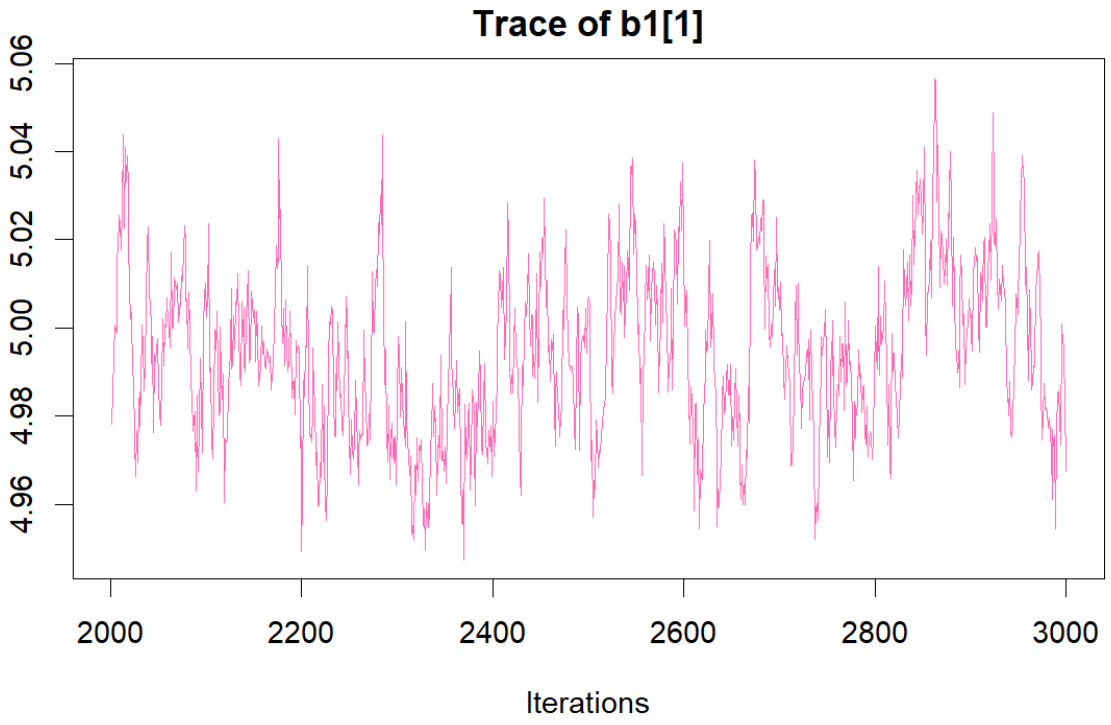
0 subjects were misclassified. My model has a 100% classification accuracy rate.

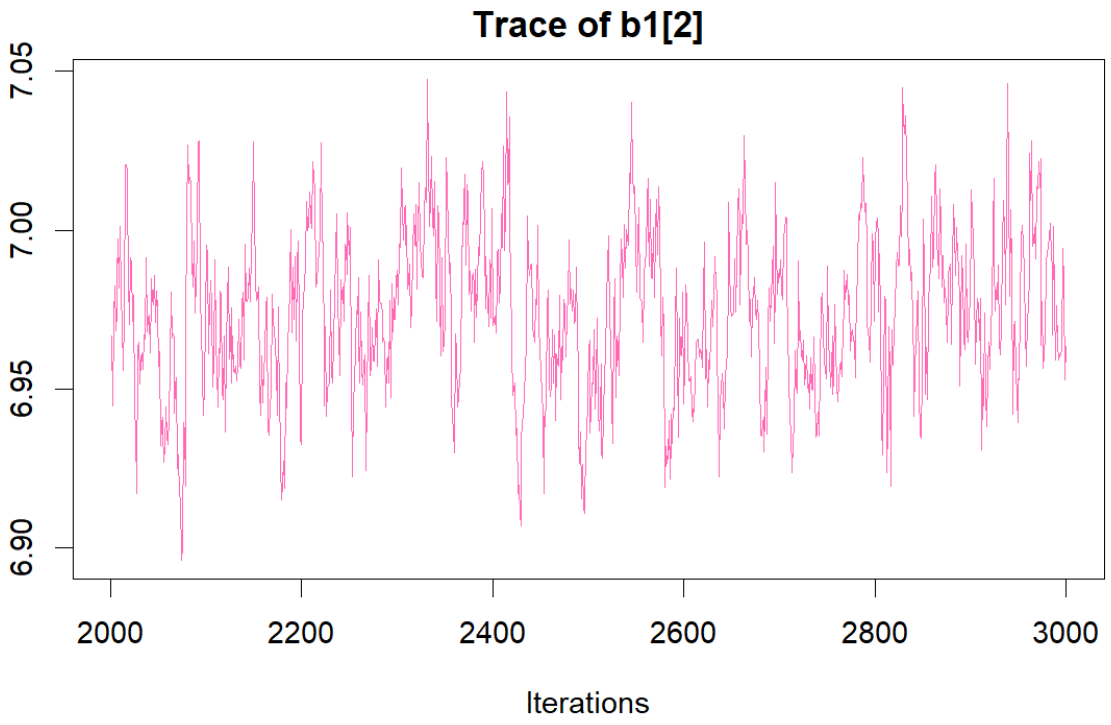

Two-Cluster Labeling for Trajectory Dataset 1
Outcome
0
10
20
30
40
50
1
2
3
4
5
6
Time


Trace of theta
0.26
0.30
0.34
2000
2200
2400
2600
2800
3000
Iterations


Trace of b0[1]
2.2
2.1
2.0
1.9
1.8
2000
2200
2400
2600
2800
3000
Iterations


Trace of b0[2]
5.3
5.2
5.1
5.0
4.9
4.8
2000
2200
2400
2600
2800
3000
Iterations


Trace of b1[1]
5.06
5.04
5.02
5.00
4.98
4.96
2000
2200
2400
2600
2800
3000
Iterations


Trace of b1[2]
7.05
7.00
6.95
6.90
2000
2200
2400
2600
2800
3000
Iterations

Dataset 2

$$Cluster\ 1\ y = 3.0702 + 7.0131$$
$$Cluster\ 2\ y = 3.0702 + 4.1543$$

```
Iterations = 2001:3000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 1000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

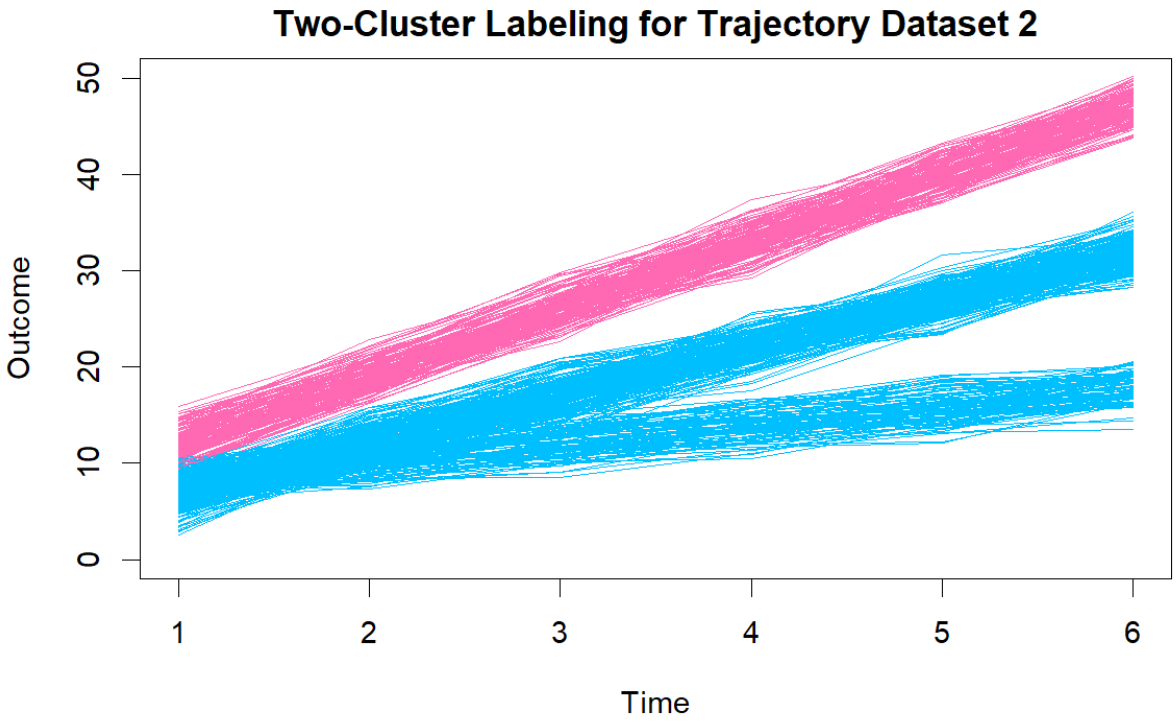
        Mean      SD  Naive SE Time-series SE
b0[1] 3.0784 0.18936 0.0059882      0.0239365
b0[2] 5.0561 0.28067 0.0088756      0.0288488
b1[1] 4.1508 0.04841 0.0015309      0.0056489
b1[2] 7.0158 0.07235 0.0022880      0.0074621
theta 0.3031 0.02115 0.0006689      0.0008065

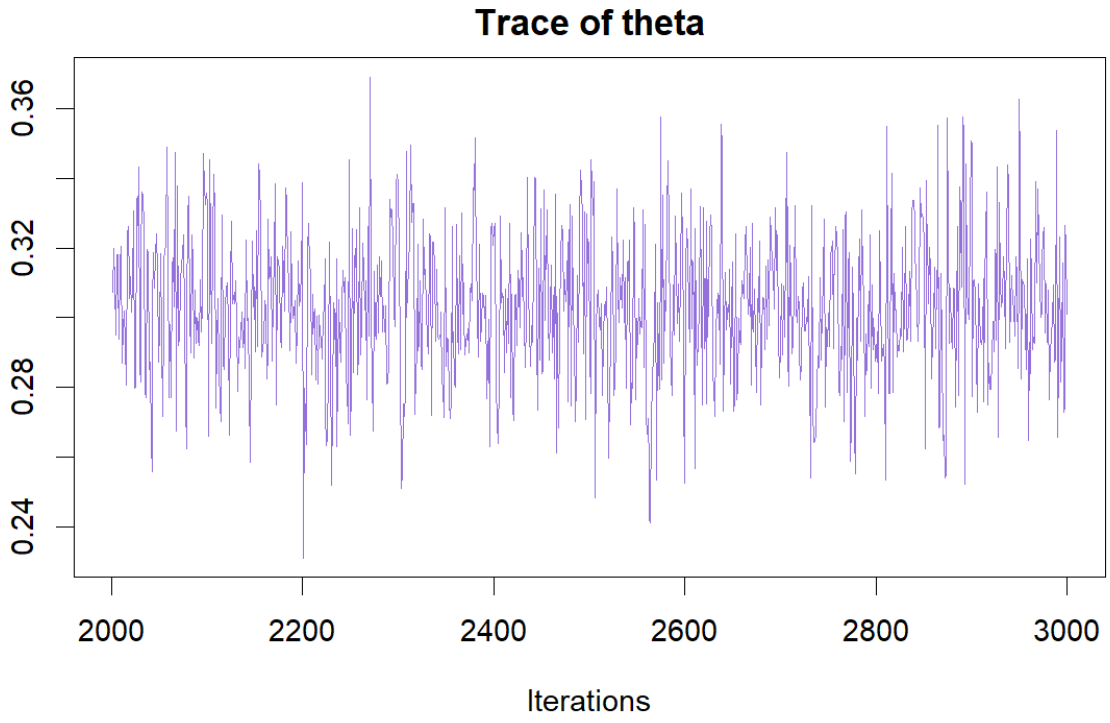
2. Quantiles for each variable:

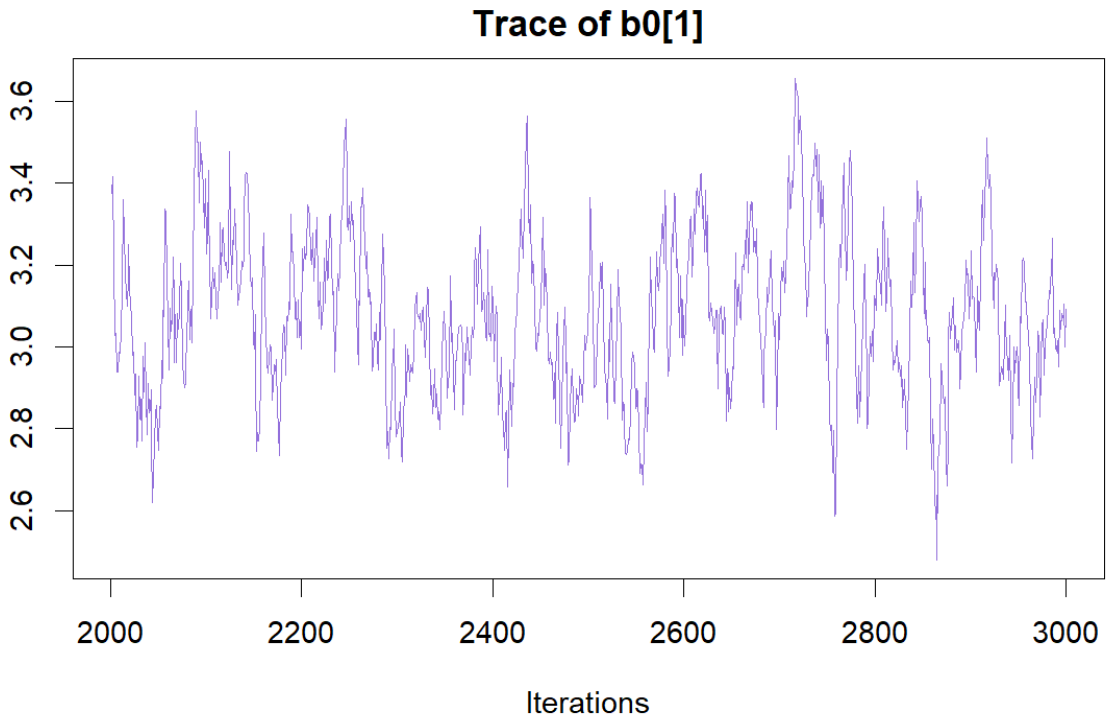
        2.5%    25%    50%    75%  97.5%
b0[1] 2.7330 2.9499 3.0702 3.2049 3.4675
b0[2] 4.5009 4.8760 5.0634 5.2573 5.5702
b1[1] 4.0562 4.1187 4.1543 4.1841 4.2404
b1[2] 6.8807 6.9671 7.0131 7.0639 7.1638
theta 0.2629 0.2896 0.3022 0.3177 0.3446
```

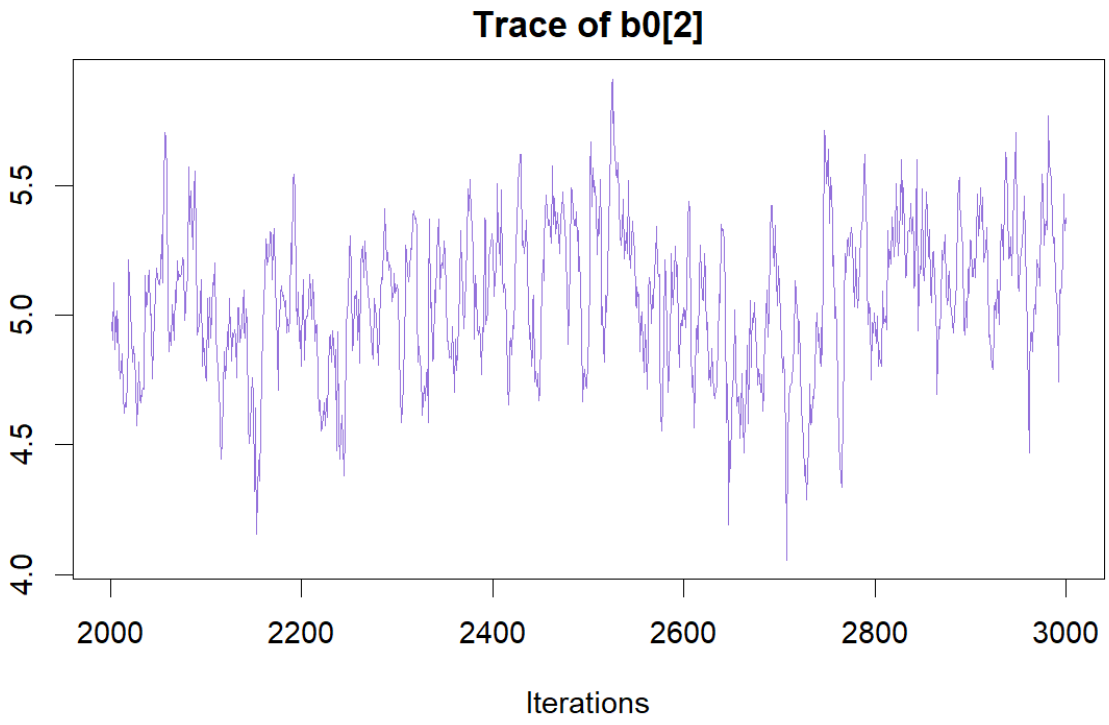
		Classification	
		1	2
True Cluster	1	249	0
	2	0	151
	3	100	0

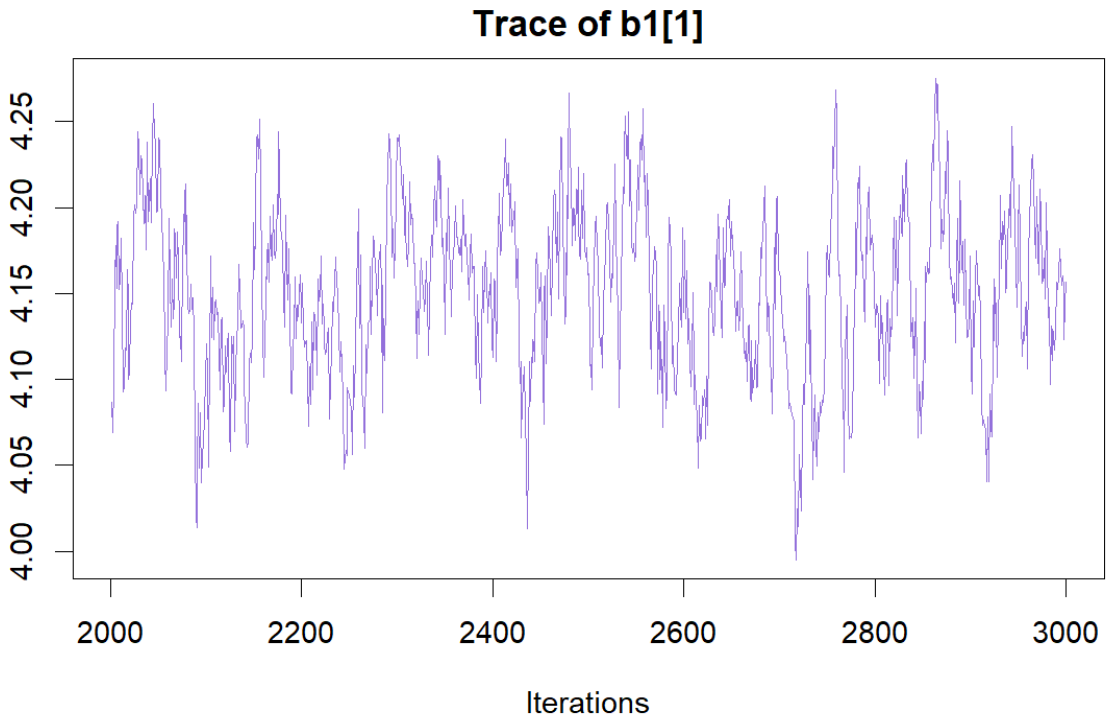
100 subjects were misclassified because the model did not have an option to assign to a group 3. Clusters 1 and 2 were all accurately assigned. My model has a 80% classification accuracy rate.

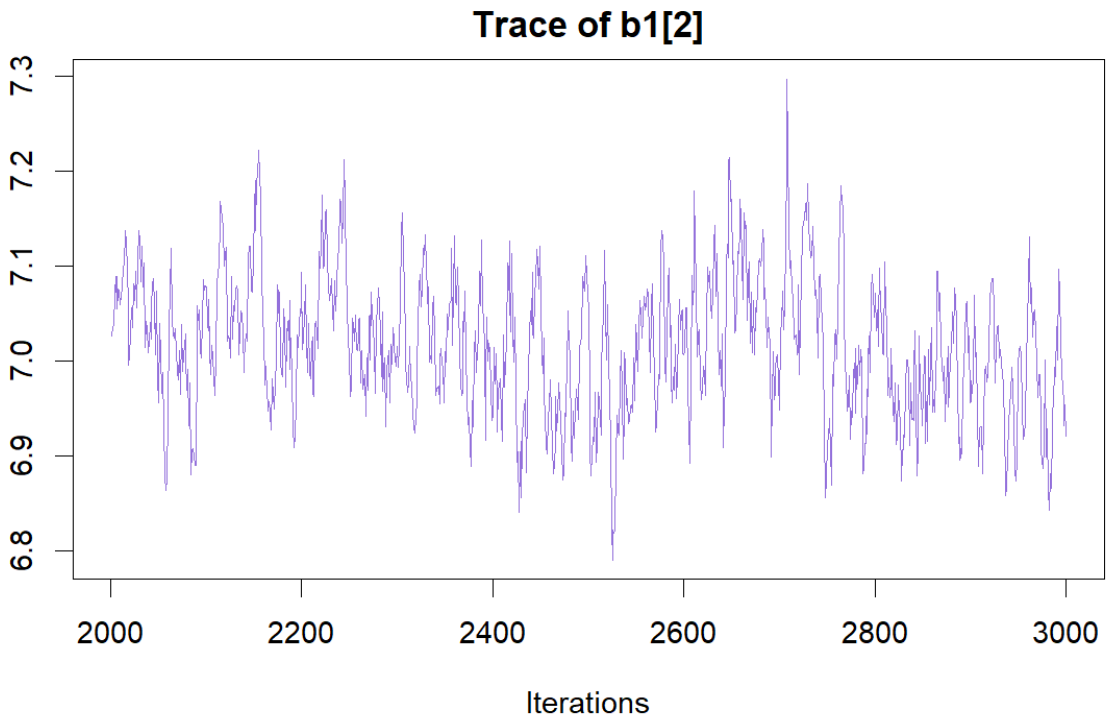

Two-Cluster Labeling for Trajectory Dataset 2
Outcome
0
10
20
30
40
50
1
2
3
4
5
6
Time


Trace of theta
0.24
0.28
0.32
0.36
2000
2200
2400
2600
2800
3000
Iterations


Trace of b0[1]
2.6
2.8
3.0
3.2
3.4
3.6
2000
2200
2400
2600
2800
3000
Iterations


Trace of b0[2]
4.0
4.5
5.0
5.5
2000
2200
2400
2600
2800
3000
Iterations


Trace of b1[1]
4.00
4.05
4.10
4.15
4.20
4.25
2000
2200
2400
2600
2800
3000
Iterations


Trace of b1[2]
6.8
6.9
7.0
7.1
7.2
7.3
2000
2200
2400
2600
2800
3000
Iterations

Analyze the same data sets using 3 clusters. Compare the clusters assignment generated with MCMC and the true cluster label and discuss the accuracy of cluster membership estimated with the MCMC. Include a plot of the trajectories and color them based on the cluster label estimated with the MCMC.

Dataset 1

$$Cluster\ 1\ y = 2.8890 + 4.9918$$
$$Cluster\ 2\ y = 1.3173 + 4.9915$$
$$Cluster\ 3\ y = 5.0254 + 6.9807$$

```
1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

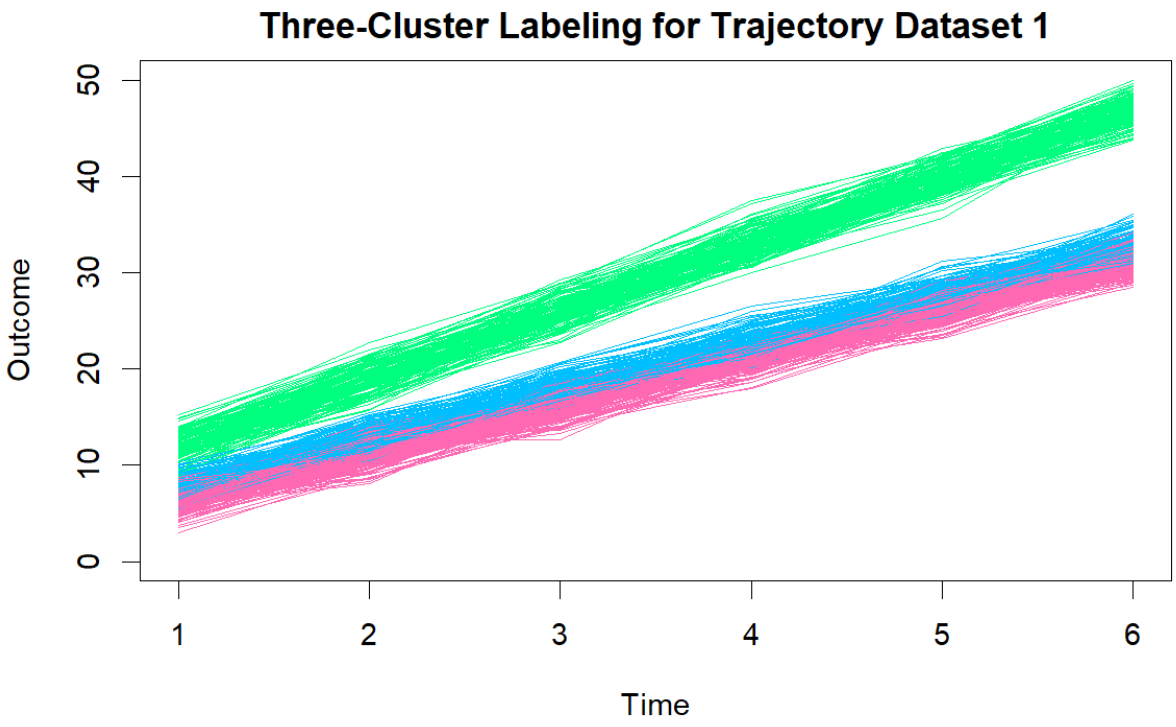
           Mean      SD  Naive SE Time-series SE
b0[1]    2.8797 0.13549 0.0042845      0.0198858
b0[2]    1.3149 0.09821 0.0031056      0.0118874
b0[3]    5.0235 0.08782 0.0027772      0.0085452
b1[1]    4.9906 0.02614 0.0008266      0.0028683
b1[2]    4.9905 0.02315 0.0007322      0.0025040
b1[3]    6.9803 0.02276 0.0007198      0.0021566
theta[1] 0.3158 0.03767 0.0011913      0.0050825
theta[2] 0.3844 0.03843 0.0012154      0.0047857
theta[3] 0.2998 0.02058 0.0006509      0.0006509

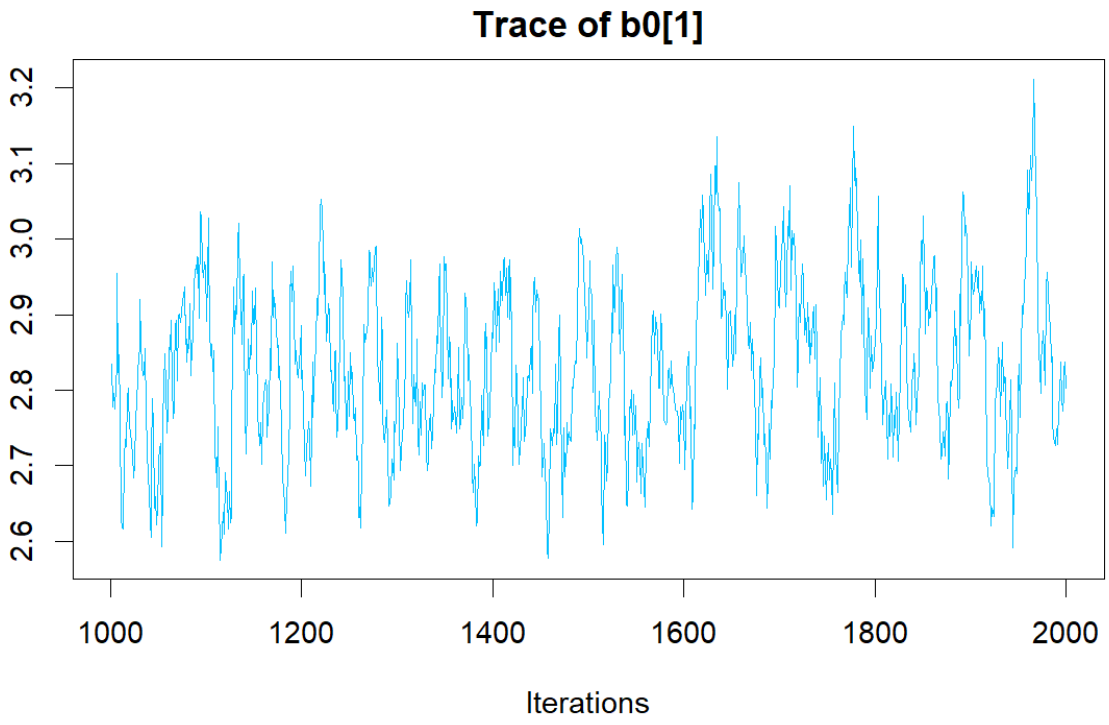
2. Quantiles for each variable:

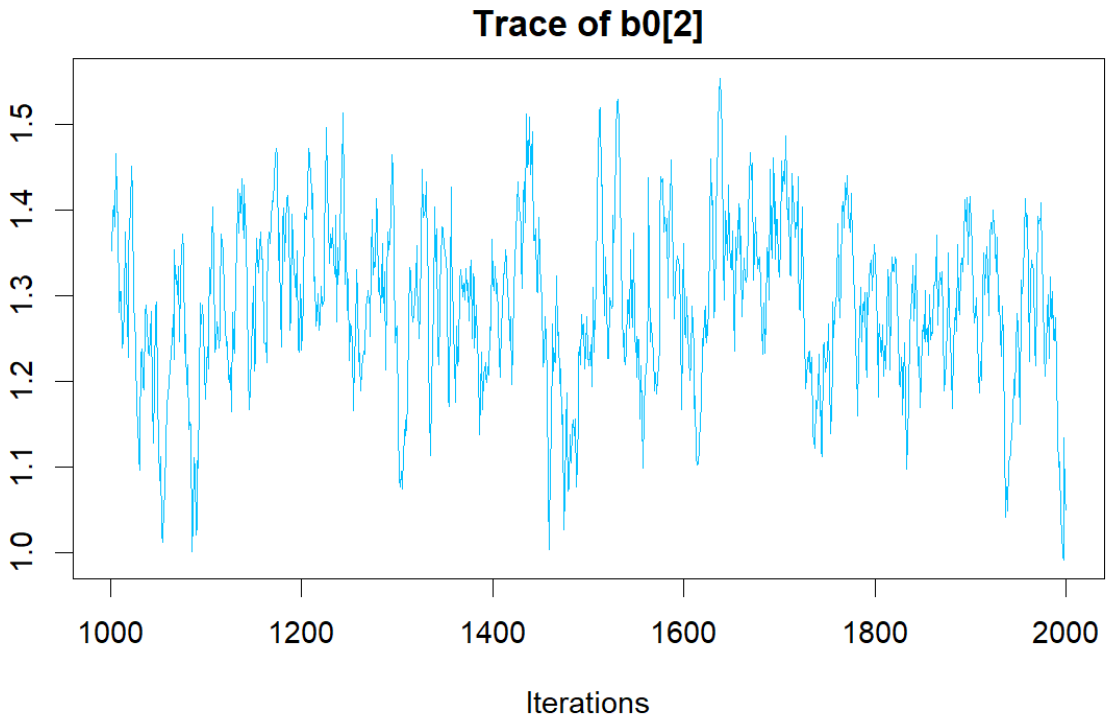
           2.5%    25%    50%    75%  97.5%
b0[1]    2.5917 2.7924 2.8890 2.9685 3.1249
b0[2]    1.1263 1.2455 1.3173 1.3822 1.4959
b0[3]    4.8530 4.9661 5.0254 5.0820 5.2018
b1[1]    4.9416 4.9723 4.9918 5.0087 5.0395
b1[2]    4.9436 4.9743 4.9907 5.0065 5.0348
b1[3]    6.9336 6.9661 6.9807 6.9941 7.0264
theta[1] 0.2477 0.2908 0.3135 0.3403 0.3936
theta[2] 0.3057 0.3598 0.3870 0.4112 0.4514
theta[3] 0.2602 0.2858 0.2992 0.3135 0.3402
```

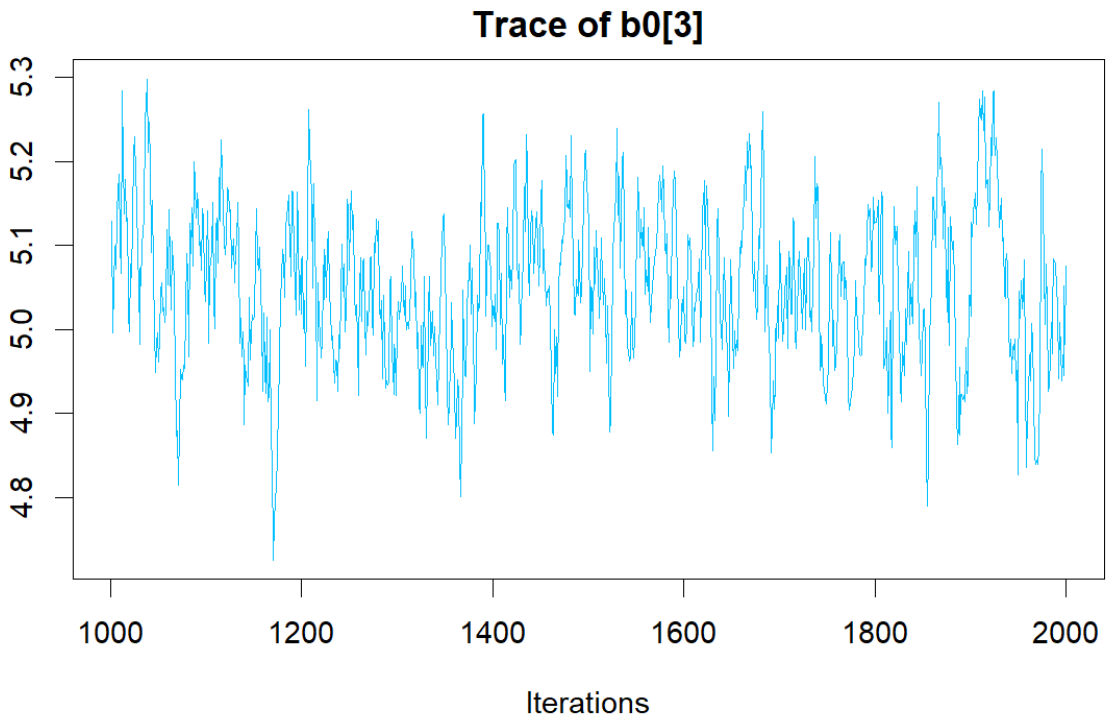
		Classification		
		1	2	3
True Cluster	1	150	200	0
	2	0	0	150

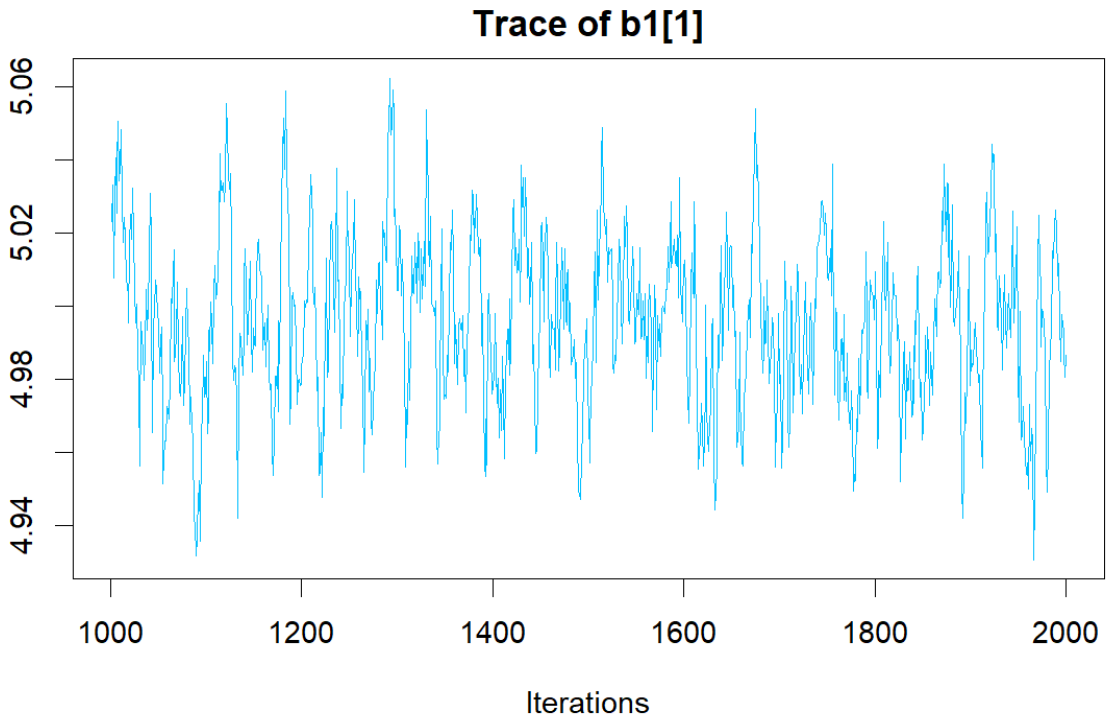
150 subjects were misclassified because the actual underlying distribution did not have a third cluster. The model classified the actual second cluster as cluster 3 and split the actual first cluster into clusters 1 and 2. My model has a 30% classification accuracy rate.

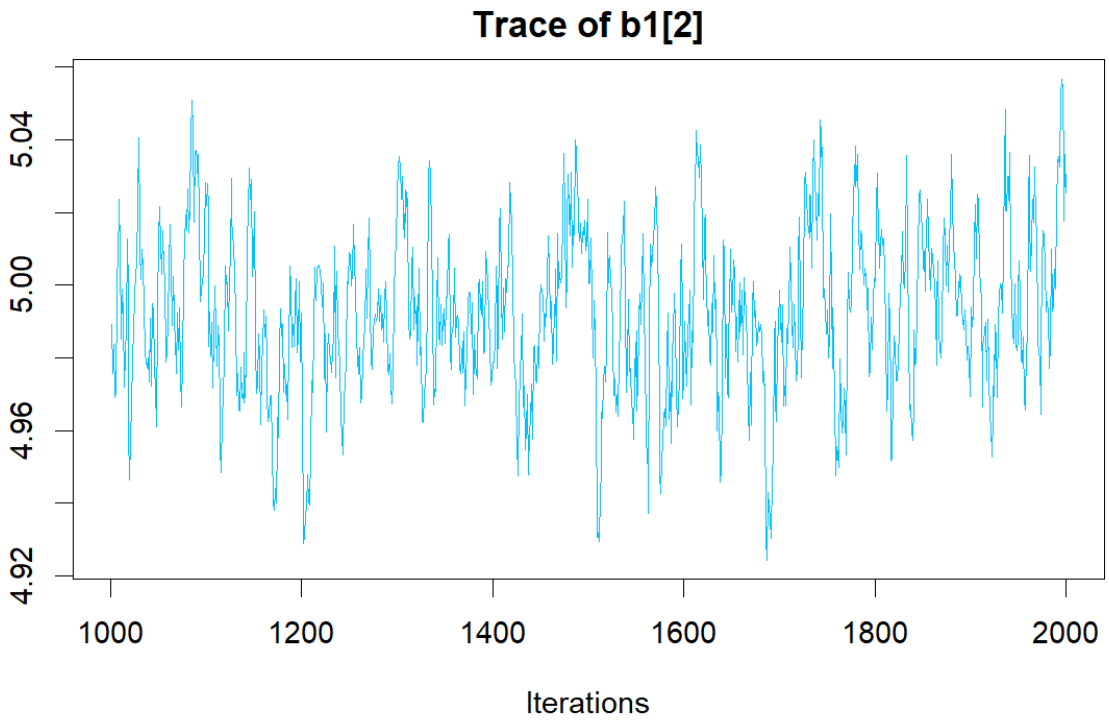

Three-Cluster Labeling for Trajectory Dataset 1
Outcome
0
10
20
30
40
50
1
2
3
4
5
6
Time

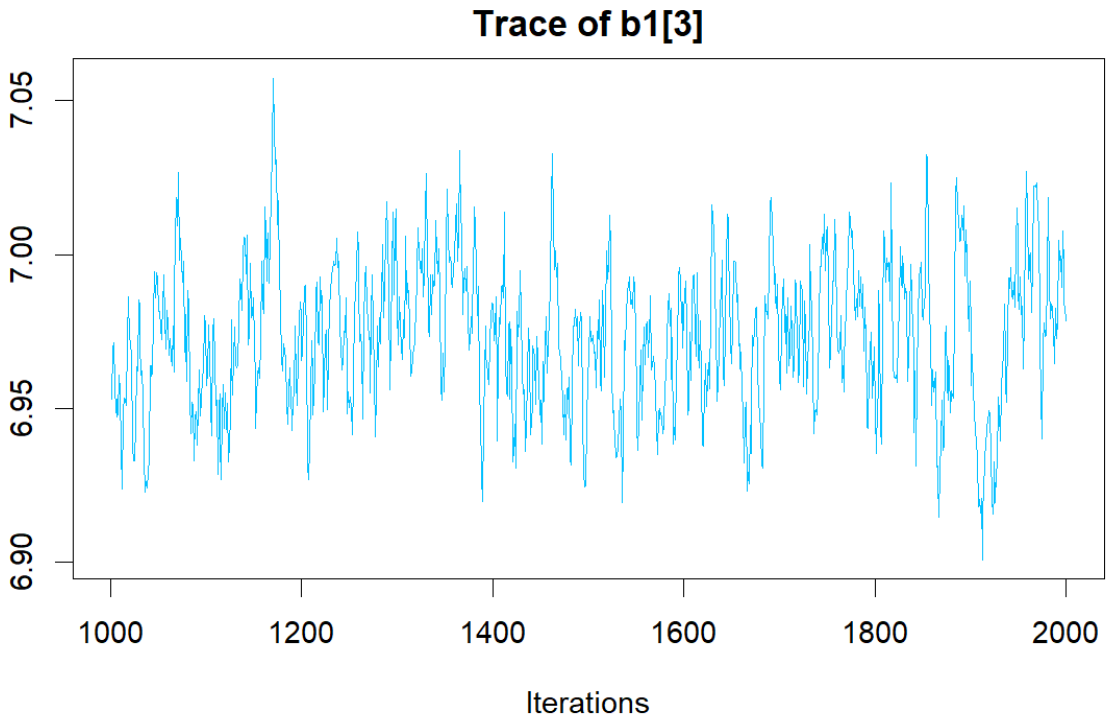

Trace of b0[1]
2.6
2.7
2.8
2.9
3.0
3.1
3.2
1000
1200
1400
1600
1800
2000
Iterations

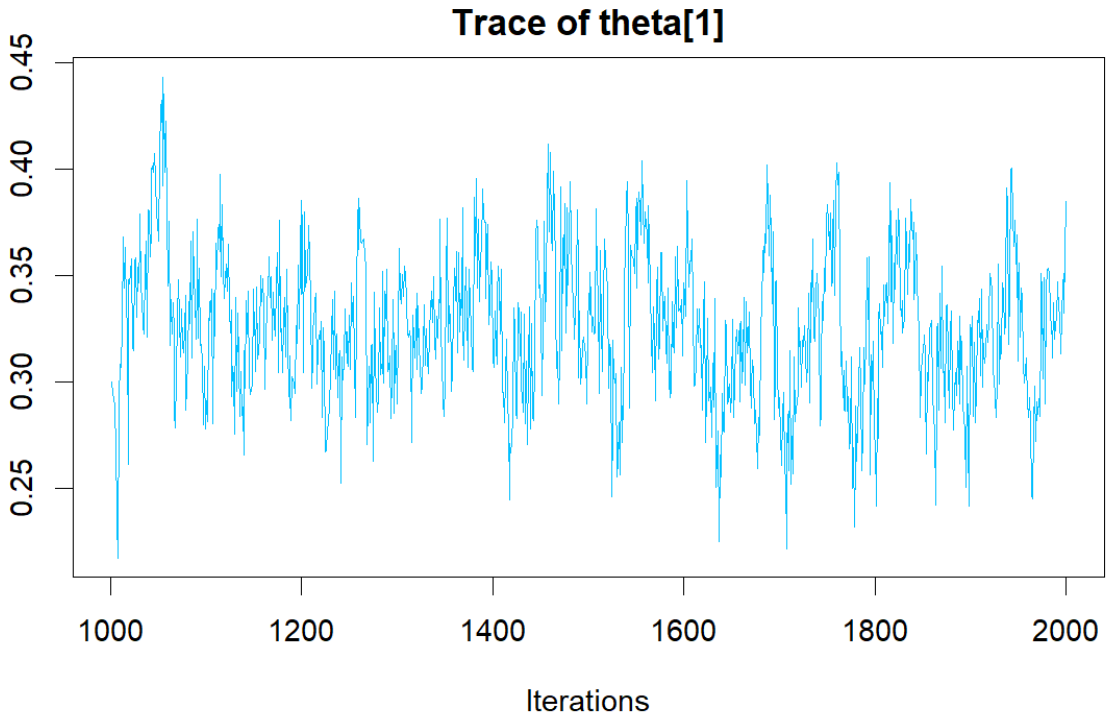

Trace of b0[2]
1.0
1.1
1.2
1.3
1.4
1.5
1000
1200
1400
1600
1800
2000
Iterations

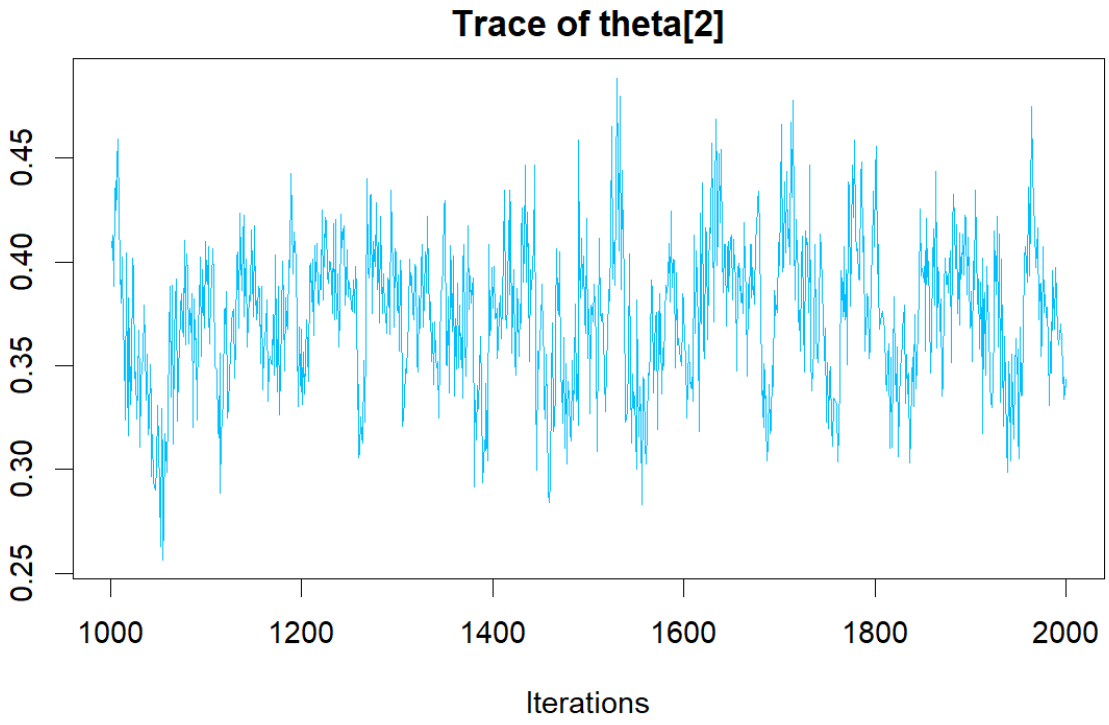

Trace of b0[3]
4.8
4.9
5.0
5.1
5.2
5.3
1000
1200
1400
1600
1800
2000
Iterations

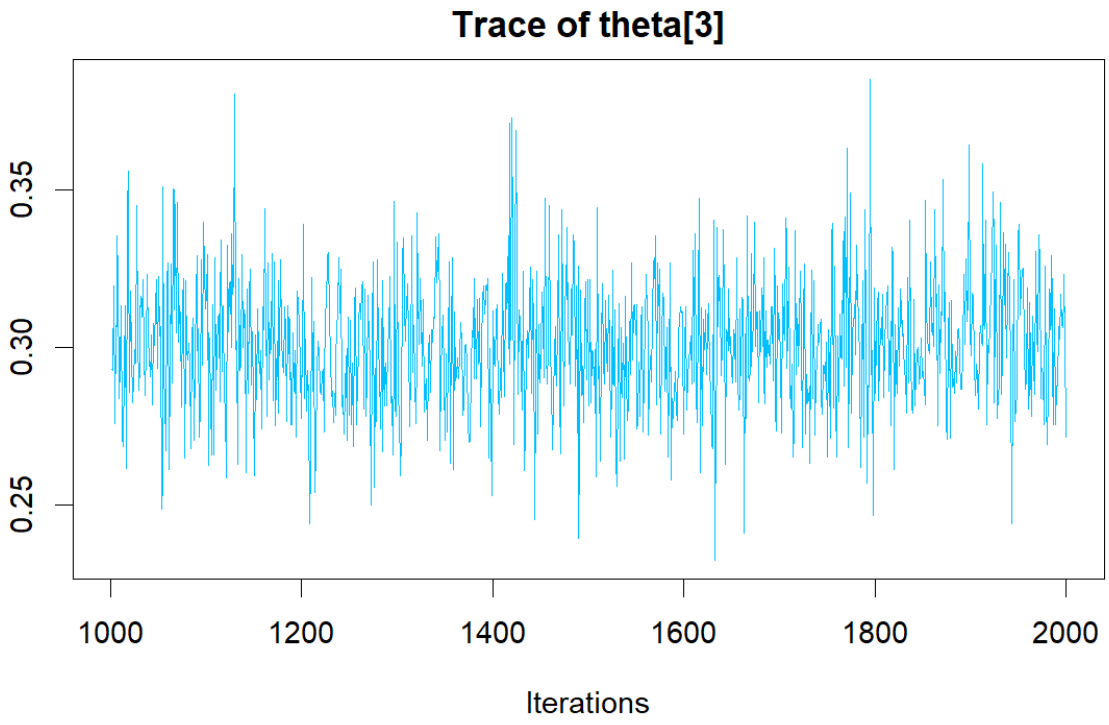

Trace of b1[1]
5.06
5.02
4.98
4.94
1000
1200
1400
1600
1800
2000
Iterations


Trace of b1[2]
5.04
5.00
4.96
4.92
1000
1200
1400
1600
1800
2000
Iterations


Trace of b1[3]
7.05
7.00
6.95
6.90
1000
1200
1400
1600
1800
2000
Iterations


Trace of theta[1]
0.45
0.40
0.35
0.30
0.25
1000
1200
1400
1600
1800
2000
Iterations


Trace of theta[2]
0.25
0.30
0.35
0.40
0.45
1000
1200
1400
1600
1800
2000
Iterations


Trace of theta[3]
0.25
0.30
0.35
1000
1200
1400
1600
1800
2000
Iterations

Dataset 2

$$Cluster\ 1\ y = 1.9234 + 5.0146$$
$$Cluster\ 2\ y = 5.0635 + 5.0635$$
$$Cluster\ 3\ y = 5.9592 + 2.0015$$

```
Iterations = 1001:2000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 1000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

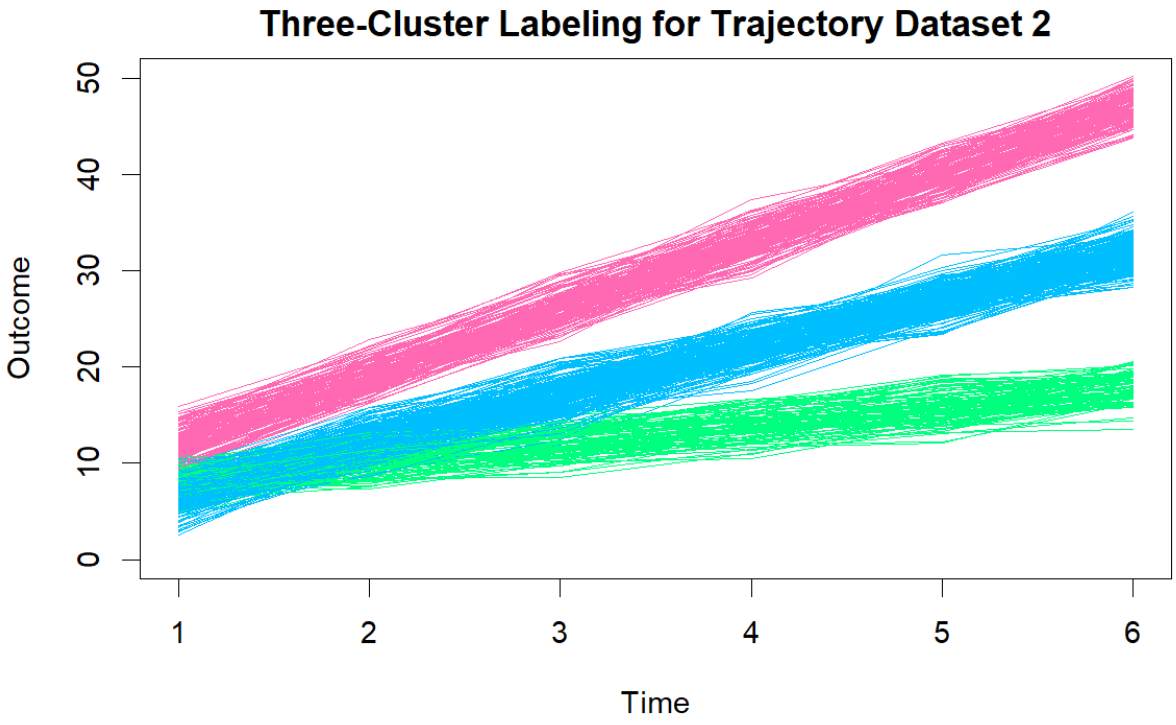
           Mean      SD  Naive SE Time-series SE
b0[1]    1.9226 0.08438 0.0026683      0.0080466
b0[2]    5.0627 0.11265 0.0035625      0.0106557
b0[3]    5.9549 0.12407 0.0039233      0.0112166
b1[1]    5.0145 0.02148 0.0006793      0.0020317
b1[2]    7.0143 0.02866 0.0009064      0.0027388
b1[3]    2.0011 0.03180 0.0010055      0.0028277
theta[1] 0.4969 0.02208 0.0006983      0.0006983
theta[2] 0.3023 0.02059 0.0006510      0.0006510
theta[3] 0.2008 0.01807 0.0005715      0.0005970

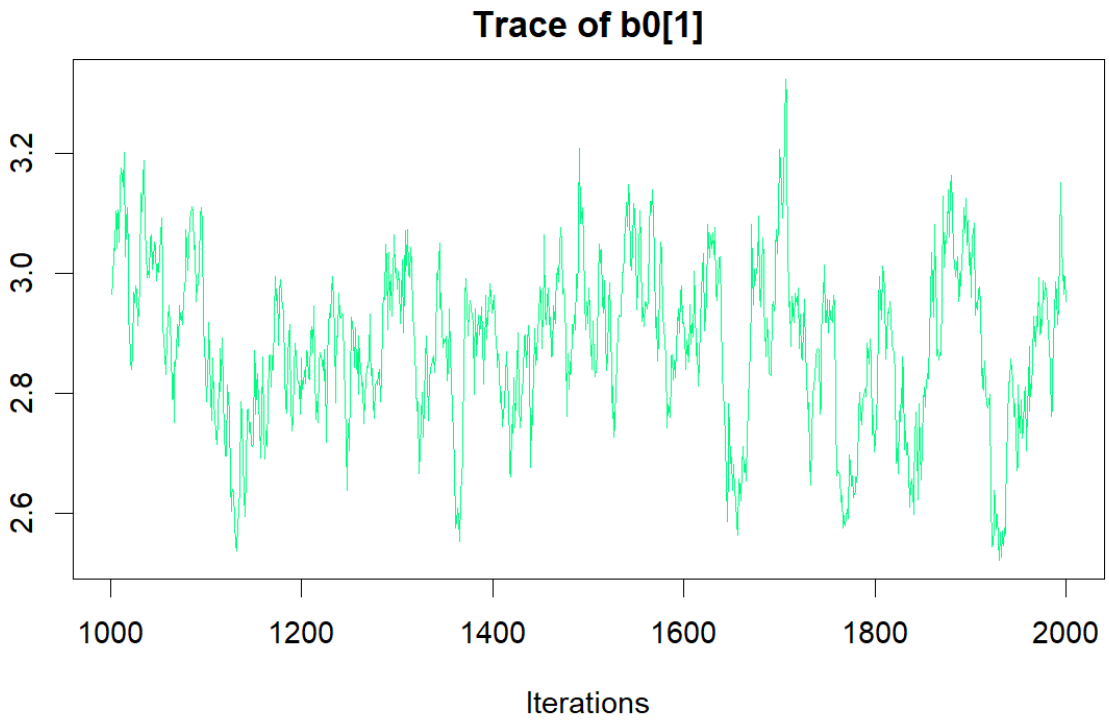
2. Quantiles for each variable:

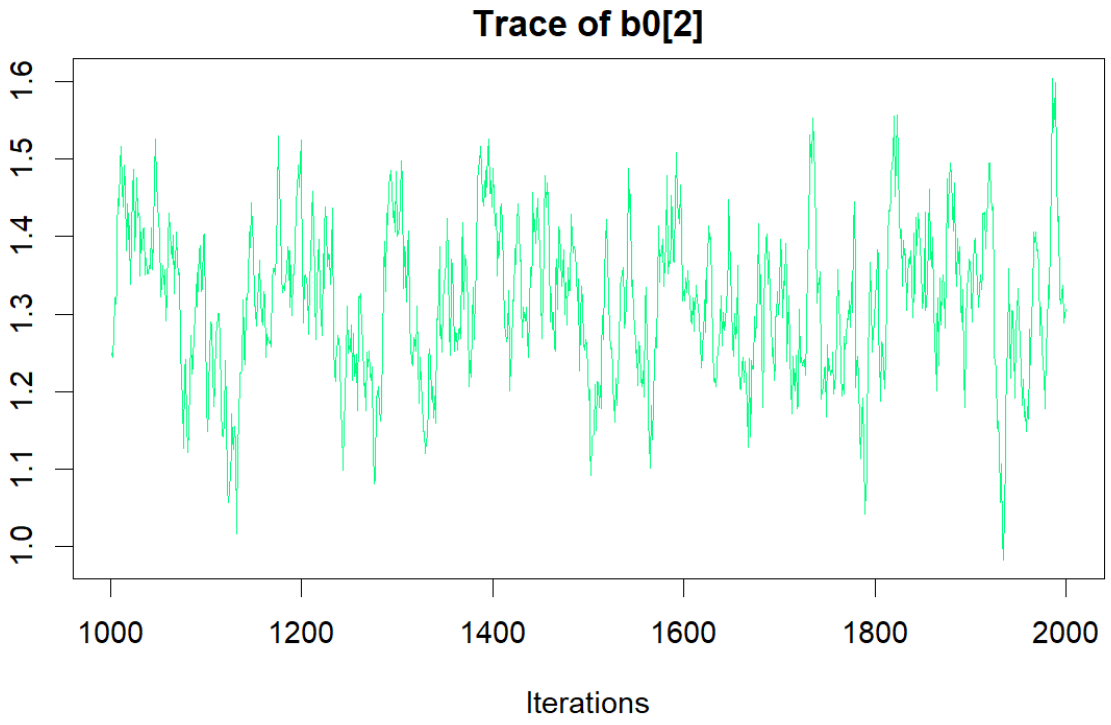
           2.5%    25%    50%    75%  97.5%
b0[1]    1.7579 1.8688 1.9234 1.9758 2.0886
b0[2]    4.8411 4.9865 5.0635 5.1375 5.2917
b0[3]    5.7087 5.8735 5.9592 6.0344 6.2006
b1[1]    4.9719 5.0009 5.0146 5.0276 5.0575
b1[2]    6.9566 6.9955 7.0149 7.0334 7.0703
b1[3]    1.9397 1.9799 2.0015 2.0223 2.0653
theta[1] 0.4524 0.4819 0.4977 0.5117 0.5399
theta[2] 0.2607 0.2880 0.3029 0.3160 0.3427
theta[3] 0.1651 0.1887 0.2008 0.2115 0.2362
```

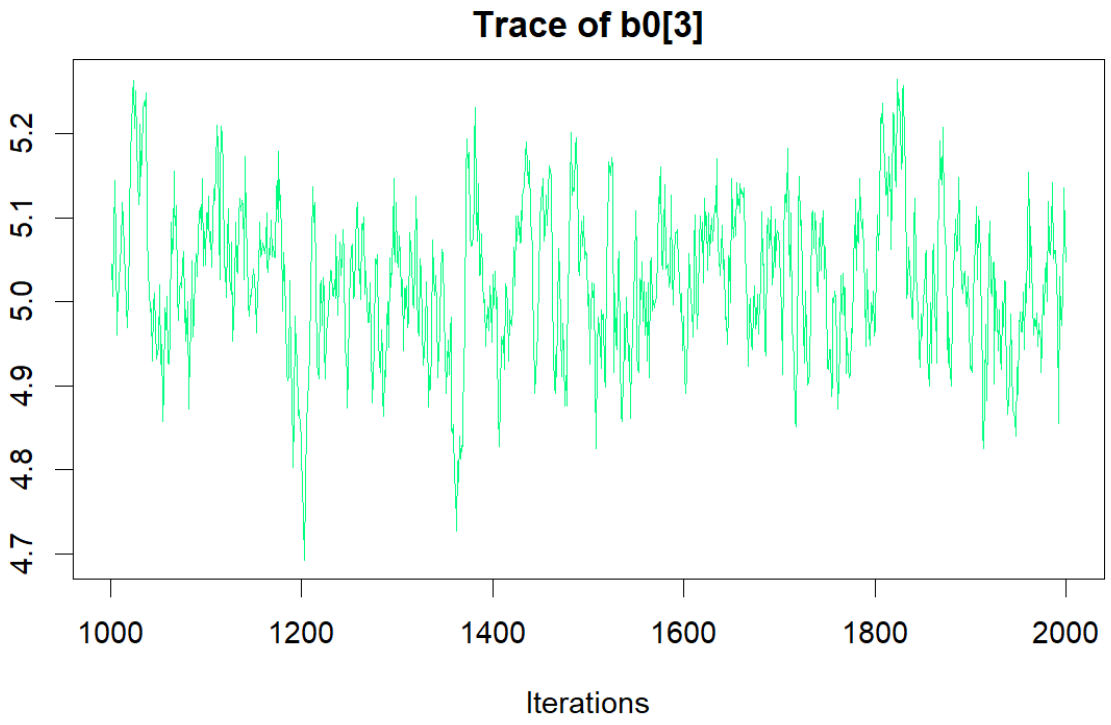
		Classification		
		1	2	3
True Cluster	1	249	0	0
	2	0	151	0
	3	0	0	100

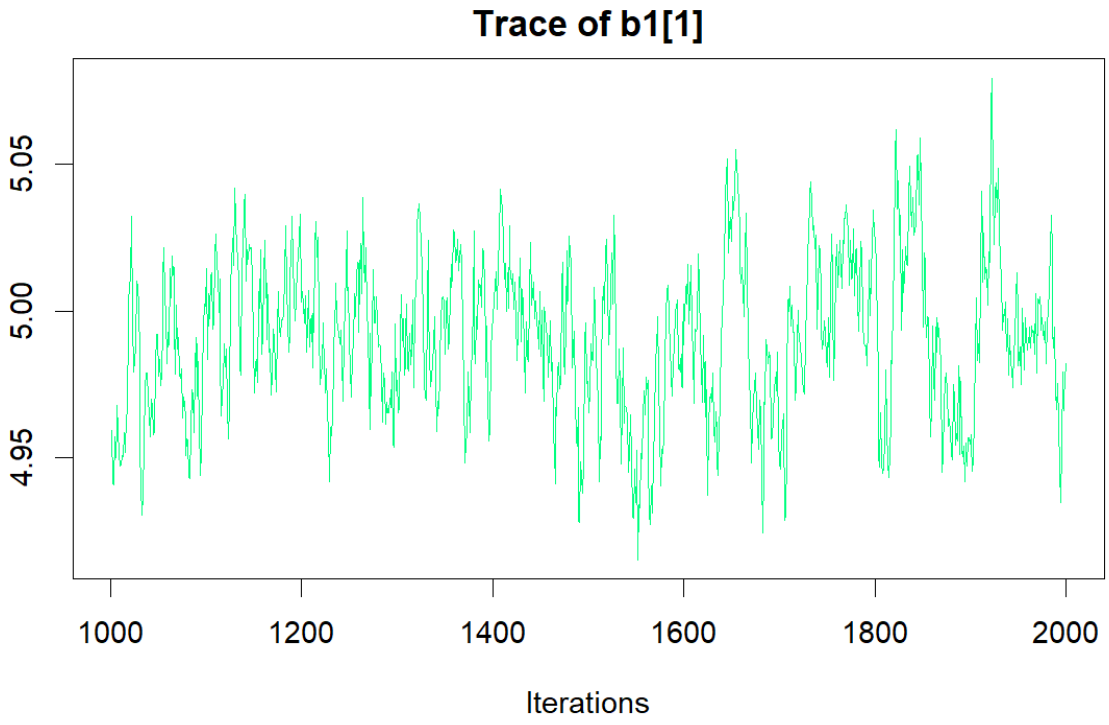
0 subjects were misclassified. My model has a 100% classification accuracy rate.

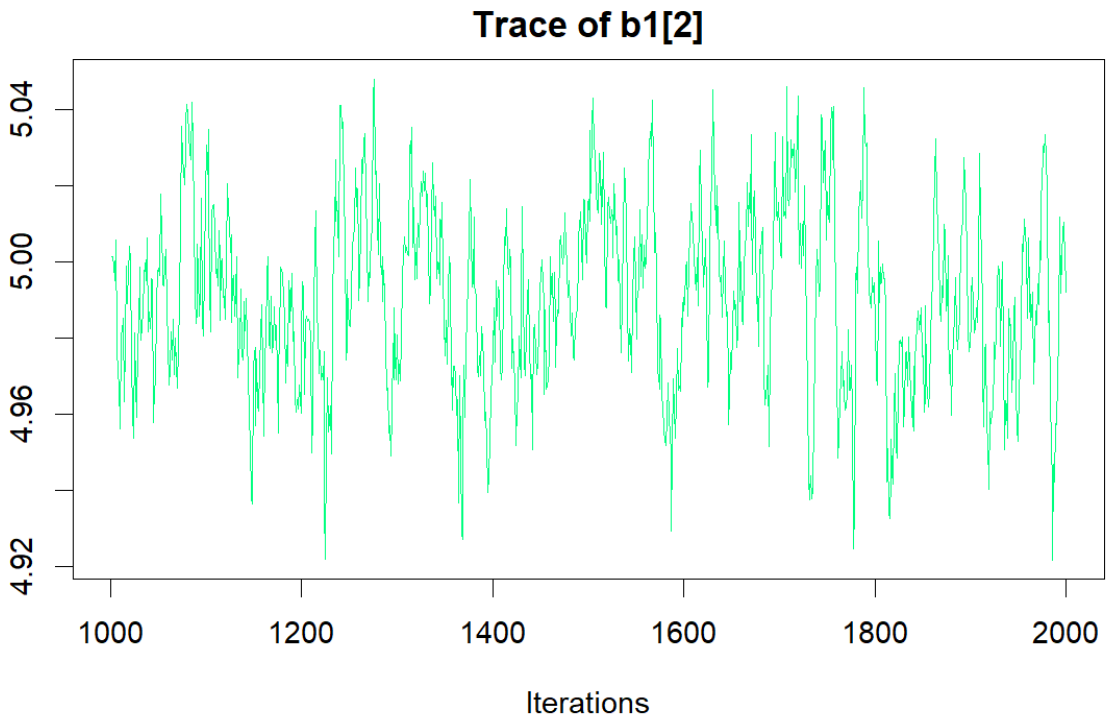

Three-Cluster Labeling for Trajectory Dataset 2
Outcome
0
10
20
30
40
50
1
2
3
4
5
6
Time

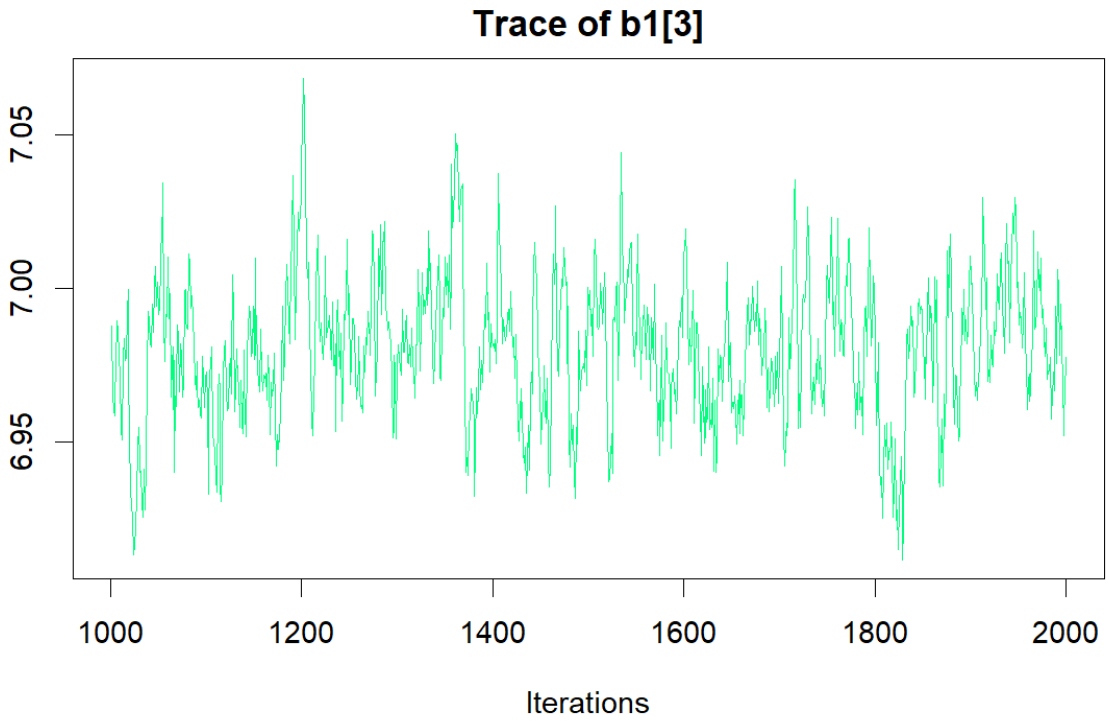

Trace of b0[1]
2.6
2.8
3.0
3.2
1000
1200
1400
1600
1800
2000
Iterations

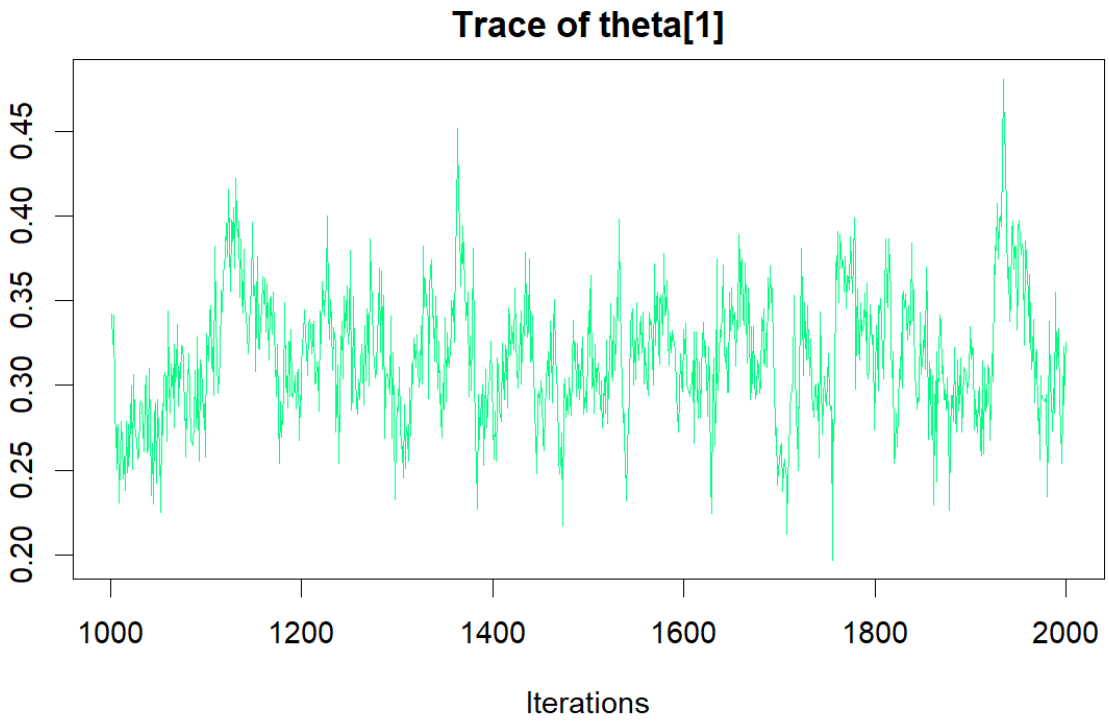

Trace of b0[2]
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1000
1200
1400
1600
1800
2000
Iterations

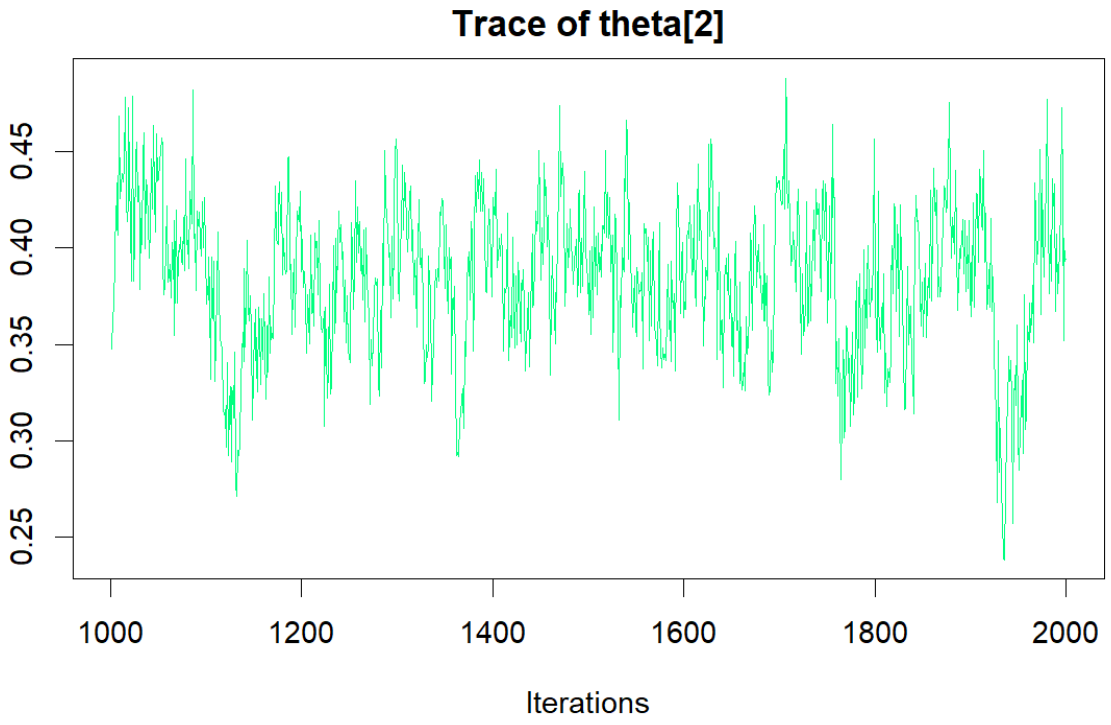

Trace of b0[3]
4.7
4.8
4.9
5.0
5.1
5.2
1000
1200
1400
1600
1800
2000
Iterations

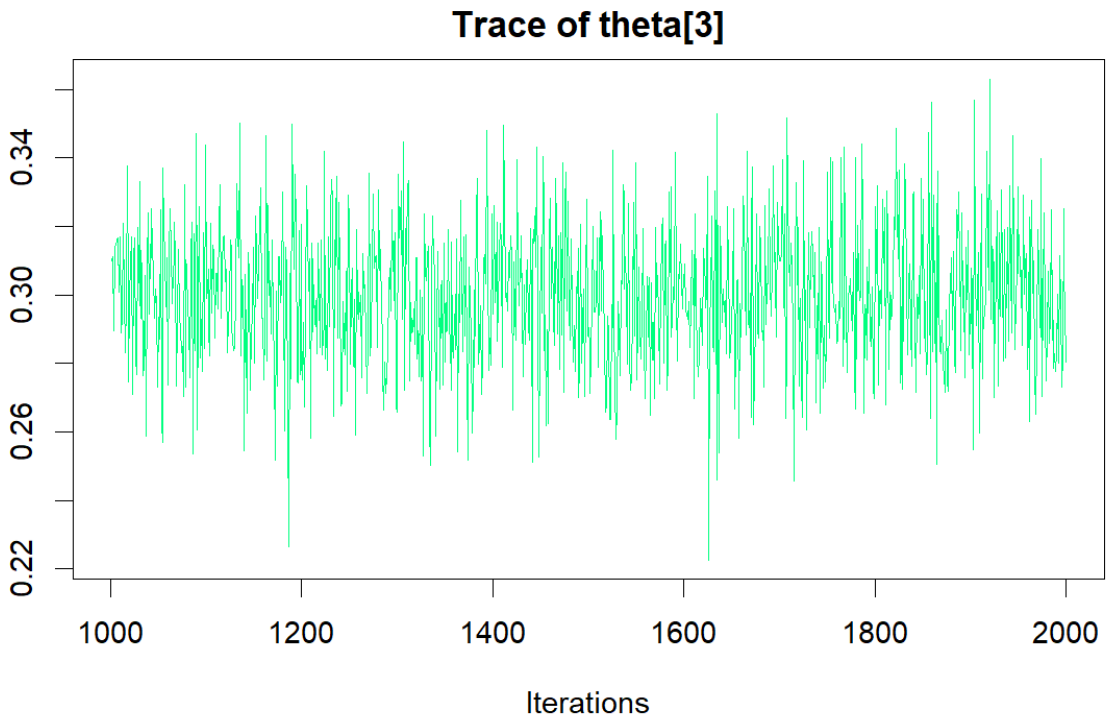

Trace of b1[1]
5.05
5.00
4.95
1000
1200
1400
1600
1800
2000
Iterations


Trace of b1[2]
5.04
5.00
4.96
4.92
1000
1200
1400
1600
1800
2000
Iterations


Trace of b1[3]
7.05
7.00
6.95
1000
1200
1400
1600
1800
2000
Iterations

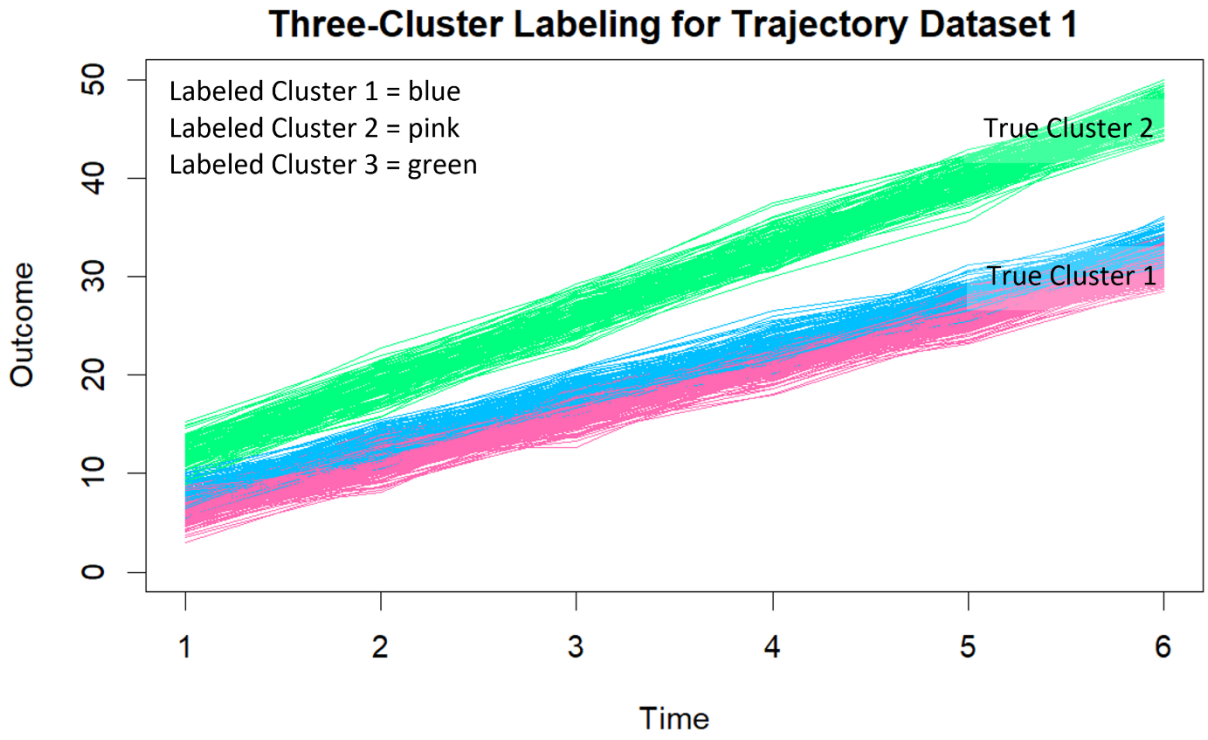

Trace of theta[1]
0.45
0.40
0.35
0.30
0.25
0.20
1000
1200
1400
1600
1800
2000
Iterations


Trace of theta[2]
0.25
0.30
0.35
0.40
0.45
1000
1200
1400
1600
1800
2000
Iterations


Trace of theta[3]
0.22
0.26
0.30
0.34
1000
1200
1400
1600
1800
2000
Iterations

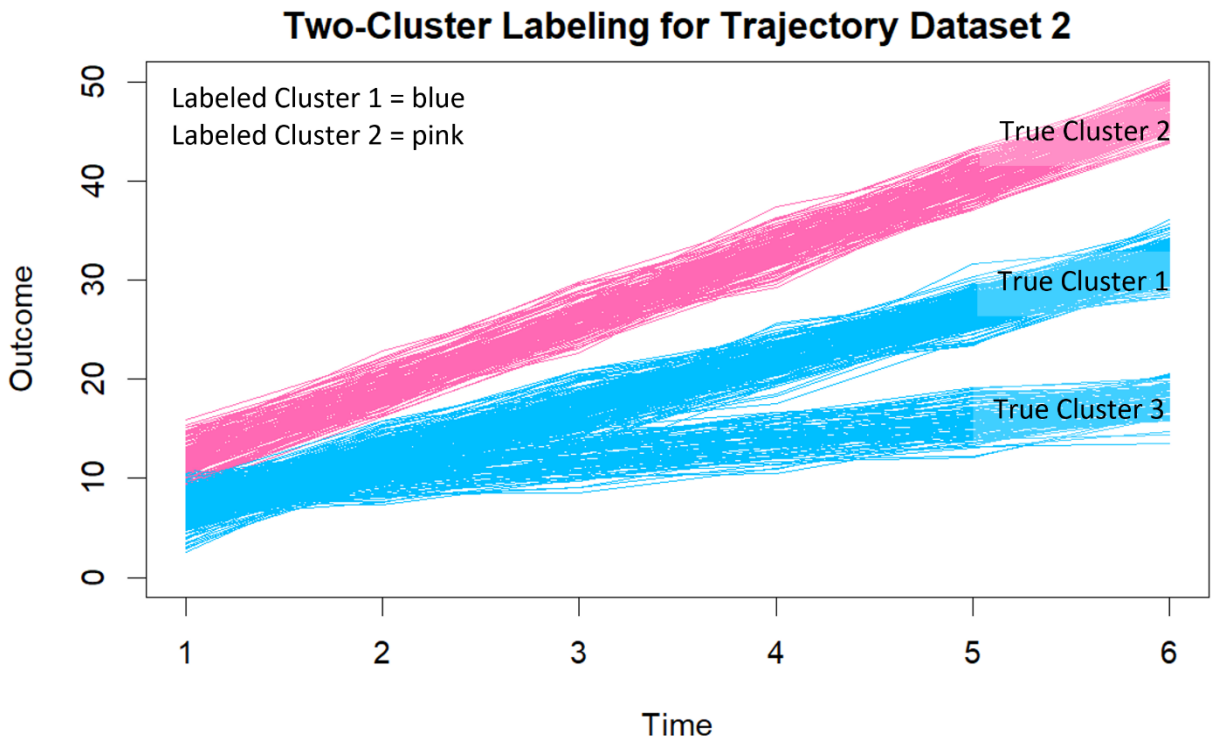
Discuss the results of all analyses.

The first dataset was made of two distributions, $y = 2 + 5x, y = 5 + 7x$, and the second dataset was made of the same two distributions plus one more, $y = 6 + 2x$. A 100% accuracy classification rate was achieved when a two-cluster model was applied to the first dataset and a three-cluster model was applied to the second dataset.



		Classification		
		1	2	3
True Cluster	1	150	200	0
	2	0	0	150

When a three-cluster model was forced onto the first dataset, the model separated the first distribution into two clusters, as seen in the graph above. 150 subjects out of the 350 subjects in the first distribution were classified correctly, the other 200 were classified as cluster 2. This model had a 30% accuracy rate.



		Classification	
		1	2
True Cluster	1	249	0
	2	0	151
	3	100	0

When a two-cluster model was forced onto the second dataset, the model combined the first and third distributions together into one cluster because their trajectories were closer to each other than the second trajectory, as seen in the graph above. All 249 subjects in the first distribution and all 151 subjects in the second distribution were classified correctly. The 100 subjects from the third distribution were misclassified as cluster 1. This model had an 80% accuracy rate.