Task A: Analysis of the “Old Faithful” geyser data set with GMMs

ID: 10265128
Shrayansh Jyoti

Background[1]:

Old Faithful is a cone geyser located in Yellowstone National Park in Wyoming, United States. It was named in 1870 during the Washburn-Langford-Doane Expedition and was the first geyser in the park to receive a name. It is a highly predictable geothermal feature, and has erupted every 44 minutes to two hours since 2000. More than 1,000,000 eruptions have been recorded. Harry Woodward first described a mathematical relationship between the duration and intervals of the eruptions in 1938. More than 1,000,000 eruptions have been recorded. Harry Woodward first described a mathematical relationship between the duration and intervals of the eruptions in 1938.

Description of the data:

This paper describes the analysis of some data on the Old Faithful geyser. The data consists of 272 pairs of measurements, with two variables:

- eruption i.e., the duration of each eruption
- waiting i.e., the time interval between successive intervals.

Use code[#]:

```
library(MASS)
library(ggplot2)
data(faithful)
faithful1 = faithful[faithful$eruptions<3.2.5,]
faithful2 = faithful[faithful$eruptions>3.2.5,]
summary(faithful1)
summary(faithful2)
cor(faithful1$eruptions, faithful1$waiting)
[1] 0.3511609
cor(faithful2$eruptions, faithful2$waiting)
[1] 0.3537682
duration = faithful$eruptions
waiting = faithful$waiting
plot <- ggplot(faithful, aes(x=duration, y=waiting)) + geom_point()
```

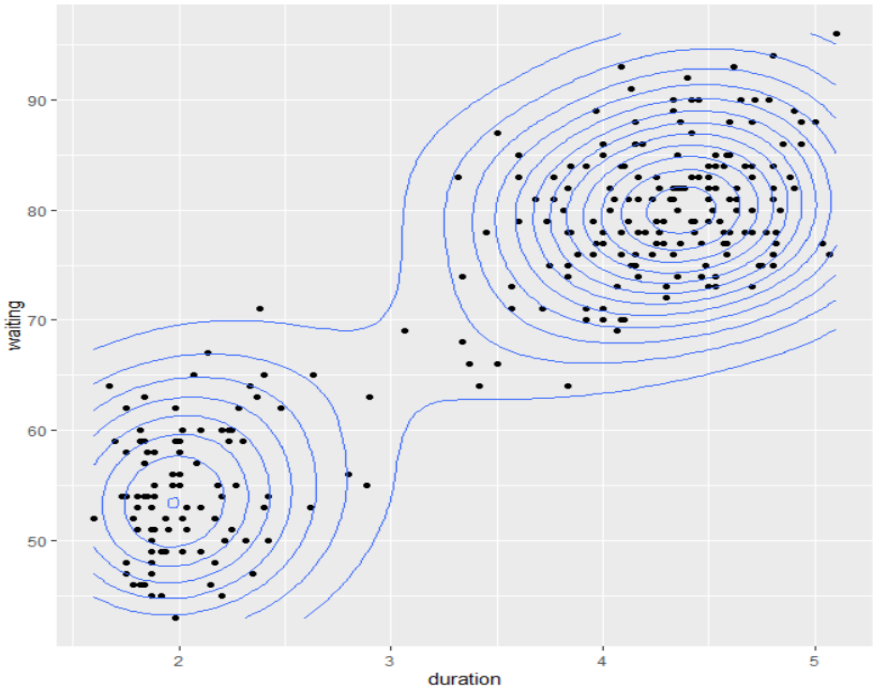
```
> summary(faithful1)
   eruptions        waiting
 Min.   :1.600   Min.   :43.00
 1st Qu.:1.833   1st Qu.:50.00
 Median :1.983   Median :54.00
 Mean   :2.049   Mean   :54.64
 3rd Qu.:2.200   3rd Qu.:59.00
 Max.   :3.067   Max.   :71.00
> summary(faithful2)
   eruptions        waiting
 Min.   :3.317   Min.   :64.00
 1st Qu.:4.037   1st Qu.:76.00
 Median :4.341   Median :80.00
 Mean   :4.298   Mean   :80.05
 3rd Qu.:4.583   3rd Qu.:84.00
 Max.   :5.100   Max.   :96.00
```

Firstly, we use the summary() function to get a basic idea of the data. We have divided the data into two groups, by using eruption time = 3.25 seconds as the dividing line. This choice is based on observing the data. We can see from the summary that waiting for 54 minutes we can expect to see an eruption of about 2 seconds while if we wait for 80 minutes we will likely see an eruption of about 4.5 seconds. This suggests that there might be a positive correlation between the two variables, however, upon calculation we see quite a weak correlation for each one, being between 3 and 3.5. Now we shall plot the function.

```
plot + geom_density_2d()
```

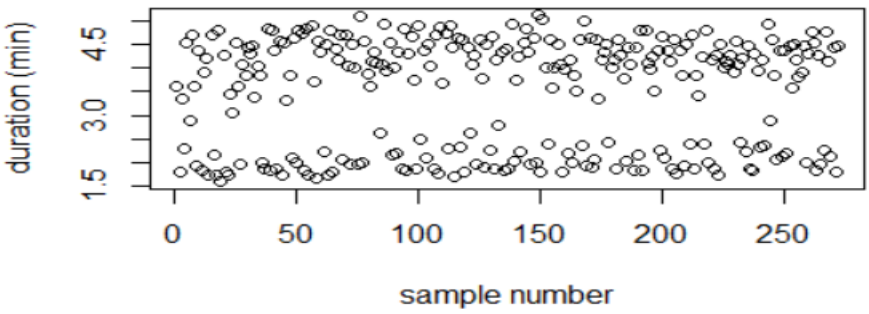
References: 1. https://en.wikipedia.org/wiki/Old_Faithful

: I have omitted ">" so it is easier to copy and paste the code in R, otherwise, it gives an error Also, any code on the further pages should be considered as a continuation of this one.



The graph shows a scatter plot of the data with a 2D contour plot superimposed on it. Observing the graph one can see why duration = 3.25 was chosen as the dividing line. The graph appears to have two clusters of elliptical nature suggesting two bivariate normal distributions of similar volume, shape, and orientation. We also see that there are more data points in the second group So eruptions are more likely

```
plot(duration, xlab="Sample Number", ylab="Duration (min)")
```



Furthermore, we can also analyse the data univariately and we see that plotting eruption times also shows two sub-groups within the data. Now we shall move on the formally analysing the data.

Analysing the data using Gaussian Mixture Model[3]:

A Gaussian mixture model is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters [2]. The model parameters can be estimated using the *Expectation-Maximization* (EM) algorithm initialized by hierarchical model-based clustering. Each cluster is characterised by three parameters:

- Mean vector : μ_k
- Covariance vector : $\mathbf{\Sigma_k}$
- Each point has a probability of belonging to each cluster

We start by calling on the Mclust package in R and using it to fit the data by EM algorithm.

```
library(mclust)
fit <- Mclust(faithful)  # Model-based clustering
summary(fit)
```

References: 2. https://scikit-learn.org/stable/modules/mixture.html
3. https://www.datanovia.com/en/lessons/model-based-clustering-essentials/

```
> summary(fit)
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust EEE (ellipsoidal, equal volume, shape and orientation) model with 3
components:

 log-likelihood   n df       BIC       ICL
      -1126.326 272 11 -2314.316 -2357.824

Clustering table:
  1   2   3
 40  97 135
```

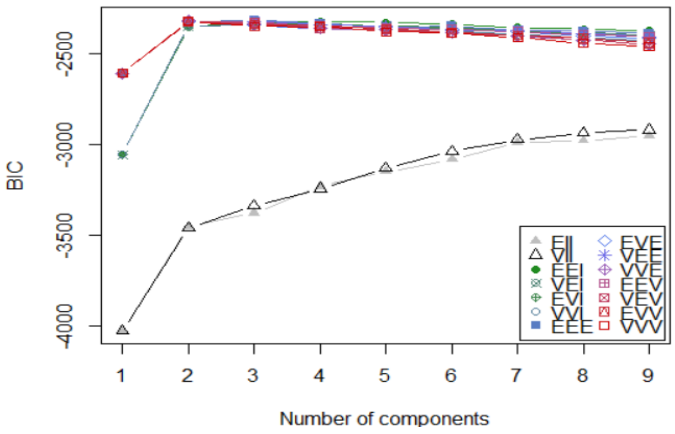
This gives us a plethora of information including the number of clusters, the model, the BIC, the log-likelihood and the number of data points in each cluster.

Now we know that our initial assumption was WRONG! And that there are three ellipsoidal clusters of equal volume, shape, and size in p-dimensions. To confirm the number of clusters we find the BIC[4].

```
bic <- mclustBIC(faithful)
plot(bic)
```

We see that it is not clearly discernible from the graph that the number of clusters is indeed 3. So we find the summary.

```
summary(bic)
```




```
> summary(bic)
Best BIC values:
             EEE,3        VVE,2        VEE,3
BIC      -2314.316 -2320.432980 -2322.103490
BIC diff     0.000    -6.116684    -7.787194
```

We can clearly see that the model EEE with 3 clusters has the maximum BIC therefore it is the preferred model. Now we plot the function in several ways:

```
plot(fit)
```

This gives us four options, we can select the one we wish. Selecting 2, gives us the cluster plot for the classification of the model.

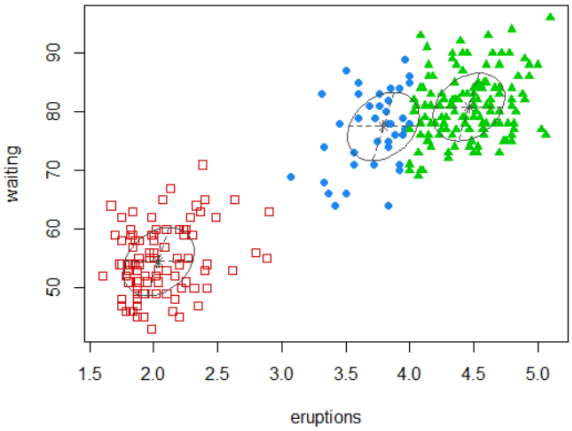
```
Model-based clustering plots:

1: BIC
2: classification
3: uncertainty
4: density

Selection:
```

1. Classification cluster graph:
   We are now sure that there are three clusters. To discern which one is which we use this code: faithful[1:3,]
   head(fit$classification, 3)

```
> faithful[1:3,]
  eruptions waiting
1     3.600      79
2     1.800      54
3     3.333      74
> head(fit$classification, 3)
1 2 3
1 2 1
```



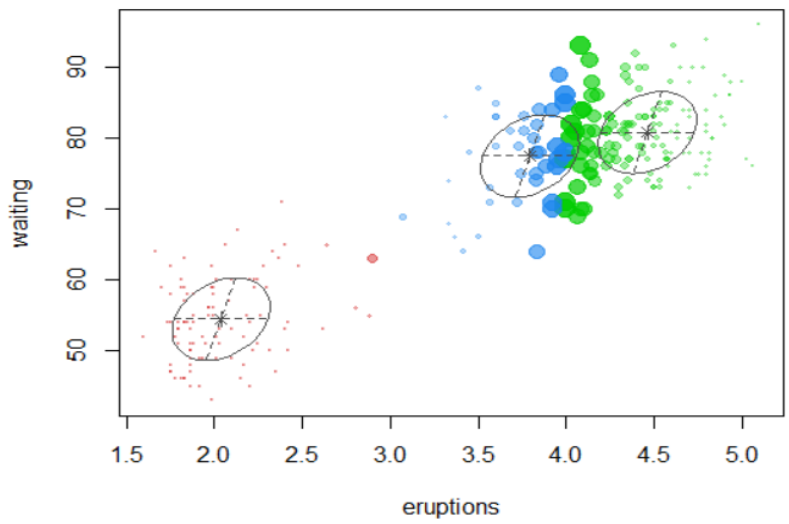
   We can see that the element is the 'blue' cluster which is the 1st cluster. Similarly, the second element is in the 'red' cluster which is the 2nd cluster. Therefore, the 'green' cluster is the 3rd cluster. Hence, using the clustering table we discern that the 'red' cluster

---

References: 4. https://www.datanovia.com/en/lessons/model-based-clustering-essentials/

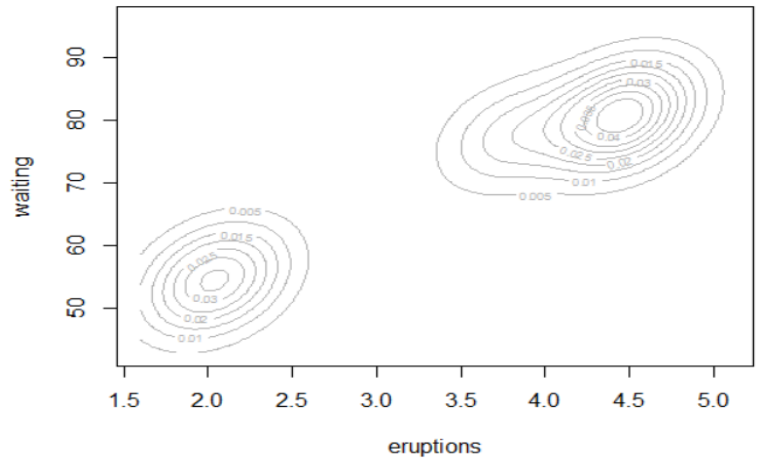
has 97 data points in it. The ‘blue’ cluster has 40 data points in it. And the ‘green’ cluster has 135 data points in it. As can be seen in the graph all three clusters have the same covariance structure which implies they have homogeneous covariance matrices.

2. Uncertainty cluster graph:
   Selecting 3 on the menu gives us this graph.



This graph shows the uncertainty inherent in the process of using conditional probability from EM to classify in model-based cluttering. Each data point has a probability associated with each cluster and therefore, the uncertainty arises. We can see that the ‘blue’ and ‘green’ clusters have much more uncertainty that the ‘red’ one, especially at the juncture of their overlap which is expected as these points are outliers.

3. Density graph:



Selecting 4 on the menu.
This graph shows the density of the clusters. The ‘green’ cluster is the most densely packed one as it has the most number of data points. Then, it is the ‘red’ followed by the ‘ blue’ which is quite sporadic and overlaps with the ‘green’ one. The ‘red’ cluster is the most uniformly dense of the three.

Conclusion:

Initially, casually observing the data we had we had made a couple follies, a positive correlation and a two cluster assumption. Both turned out to be wrong upon further analysis of the data. That does not mean that casual observation of a data is futile but rather that it gives us the right footing to approach further analysis of the given data. Even though the correlation might be quite weak there is clearly a relation between the two quantities. Perhaps, this works as a clue for the internal operations of the geyser. It may be that the more time it takes to erupt, the more the pressure builds up and therefore the eruption lasts longer. Since using the Gaussian Mixture Method along with the Expectation-Maximization algorithm provided us with three clusters one can then go one to discern depending on the waiting time in which cluster does a specific data point fit into and the likelihood of an eruption around that time. We also saw that it was less likely for an eruption to go off before 70 minutes than after. This knowledge will be extremely useful to anyone studying the workings of the Old Faithful geyser.