MATH38161 Coursework

Wei Chung-Yu 10680317

Nov 2023

1 Dataset

This project is part of the MATH38161 Coursework for task B. The document presents an examination of data related to the Old Faithful Geyser located in Yellowstone National Park, Wyoming, USA. This area, renowned for geyser watching, is the park's most convenient and welcoming spot. It features bench seating, a spacious parking lot, and an impressive visitor center equipped with facilities to monitor eruption details such as timing, height, and duration, enabling the prediction of the next eruption.[1].

The `geyser` data set in the `mass` R package comprises 299 sets of measurements, specifically two distinct features should be taken into account:

1. The time gaps between the commencement of consecutive eruptions.
2. The duration of the ensuing eruption.

```
> library(MASS)    # Loads the "MASS" package into R.
> dim(geyser)      # The dimension of the 'geyser' data set.
[1] 299   2
> head(geyser)     # This code displays the first 6 lines of the data set.
  waiting duration
1      80 4.016667
2      71 2.150000
3      57 4.000000
4      80 4.000000
5      75 4.000000
6      77 2.000000
```

Listing 1: Geyser data

2 Analyse method

We analyse the `geyser` dataset using Gaussian Mixture Models (GMMs). To perform the GMM analysis, we used the `mclust` package. This package offers capabilities for estimations using the EM algorithm in normal mixture models, encompassing diverse covariance structures, along with functions for simulating data from these models.

Step1 prepare the data by centering and standardising the data to ensure that the variables have a similar scale for GMMs.
Step2 scale the data by subtracting the mean and dividing the points by the standard deviation.
Step3 visualise the clustering result, a plot is generated, with each point assign a unique color corresponding to its cluster.

Step4, We identify the optimal number of clusters by fitting GMMs with different numbers of clusters and comparing their Bayesian Information Criterion (BIC) and their Akaike Information Criterion (AIC) to evaluate the effectiveness of the models. Both serve as tools for model selection, with AIC emphasising a balance between fit and complexity, while BIC places greater emphasis on model simplicity. In the last step, we compare the models by using the BIC and AIC criteria and find out which number of clusters have the lowest value for each criteria[2].

3 Results and discussion

Initially in listing2, the data is processed through centering and standardisation. Subsequently, Next, we generate 10 distinct models. This process assists in identifying the most optimal model by examining the BIC and AIC values. The optimal number of clusters corresponds to the point where both BIC and AIC values are at their lowest, and this will be the selected model. Finally, the `Mclust` function, sourced from the previously installed package, is employed to apply the EM algorithm to the processed data.

```
> library("mclust")
> # Center and standardise the variables.
> X.geyser <- scale(geyser, scale = TRUE)
> # Fit 10 models with clusters ranging from 1 to 10.
> models <- lapply(1:10, function(G) Mclust(X.geyser, G = G, verbose = FALSE))
> # Compute AIC.
> aicmodel <-function(model) {aic <- 2*model$df-2*model$loglik
+ return(aic)}
> # Compute BIC and AIC values for each model.
> BIC_dataset <- sapply(models, function(model) BIC(model))
> AIC_dataset <-sapply(models, compute_aic)
> # Find the best model using BIC data.
> gmm.geyser_BIC <- models[[which.min(BIC_data)]]
> summary(gmm.geyser_BIC)
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VVI (diagonal, varying volume and shape) model with 4
components:

 log-likelihood   n df       BIC       ICL
      -502.1488 299 19 -1112.606 -1142.556

Clustering table:
 1  2  3  4
90 17 98 94
> mclustModelNames("VVI")  # Explains the VVI model
$model
[1] "VVI"

$type
[1] "diagonal, varying volume and shape"
```

Listing 2: Information of Dataset

In listing 2, From Line19 to Line27, we can check the number of points in the dataset for each cluster. From Line28 to Line33, The table components in listing4 encompass diverse details concerning the Gaussian Mixture Model, incorporating information such as the Bayesian Information Criterion (BIC) value for the fitted model. This value serves as an indicator of the model's fit, taking into account the

number of clusters. It is noteworthy that the most suitable model is identified as the `mclust` type VVI and the optimal number of cluster according to BIC criterion is 4.
Correspondingly, the shapes of the clusters are diagonal in arbitrary size and shape. The selection of the VVI model is likely attributed to its optimal fit to the `X.geyser` data according to the BIC. Consequently, for the provided data[3]:

- **V (Volume):** The clusters exhibit diverse volumes, implying notable variations in spread or standard deviation within each cluster.
- **V (Shape):** The clusters display diverse shapes, suggesting that the distribution within each cluster may have varying degrees of elongation or skewness.
- **I (Diagonal Covariance Matrix):** This implies that the algorithm identified relatively uncorrelated variables within each cluster, simplifying the model by assuming a diagonal covariance matrix.

In summary, the `mclust` function's automatic model selection process likely chose the VVI model due to its optimal fit to the `X.geyser` data, as indicated by the BIC. This model provides flexibility in both the size and shape of the clusters, while assuming uncorrelated variables within each cluster.

```
# Plot the classification
plot(gmm.geyser, what = "classification")
# Add legend
legend("bottomleft", legend=paste("Cluster", 1:num_clusters), col=1:num_
    clusters, pch=0.5, title="Clusters")
```

Listing 3: Geyser Data with 4 Clusters

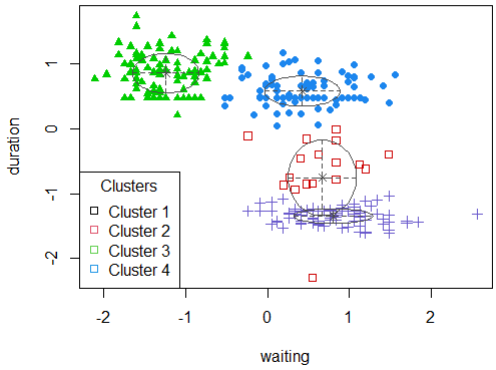

Figure 1: Output of Geyser Data with 4 clusters.

The graph depicted in Figure 1 represents the Gaussian Mixture Model (GMM) in which data points from the geyser dataset are categorized into various colors corresponding to their clusters, as indicated in the lower-left section of the chart. The visualisation indicates the presence of four distinct groups with minimal overlap, suggesting effective data separation by the GMM.

```
> # Find the suitable model for AIC
> gmm.geyser_AIC <- models[[which.min(AIC_data)]]
> summary(gmm.geyser_AIC)
```

```
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VEV (ellipsoidal, equal shape) model with 8 components:

 log-likelihood   n df       BIC       ICL
      -456.7189 299 40 -1141.455 -1254.522

Clustering table:
 1  2  3  4  5  6  7  8
11  7 78 19 64 74 18 28
> mclustModelNames("VEV") # Explains the VEV model
$model
[1] "VEV"

$type
[1] "ellipsoidal, equal shape"
> plot(gmm.geyser_AIC, what = "classification")
> legend("bottomleft", legend=paste("Cluster", 1:8), col=1:8, pch=1, title="
    Clusters")
```

Listing 4: Classification result

From the summary test using the AIC data in listing4, we can get the best model is VEV, and this model is described as ellipsoidal and equal shape.
Figure 2 is the Gaussian Mixture Model fitted for the ideal number of clusters based on the AIC data, which is 8, and it is clear to see that the groups exhibit some overlap, making it a bit challenging to discern them clearly. Also 8 clusters are differentiated into different colours based on their group.

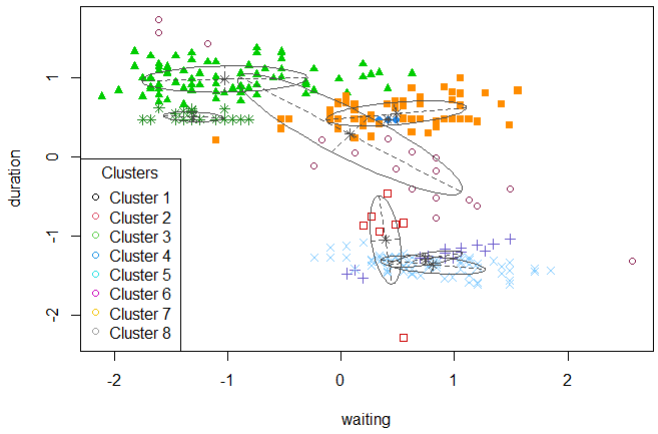

Figure 2: Output of Geyser Data with 8 clusters.

```
plot(1:10, BIC_data, type = "b", pch = 16, col = "green", xlab = "Number of
    Clusters",
     ylab = "BIC",ylim = c(950, 1600))
points(1:10, AIC_data, type = "b", pch = 16, col = "purple")

legend("bottomleft", legend = c("BIC", "AIC"), col = c("green", "purple"), pch
    = 16)
```

Listing 5: VVI model

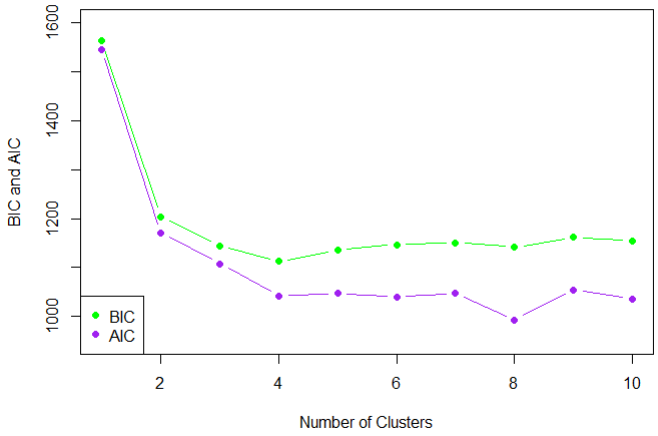

Figure 3: BIC and AIC for Gaussian Mixture Models

In the final part of this report, we are going to plot the BIC and AIC data with different number of clusters. Since BIC has the loweest value when the number of clusters is 4, and AIC has the lowest value when the nnumber of clusters is 8.
In general, BIC is more stringent in penalizing models with numerous parameters compared to AIC. Therefore, it is commonly employed when the aim is to identify a simpler, more parsimonious model. Nonetheless, both AIC and BIC serve the purpose of comparing various models and determining the most suitable one for a given dataset[4].

References

[1] Kwak-Hefferan, Elisabeth. "About Old Faithful, Yellowstone's Famous Geyser." Yellowstone National Park, 13 Apr. 2021. Available at: `https://www.yellowstonepark.com/things-to-do/geysers-hot-springs/about-old-faithful/`

[2] Korbinian Strimmer.3 Unsupervised learning and clustering page109—Multivariate Statistics and machine learning. Available at: `https://online.manchester.ac.uk/webapps/blackboard/execute/content/file?cmd=view&content_id=_15341066_1&course_id=_78684_1/`

[3] Erar, Bahar. Mixture Model Cluster Analysis under Different Covariance Mixture Model Cluster Analysis under Different Covariance Structures Using Information Complexity Structures Using Information Complexity. 2011. Available at: `https://trace.tennessee.edu/cgi/viewcontent.cgi?article=2096&context=utk_gradthes&httpsredir=1&referer=`

[4] Menear, Kevin. "Comparing Clustering Methods: Using AIC and BIC for Model Selection." Medium, 8 Feb. 2023 Available at: `https://medium.com/@kevin.menear/comparing-clustering-methods-using-aic-and-bic-for-model-selection-bf80d0d37ec2`