Model Based Cluster Analysis: Wines produced in USA (Wine Dataset)

Himanshu Gupta

1. Use model-based clustering methods to categorise the wines from the USA based on price and points rating. Can you identify any clusters that are good value for money?

Introduction:

The objective of the cluster analysis is to categorize the wines from the USA based on two features namely price and point ratings. We use model based clustering to achieve this objective which considers the data points to be a part of distribution which is an amalgamation of two or more clusters.

Cluster Analysis Using Model Based Clustering:

In general the clustering technique uses unsupervised learning to create clusters based on the similarities measure.
It is to note that k-mean clustering requires number of clusters as an input and is randomly initialized giving different set of results. So all these clustering methods decide clusters based on data merely. While model based methods use probabilities or uncertainties to assign data points to any cluster.

Here iteratively data points are fit into clusters by optimizing assignment of data points in clusters. It works in 3 steps:

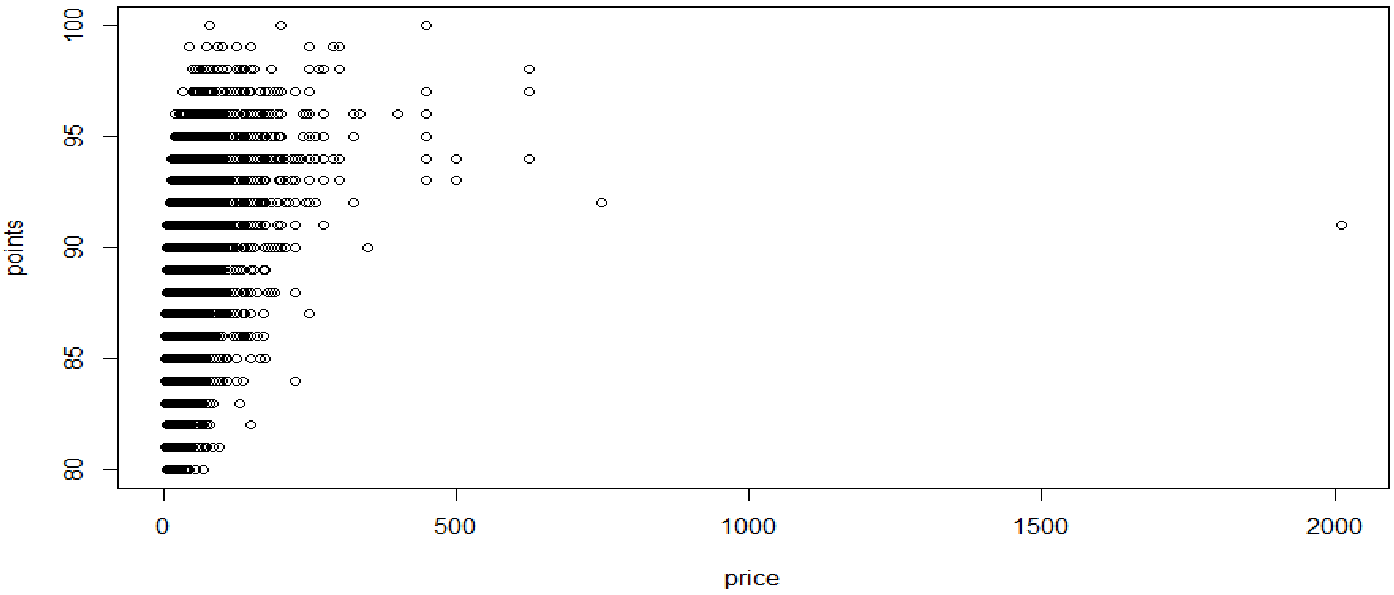
1. Select random Gaussian parameters and use to fit various data records.
2. In order to fit good number of points, distribution parameters can be updated iteratively.
3. When the local minima is achieved using iterative updates, the data records can be allotted region nearer to the distribution cluster.

Data Analysis: Wine Review Dataset Kaggle

We loaded the wine data into R workbook. For modelling the data as a normal finite mixture we used mclust package for analysis. As per problem, we filtered out the data points corresponding to USA and selected the required features as points and price for our analysis. The records having 'NA' values were dropped as they were not providing much information. We did exploratory analysis but knowing the insight about the dataset given.

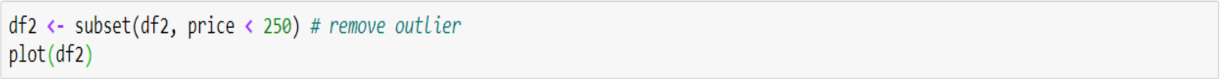
The graph has been plotted between Points vs. Price to find the correlation between these features as below:

Fig-15



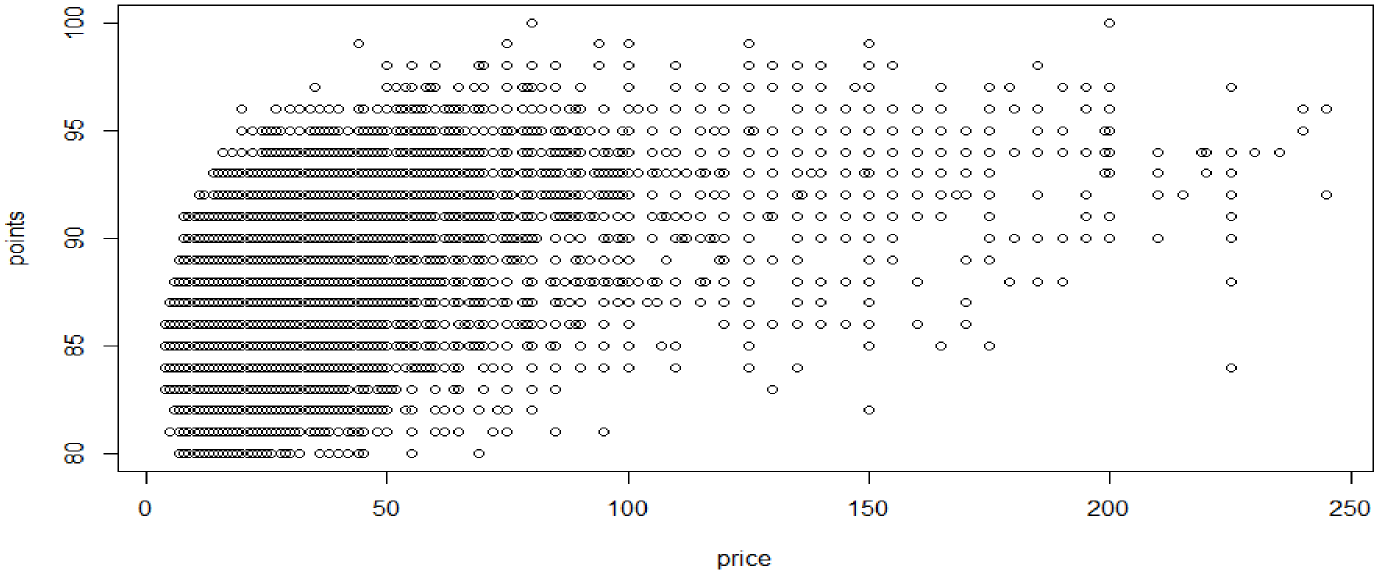
It can be inferred from above that there is no relation between variables points and price. Also we may observe few outliers in the above plot which may restrict the correct cluster formation. So outliers were removed as shown below:

Code snippet: (Removing outliers)



```
df2 <- subset(df2, price < 250) # remove outlier
plot(df2)
```

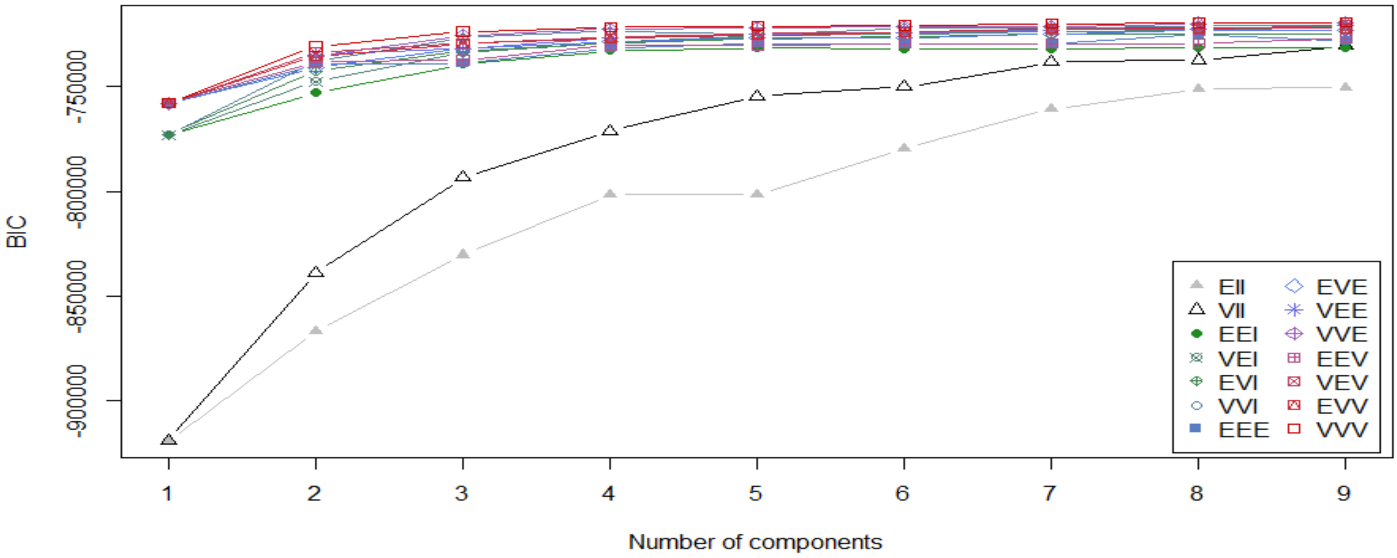
Removed outliers having value equal to and more than 350 (2000 one is removed): (Fig-16)



It can be observed that data has been cleaned and now data modelling can be performed. Using the function mclust, the data can be fit to obtain a model. We haven't pass any parameter to this function, so all the default values will be used to fit Gaussian finite model using EM algorithms. Also the covariance matrix is used to depict the geometry of clusters that is Volume, shape and orientation of clusters. These are termed as constraints and combination of these yields models: ellipsoidal, spherical and diagonal. Now function mclust will identify the models out of 14 which best describe the data. For model selection BIC (Bayesian information criterion) is used which penalizes the complexity of models with large number of clusters. When using the optimal number of cluster 'G' as 'null' will impose mclust to use 1-9 clusters and choose components based on BIC. It is important to note that lower BIC will give better fit and increase in likelihood will overfit the model.

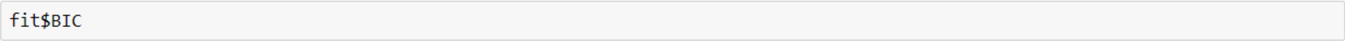
Below is the plot for BIC vs. number of components:

Fig-17



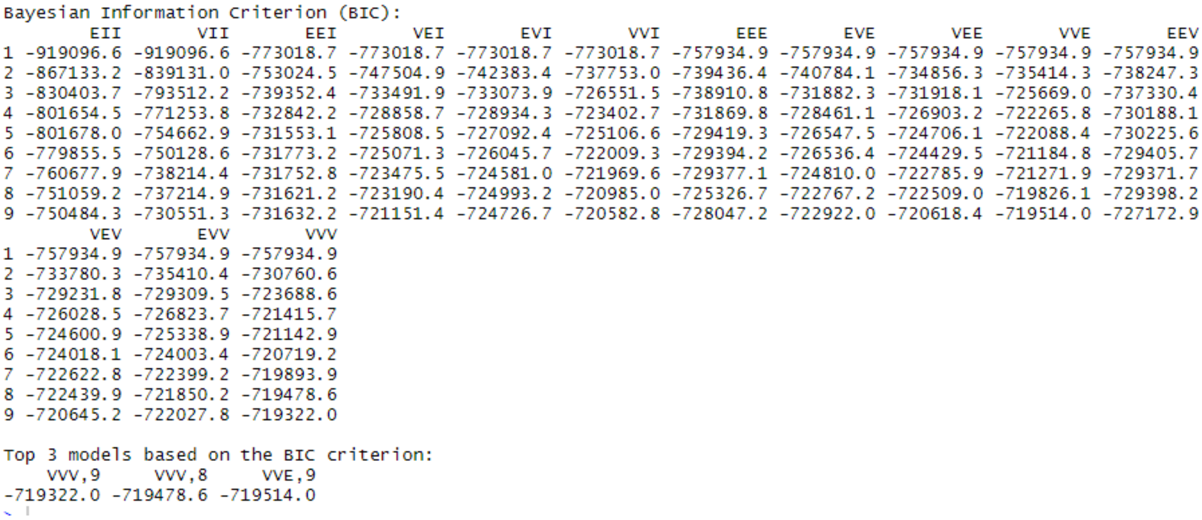
The plotted figure above represents number of clusters, various covariance structures and their respective BIC values. The line graph plotted shows the increase in BIC values as the number of clusters increases. To describe BIC values for all 14 models we use:

Code to describe BIC for all models:



```
fit$BIC
```

Results:



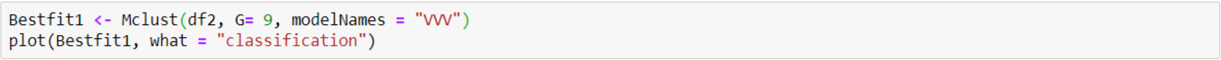
```
Bayesian Information Criterion (BIC):
        EII       VII       EEI       VEI       EVI       VVI       EEE       EVE       VEE       VVE       EEV
1 -919096.6 -919096.6 -773018.7 -773018.7 -773018.7 -773018.7 -757934.9 -757934.9 -757934.9 -757934.9 -757934.9
2 -867133.2 -839131.0 -753024.5 -747504.9 -742383.4 -737753.0 -739436.4 -740784.1 -734856.3 -735414.3 -738247.3
3 -830403.7 -793512.2 -739352.4 -733491.9 -733073.9 -726551.5 -738910.8 -731882.3 -731918.1 -725669.0 -737330.4
4 -801654.5 -771253.8 -732842.2 -728858.7 -728934.3 -723402.7 -731869.8 -728461.1 -726903.2 -722265.8 -730188.1
5 -801678.0 -754662.9 -731553.1 -725808.5 -727092.4 -725106.6 -729419.3 -726547.5 -724706.1 -722088.4 -730225.6
6 -779855.5 -750128.6 -731773.2 -725071.3 -726045.7 -722009.3 -729394.2 -726536.4 -724429.5 -721184.8 -729405.7
7 -760677.9 -738214.4 -731752.8 -723475.5 -724581.0 -721969.6 -729377.1 -724810.0 -722785.9 -721271.9 -729371.7
8 -751059.2 -737214.9 -731621.2 -723190.4 -724993.2 -720985.0 -725326.7 -722767.2 -722509.0 -719826.1 -729398.2
9 -750484.3 -730551.3 -731632.2 -721151.4 -724726.7 -720582.8 -728047.2 -722922.0 -720618.4 -719514.0 -727172.9
        VEV       EVV       VVV
1 -757934.9 -757934.9 -757934.9
2 -733780.3 -735410.4 -730760.6
3 -729231.8 -729309.5 -723688.6
4 -726028.5 -726823.7 -721415.7
5 -724600.9 -725338.9 -721142.9
6 -724018.1 -724003.4 -720719.2
7 -722622.8 -722399.2 -719893.9
8 -722439.9 -721850.2 -719478.6
9 -720645.2 -722027.8 -719322.0

Top 3 models based on the BIC criterion:
    VVV,9     VVV,8     VVE,9 
-719322.0 -719478.6 -719514.0 
```

From above BIC details, it can be observed that top three models which best fits are (VVV, 9), (VVV, 8) and (VVE, 9) while model VVV having group size 9 has highest BIC values. The model's geometry is decided based on volume, shape and orientation. It can be seen that models having general or ellipsoidal cluster, fits the given wine data best as compared to diagonal and spherical.

The mclust function is also use to depict the classification and density estimation as well. Using the model VVV with group size 9, the different plots are:

Classification: Best fit where G = 9

Code snippet

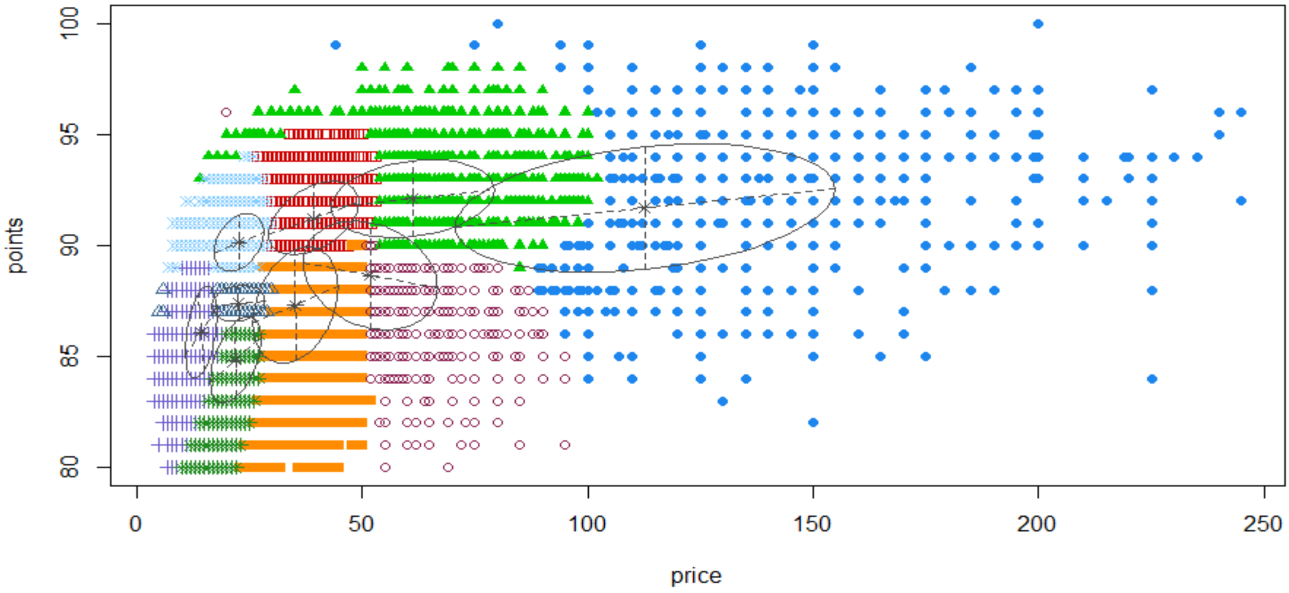


```
Bestfit1 <- Mclust(df2, G= 9, modelNames = "VVV")
plot(Bestfit1, what = "classification")
```

Graph: **Fig-18**

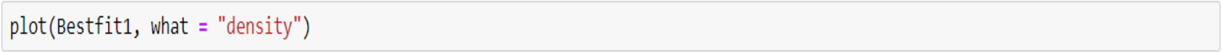
Model Based Cluster Analysis: Wines produced in USA (Wine Dataset)

Himanshu Gupta



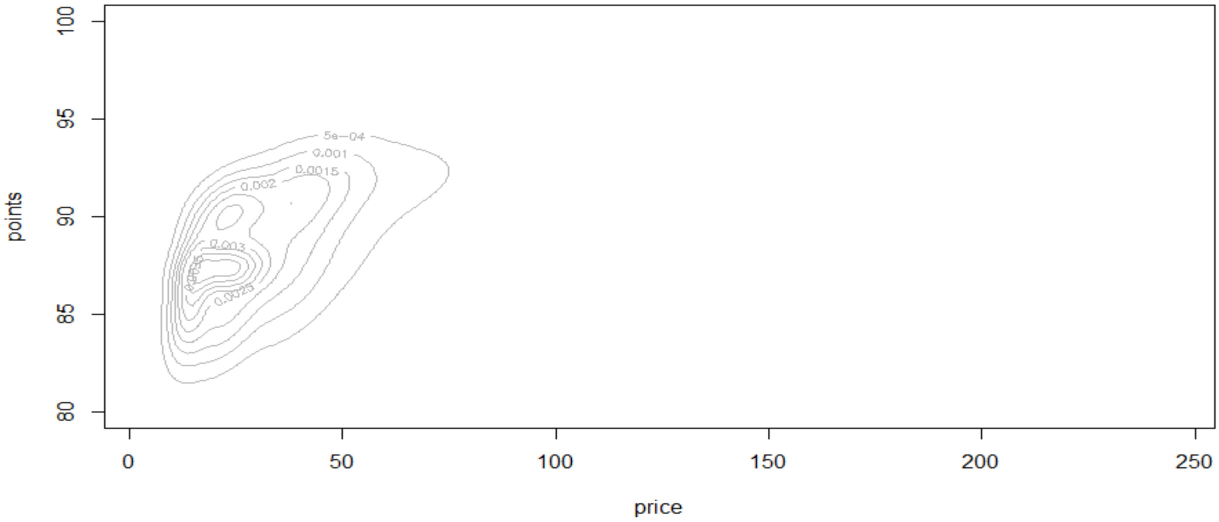
Density Estimation:

Code:



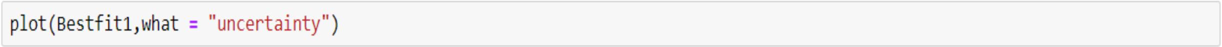
```
plot(Bestfit1, what = "density")
```

Graph: **Fig-19**



Uncertainty:

Code:

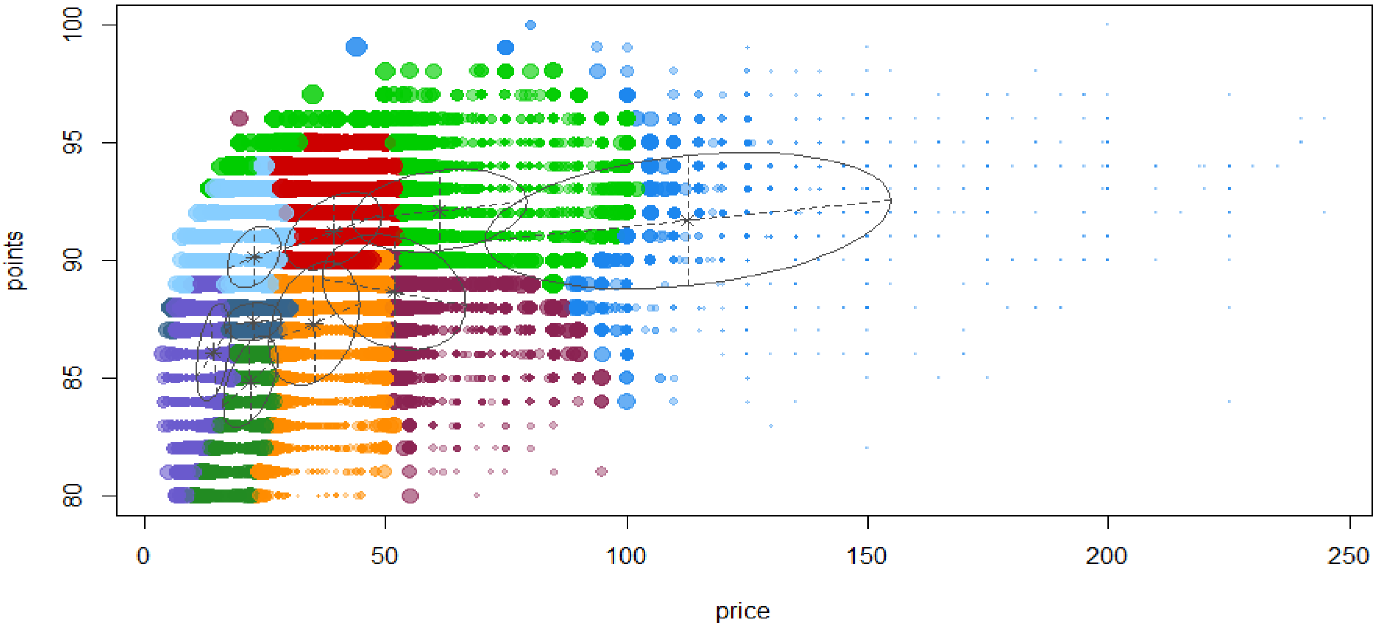


```
plot(Bestfit1,what = "uncertainty")
```

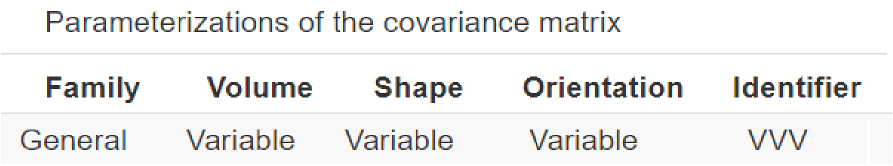
Graph: Fig-20

Model Based Cluster Analysis: Wines produced in USA (Wine Dataset)

Himanshu Gupta



Below shows various contarints with best **fitted model VVV, having group size 9**.

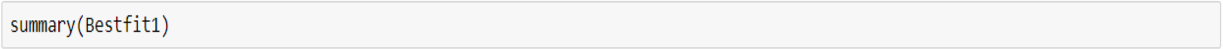


Parameterizations of the covariance matrix

Family	Volume	Shape	Orientation	Identifier
General	Variable	Variable	Variable	VVV

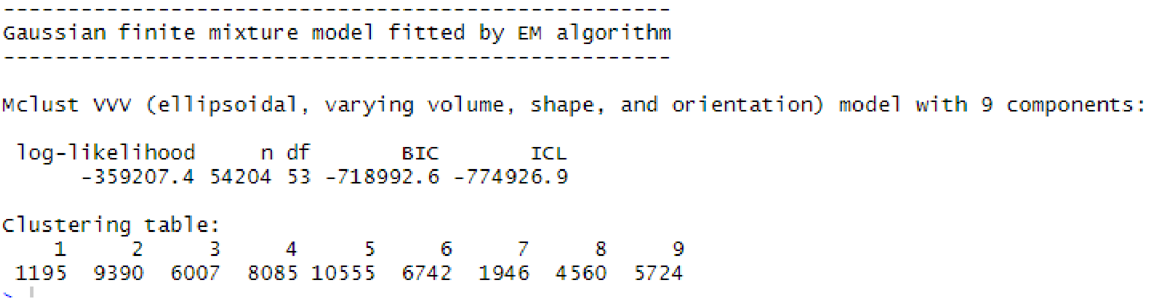
Conclusion:

Code:



```
summary(Bestfit1)
```

Result:



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

Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 9 components:

 log-likelihood     n df       BIC       ICL
      -359207.4 54204 53 -718992.6 -774926.9

Clustering table:
    1     2     3     4     5     6     7     8     9
 1195  9390  6007  8085 10555  6742  1946  4560  5724
```

From various graphs plotted for this problem and above data summary, we can estimate that clusters having high points rating and low price values can be considered as clusters that are good for money. We conclude from the summary that optimal model here **is VVV,9 with BIC = -718992.6** identifies **9 clusters** with cluster() having most number of data point s We need to find the cluster which is good value for money, i.e., more rating and less priced. For this we consider the cl assification plot. The clusters marked **on red, light sky blue colours** can be called as good value for money as they repres ent wines with good rating and fewer prices.