# Model-based clustering with mclust R package: Multivariate assessment of mathematics performance of students in Qatar

**Ali Rashash R Alzahrani [a,b], Eric J. Beh [a] and Elizabeth Stojanovski[a]**

*[a] School of Information and Physical Sciences, University of Newcastle, Newcastle, Australia, [b] Mathematical Sciences Department, College of Applied Sciences, Umm Al-Qura University*
*Email: alirashashr.alzahrani@uon.edu.au*

**Abstract:** This study demonstrates how model-based clustering can be undertaken using `mclust`, a contributed R package, to examine factors influencing mathematics performance of high school students in Qatar. Although there are numerous cluster analysis approaches, this paper highlights the intricacies, assumptions, limitations, benefits and pitfalls of clustering using a model-based approach, and how the inherent inadequacies of other clustering approaches can be better explored using model-based methods. Moreover, this paper demonstrates how the `mclust` package can be used to concurrently analyse and compare different models, in order to select the preferred clustering model according to the Bayesian information criterion, and to estimate parameters of the associated model using maximum likelihood estimation. The benefit of selecting a prior to avoid model-based clustering estimation singularity- and degeneracy-related issues offers an alternative approach to improve the rate of convergence. The results from applying model-based clustering using `mclust` to educational data that examines the mathematics performance of secondary students in Qatar will be used to identify factors that influence mathematics performance for different clusters of students, to help facilitate potential adoptions of the most appropriate remedial teaching strategies to implement to enhance learning. Furthermore, the results can help teachers to identify groups of students whose performance in different subject areas is likely to be affected by certain factors, thereby helping them to reduce potentially undesirable learning outcomes.

***Keywords:*** *Hierarchical clustering, mclust, Bayesian information criterion, Model-based clustering*

## 1. INTRODUCTION

Various clustering approaches exist, including density-based, distribution- or model-based clustering (Fraley & Raftery, 2002), centroid (Niraj et al., 2013), k-means (Hartigan & Wong, 1979; Lloyd, 1982) and hierarchical clustering (Gordon, 1987). The focus of this paper is to describe the model-based clustering methodology using the `mclust` (Fraley et al., 2020; R Core Team, 2020). The `mclust` employs finite normal mixture modelling to determine model-based clustering. By treating the entire population as a mixture of sub-populations, with the latter identified as clusters, individual elements of this mixture are modelled by conditional probability distributions. This method therefore models individual elements of this mixture by means of conditional probability distributions. The `mclust` implements several different models using maximum likelihood estimation and the Bayesian criterion for selecting the model that is most likely to determine the appropriate number of clusters. As described by Grün (2018) and Fraley et al. (2012), the versatility of the model-based clustering methodology is evident by its diverse applications. By applying the model-based clustering approach, geometric characteristics of the clusters including the orientation, shape and volume of the clusters become of interest. These aspects are determined based on the covariance matrices which are typically approximated based on the underlying data, with the distribution of each group in the mixture model either ellipsoidal, diagonal or spherical.

## 2. METHODS

### 2.1. Data source

The variables assessed for the present study were based on Programme for International Student Assessment data collected from secondary school students in Qatar (OCED, 2013). A total of 10,966 students from 154 schools participated in the study. Each student was asked a series of questions related to learning mathematics in the classroom and based on their responses, their a measure of their mathematics teachers' use of 12 different teaching strategies was formed. Each teaching strategy was scored from 1 to 4 based on the extent to which the teacher utilised the technique, using a scale from 'never' to 'always'. A measure of parental education levels was also provided as a proxy measure of socio-economic status, with measures scored from 0 to 6, ranging from 'less than primary school' to 'postgraduate level' (Alzahrani and Stojanovski, 2019). A students' performance in mathematics was determined from a questionnaire comprising 35 mathematics questions. Students' mathematics performance along with the student demographic variables and teaching strategies were analysed to establish how these varied using model-based clustering approaches.

### 2.2. Parameterisation of covariance matrices in `mclust`

The most general covariance matrix for a mixture of G clusters is defined by $\Sigma_k$, where $k$ represents the $k^{th}$ group among G clusters in the mixture model, so that each cluster has its own covariance matrix. Here, $\Sigma_k$ is a P × P symmetric covariance matrix that contains variances for the P variables along the main diagonal, and covariances between pairs of variables along the off-diagonal for the $k^{th}$ cluster. Following the method of Banfield and Raftery (1993) and Celeux and Govaert (1995), the within-cluster covariance matrix can be decomposed by $\Sigma_k = \lambda_k D_k A_k D_k^T$. The orthogonal matrix of eigenvectors is denoted by $D_k$, a P × P matrix. Corresponding to each eigenvector, the diagonal matrix of scaled eigen-values of $\Sigma_k$ is denoted as $A_k$, is a P × P diagonal matrix. This is scaled to have a determinant equal to 1, with a scalar value $\lambda_k$ factored from the eigen-value matrix to ensure the determinant is equal to 1.

The covariance matrix has several possible parameterisations that can be considered. `mclust` employs unique individual identifiers for each possible covariance matrix parameterisation; *I* for coordinate axes, *V* for variable and *E* for equal, with 14 different combinations by including different constraints on volume, shape and orientation of clusters. Table 1 summarises the available options for each model obtained from the decomposition of $\Sigma_k$.

**Table 1.** Within-Cluster Covariance Matrix Parameterisations for Multidimensional data available in `mclust`, alongside corresponding Geometrics features

| Model | $\Sigma_k$ | Distribution* | Volume** | Shape*** | Orientation**** |
|---|---|---|---|---|---|
| EII | $\lambda I$ | Spherical | Equal | Equal | Not available |
| VII | $\lambda_k I$ | Spherical | Variable | Equal | Not available |
| EEI | $\lambda A$ | Diagonal | Equal | Equal | Coordinate axes |
| VEI | $\lambda_k A$ | Diagonal | Variable | Equal | Coordinate axes |

| | | | | | |
|---|---|---|---|---|---|
| EVI | $\lambda A_k$ | Diagonal | Equal | Variable | Coordinate axes |
| VVI | $\lambda_k A_k$ | Diagonal | Variable | Variable | Coordinate axes |
| EEE | $\lambda DAD^T$ | Diagonal | Equal | Equal | Equal |
| EVE | $\lambda DA_k D^T$ | Ellipsoidal | Equal | Variable | Equal |
| VEE | $\lambda_k DAD^T$ | Ellipsoidal | Variable | Equal | Equal |
| VVE | $\lambda_k DA_k D^T$ | Ellipsoidal | Variable | Variable | Equal |
| EEV | $\lambda D_k AD_k^T$ | Ellipsoidal | Equal | Equal | Variable |
| VEV | $\lambda_K D_k AD_k^T$ | Ellipsoidal | Variable | Equal | Variable |
| EVV | $\lambda D_k A_K D_k^T$ | Ellipsoidal | Equal | Variable | Variable |
| VVV | $\lambda_K D_k A_K D_k^T$ | Ellipsoidal | Variable | Variable | Variable |

* Spherical, diagonal and ellipsoidal refer to the distribution restrictions in each of P variables' directions.
** In terms of Volume, 'equal' indicates that all k clusters are constrained to the same volume ($\lambda_k$) and 'variable' means that each k cluster could take on a different volume ($\lambda_k$).
*** In terms of Shape, 'equal' implies that each k cluster is restricted to the same shape.
**** In terms of Orientation (orthogonal matrix of eigenvectors), 'equal' means the k clusters are constrained to the same orientation ($D_k$) and 'variable' means that each k cluster can have a different orientation ($D_k$). 'Coordinate axes' means that the model is diagonal in distribution, so the ellipse orientation in all P dimensions should be oriented in one of P dimension's direction. 'Not available' means that the clusters do not have a specific orientation and are like a circle in each P dimension cross-section.

Interpreting the combination of different unique individual identifiers used by `mclust` depends on the used parameterisations. For example, EVI shows all clusters have equal (*E*) volume, varying (*V*) cluster shapes, and coordinate axes orientation (*I*). If the P variables are uncorrelated within clusters, the covariances are diagonal with parameters related to volume, shape and orientation along the coordinate axes—$\lambda_k$ and $A_k$, based on the data. In the Gaussian mixture, if the variables used in clustering are uncorrelated and have the same variation in all P directions within each cluster, the clusters are spherical where the variance of all parameters is identical in each cluster. If the variances of the P variables are different within a cluster but uncorrelated, the mixture will be diagonal. In most cases, the distribution of the Gaussian mixture has different variations and features tend to be correlated, and so the covariance matrix of the clusters is ellipsoidal. The centroid of every cluster is given by its mean ($\mu_k$), with orientation, shape and volume of each cluster a function of its covariance matrix.

### 2.3 Expectation–Maximisation Algorithm

The likelihood function for the *G* components Gaussian mixture is given by the product of mixture densities, where each density is expressed as the sum of the mixture proportions multiplied by the component densities for each group. The latter follows a multivariate normal distribution. Since the cluster membership of each observation is unknown, an iterative two-step expectation maximisation (EM) algorithm is used to obtain maximum likelihood estimates. The algorithm involves an expectation step to estimate the probabilities of cluster membership of each observation, and a maximisation step to estimate unknown model parameters, and iterates until convergence. For a Gaussian mixture model, the probability density function for an observation is given as:

$$f(\mathrm{x};\ \theta) = \sum_{k=1}^{G} \pi_k\, \phi(\mathrm{x};\ \mu_k, \Sigma_\mathrm{k})$$

where *G* is the total number of clusters, $\mu_k$is the mean vector of P variables in the data with a size of P × 1 for the $k^{th}$ cluster, $\Sigma_k$ is a P × P covariance matrix of the $k^{th}$ cluster, and $\pi_k$ is the mixture proportion for the $k^{th}$ cluster ($\sum_{k=1}^{G} \pi_k = 1$). We denote θ to be the vector of parameters for the mixture model, such that $\theta = (\pi_1, \dots, \pi_{G-1}, \mu_1, \dots, \mu_G, \Sigma_1, \dots, \Sigma_\mathrm{G})$. The term $\phi(\mathrm{x};\ \mu_k, \Sigma_\mathrm{k})$ is the density of the Gaussian distribution withmean vector, $\mu_k$, and covariance matrix $\Sigma_\mathrm{k}$ (Erar, 2011).
The EM algorithm is the algorithm typically employed for model-based clustering which uses starting values for the parameters in the model and iteratively applies two steps to estimate the parameters (Erar, 2011; Melnykov and Maitra, 2010). The expected probabilities for assigning data points to each cluster are determined by use of initial values for the parameters. The second step is the maximisation step where the assigned probabilities are used as weights to estimate optimal model parameters for each cluster. The Bayes formula is used in the expectation step, and also in each consequent step, to update the cluster probabilities for each observation. In the maximisation step, the mean and covariance matrix of Gaussian distribution are weighted with the cluster probability for each observation, with repetition of these two steps until convergence of the algorithm, at which point the parameter maximum likelihood estimates are determined. Selecting an

appropriate set of initial values to start the EM algorithm is an important decision, as convergence of the algorithm may depend on this.

### 2.4 Maximum a posteriori simulation

If a prior distribution is used, $\mu$ and $\Sigma_k$ can be derived from the conjugate prior of the normal distribution. In this case, the parameters estimated by the EM algorithm are the 'maximum a posteriori' (MAP) estimates, as they are derived by multiplying the prior probability with the likelihood of the parameters, and hence is not maximum likelihood estimation. From the formula proposed by Fraley et al. (2012), $\mu$ is simulated by a normal distribution with mean $\mu_p$ (the mean of the data for each parameter) and variance $\Sigma/\kappa_p$, where $\kappa_p$ is the shrinkage parameter equal to 0.01, $\Sigma$ is simulated by the inverse Wishart distribution with $\nu_p$ degrees of freedom and $\Lambda_p$ is the scale of the prior distribution: $\mu|\Sigma \sim N\left(\mu_p, \frac{\Sigma}{\kappa_p}\right)$, such that $\Sigma \sim \mathrm{IW}(\nu_p, \Lambda_p)$, where $\nu_p = P + 2$, $P$ is the dimension of the data, $\Lambda_p = \frac{var(data)}{G^{2/P}}$ is the scale of the prior distribution and $G$ is the number of clusters.

### 2.5 Bayesian Regularisation

Bayesian regularisation (or prior control) is used to eliminate singularities that arise when using the EM algorithm to estimate model-based clustering (Fraley & Raftery, 2007). The approach is based on a dispersed conjugate prior and determines a MAP estimator. Prior control aids model selection by using an adaptation of a modified BIC (Lomet et al., 2012) as the parameters are estimated by maximising the posterior instead of the likelihood function. Given that the estimated parameters could differ in MAP, the value of the likelihood used in BIC could also differ from the likelihood derived by maximum likelihood estimation (Xue et al., 2017). Therefore, prior control helps to reduce singularities that are common with maximum likelihood approaches, without affecting stability of the results.

In model-based clustering parameterisations through eigen-decomposition of within-cluster covariances, the value with precise eigenvector normalisation constraint permits cross-component restrictions on the orientation, volume and shape of component mixtures that are normal. This is a type of Bayesian regularisation (Fraley & Raftery, 2002). Models could be constrained to comprise of varying or fixed variances across the components. Models with less variation have less chance of singularity issues as some elements in restricted models of decomposed matrices can't differ across clusters (Fraley & Raftery, 2007). Such components are derived using all observations, reducing the chance of failing due to zero determinants relative to models with no constraints (i.e. when volume, shape and orientation can differ) in any cluster (Fraley & Raftery, 2007).

Parameters of the scalar, diagonal and full matrix (i.e. volume, shape and orientation) need to be estimated. Inclusion of a low variance in each mixture component defining the clusters may lead to singularity issues and hence are more likely to have less model flexibility. Spherical and diagonal models have diagonal matrices and so can have singularity issues if there is a parameter in the dataset with zero variance in a cluster.

Model-based clustering was investigated using the `mclust` package to fit the data for the present study, since the data comprised a mixture of multivariate normal distributions. Given that the inverse of the covariance matrix is used in the theory of the multivariate normal distribution, the covariance matrix is decomposed into orientation, volume and shape, each of which could be chosen as variable (*V*) or constrained as equal (*E*) for all clusters. The choice of whether to constrain or treat as variable is determined by comparing the BIC in each fitted model. A greater likelihood is more common among complex models (with numerous predictors and a restricted covariance matrix). Free parameters are estimated by maximising the product of the prior and likelihood function in the two steps of the EM algorithm.

**Table 2.** Number of Free Parameters used to calculate the BIC

| Model | Free mixture proportions | Free mean parameters | Free covariance parameters | Total number of free parameters $\boldsymbol{m}$ |
|---|---|---|---|---|
| EII | G – 1 | GP | 1 | G + GP |
| VII | G – 1 | GP | G | 2 G + GP − 1 |
| EEI | G − 1 | GP | P | G – 1 + GP + P |
| VEI | G − 1 | GP | P + G − 1 | GP + P+2 G − 2 |
| EVI | G − 1 | GP | GP – G + 1 | 2GP |

| | | | | |
|---|---|---|---|---|
| VVI | G − 1 | GP | GP | 2GP + G − 1 |
| EEE | G − 1 | GP | P(P + 1)/2 | GP + G – 1 + P(P + 1)/2 |
| EEV | G − 1 | GP | GP(P + 1)/2 −(G − 1)P | GP(P + 3)/2 − (G − 1)(P − 1) |
| VEV | G − 1 | GP | GP(P + 1)/2 −(G − 1)(P − 1) | GP(P + 3)/2 − (G − 1)(P − 2) |
| VVV | G − 1 | GP | GP(P + 1)/2 | GP(P + 3)/2 + G − 1 |

For the purposes of model comparison, BIC employs a penalty term on the number of parameters in the model, to reduce the maximum likelihood by including additional parameters (Bogdan et al., 2004), in order to determine the optimal mixture model. The BIC implies an asymptotic outcome, based on the fact that the distribution of the data are assumed to belong to the exponential family. The BIC is described as:

$$BIC = 2\log(L) - m\log(n)$$

where $L$ denotes the value of the maximum likelihood function of the estimated model, $m$ the number of free estimated parameters, and $n$ the sample size. This formula includes a term that penalises for the inclusion of additional parameters, implying that redundant parameters could reduce the BIC value. The larger the BIC, the more likely the model and number of clusters adequately describes the data.

## 3. RESULTS

### 3.1. Finding the optimal number of clusters

Table 3 summarises the best fitting models based on the BIC criteria, which `mclust` recognised as the best combination of clusters to collectively capture the features of the assessed 15 variables that relate to mathematics performance of Qatar students. The best selected model was the VVV model, with seven component clusters, based on the smallest BIC value of −95,669 across all compared models. The VVV model indicates that the geometric features of the ellipsoidal distribution for k clusters could have a different volume ($\lambda_K$), shape, and orientation ($D_k$). The second best model is the VVV model with five component clusters since the associated BIC is the second smallest value (BIC= −95,852). The difference between the VVV model with seven components and the model with five components was not practically significant as the BIC was only slightly different. However, the VVV model with seven clusters is recommended by `mclust` as providing the optimal clustering of the data.

**Table 3.** BIC updated for all Models

| # | EII | VII | EEI | VEI | EVI | VVI | EEE | EEV | VEV | VVV |
|---|---|---|---|---|---|---|---|---|---|---|
| 2 | *−123,51* | *−123,4* | *−122,331* | −122,162 | −121,528 | −118,115 | −114,543 | −112,339 | −111,980 | −101,827 |
| 3 | −121,78 | −121,057 | −119,314 | −119,100 | −114,349 | −112,932 | −113,976 | −111,625 | −110,958 | −97,840 |
| 4 | −120,24 | −119,485 | −118,180 | −117,682 | −111,416 | −108,245 | −112,897 | −110,509 | −109,752 | −96,407 |
| 5 | −118,82 | −118,052 | −117,366 | −116,566 | −108,783 | −104,882 | −112,596 | −109,563 | −108,663 | −95,852 |
| 6 | −118,18 | −117,190 | −116,559 | −115,694 | −106,984 | −102,952 | −111,783 | −108,799 | −108,054 | −96,224 |
| 7 | −117,67 | −116,586 | −115,910 | −114,931 | −104,863 | −101,829 | −111,768 | −108,315 | −107,593 | **−95,669** |
| 8 | −117,17 | −116,134 | −115,264 | −114,229 | −104,211 | −101,167 | −111,746 | −108,363 | −107,400 | −96,269 |
| 9 | −116,66 | −115,704 | −114,825 | −113,817 | −103,353 | −100,386 | −111,682 | −108,166 | −107,425 | −97,117 |
| | −116,27 | −115,276 | −114,397 | −113,558 | −102,896 | −99,684 | −111,142 | −108,244 | −107,333 | −97,436 |

### 3.2. Cluster-Based analysis of Individual Parameters

Table 4 presents mathematics performance as an evaluation field in which the variables were used to estimate the mean for each of the seven clusters. The colours red, yellow and green are chosen to represent the effectiveness of teaching strategies as low, moderate and high, respectively. For example, in cluster 7 (C VII), ‘attributions to failure’ and ‘disciplinary climate’ had variable averages of 1.902 and 1.492 respectively, out of 4. This meant that, although cluster 7 included most of the variables with high implementation of most teaching strategies and, on average, with students with moderately educated parents, there was also a low ‘disciplinary climate’, low ‘attribution to failure’, as well as low ‘vignette teacher support’, and notably and socio economic status, which was measured by family wealth, all of which influenced student mathematics performance. Students in this cluster recorded the lowest mathematics performance, compared with students in other clusters. With the maximum value of 4 out of 4, cluster 5 (C V) recorded the highest values for ‘mathematics teaching’ and ‘teacher support’. However, the evaluation field (mathematics performance) for this cluster was not as high

as may be expected, as this cluster also comprised of students with parents with low parental education levels. This could mean that notable teaching strategies for students who, on average, come from backgrounds of lower socio-economic status, does not necessarily lead to notable improvements in mathematics performance as these can be outweighed by socio-economic backgrounds of students. This could be due to these students having, on average, parents with lower education levels, parents who could potentially be less aware of the value of mathematics to future career paths and may also be less able to assist their students in these subject areas. There could hence be a lower emphasis on learning and hence on performing well in mathematics among such students based on their home environments.

**Table 4.** Estimated Means in each cluster and each evaluation field

| | C I (599) | C II (100) | C III (359) | C IV (679) | C V (210) | C VI (1,034) | C VII (61) |
|---|---|---|---|---|---|---|---|
| Variables | 20.2% | 3.2% | 11.4% | 22.4% | 6.8% | 33.9% | 2.0% |
| Attributions to failure | 2.601 | 3.153 | 2.578 | 2.438 | 2.796 | 2.522 | 1.902 |
| Mathematics teaching | 3.215 | 3.052 | 3.229 | 3.124 | 4.000 | 3.186 | 3.882 |
| Teacher-directed instructions | 3.076 | 2.667 | 3.058 | 2.993 | 3.912 | 3.045 | 3.692 |
| Student orientation | 2.610 | 2.654 | 2.408 | 2.494 | 3.576 | 2.449 | 3.500 |
| Formative assessment | 2.840 | 2.714 | 2.797 | 2.690 | 3.818 | 2.774 | 3.451 |
| Cognitive activation | 2.929 | 2.365 | 2.920 | 2.900 | 3.648 | 2.943 | 3.697 |
| Disciplinary climate | 2.578 | 3.080 | 2.712 | 2.663 | 2.998 | 2.664 | 1.492 |
| Vignette teacher support | 2.697 | 2.743 | 2.826 | 2.757 | 3.039 | 2.848 | 2.447 |
| Teacher support | 3.214 | 2.433 | 3.226 | 3.186 | 4.000 | 3.233 | 3.729 |
| Vignette classroom management | 2.973 | 2.614 | 3.020 | 3.033 | 3.084 | 3.032 | 3.144 |
| Classroom management | 3.028 | 2.301 | 2.958 | 2.970 | 3.330 | 2.969 | 3.348 |
| Student–teacher relations | 2.959 | 2.788 | 2.985 | 2.962 | 3.474 | 2.993 | 3.468 |
| Socio-economic status | 1.050 | 1.214 | 1.157 | 1.400 | 1.305 | 1.194 | 1.034 |
| Mother qualification | 3.173 | 4.279 | 4.139 | 3.463 | 3.404 | 6.000 | 4.721 |
| Father qualification | 3.326 | 4.314 | 6.000 | 3.696 | 3.927 | 6.000 | 5.033 |
| Mathematics Performance | 0.210 | 0.156 | 0.288 | 0.206 | 0.187 | 0.311 | 0.148 |

## 4. DISCUSSION

The uptake of most mathematics teaching strategies was low to moderate in most clusters. Cluster 6 was associated with the highest mathematics performance scores but lower utilisation of mathematics teaching strategies. This could mean that students with parents with higher education levels could perform notably better in mathematics, irrespective of how mathematics content is presented in class.

This study required a conceptual framework to capture a diverse range of variables that could potentially relate performance in mathematics to multiple variables to describe the interactions between various aspects of the learning environment in terms of learning mathematics. To achieve this, Bayesian regularisation was used in an attempt to reduce convergence issues which are typical with other clustering methods. This paper has demonstrated how model-based clustering using `mclust` can be used to concurrently analyse different models, to identify the most suitable model by selecting the best fitting clustering model using the BIC criterion and estimating parameters via maximum likelihood estimation.

From a statistical point of view, this paper has explained the importance of the volume, shape and orientation of clusters in the model-based clustering approach, and the way these interact to influence the distribution (ellipsoidal, diagonal or spherical) of each cluster. The EM algorithm was used to estimate the values of cluster memberships and a prior was used to control model-based clustering estimation to enhance results. These models were investigated in an educational context to identify factors that influence mathematics performance in different clusters of students by adaptation of the best fitted model. Furthermore, these results can help teachers to better identify groups of students whose performance in various subject areas is more likely to be affected by certain factors—thereby helping them to consider avenues to explore to improve student learning.

Future research that stems from this work can include investigation of the dimension reduction function in the mclust package to visualise results of clustering results in dimensions lower than that of the original dataset.

## REFERENCES


Alzahrani, A.R. and Stojanovski, E. (2019). Evaluation of mathematics teaching strategies in Australian high schools. In Elsawah, S. (ed.) MODSIM2019, 23rd International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2019, pp. 905–910. ISBN: 978-0-9758400-9-2. **https://doi.org/10.36334/modsim.2019.J9.alzahrani**

Banfield, J. and Raftery, A. E. (1993). Model-based Gaussian and non Gaussian clustering. Biometrics, 49, 803–821.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated classification likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(7), 719–725.

Bogdan, M., Ghosh, J. K., and Doerge, R. W. (2004). Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. Genetics, 167(2), 889–999.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. Pattern Recognition, 28, 781–793.

Erar, B. (2011). Mixture model cluster analysis under different covariance structures using information complexity. Master's Thesis, University of Tennessee. **https://trace.tennessee.edu/utk_gradthes/968**

Fraley, C., and Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. Journal of the American Statistical Association, 97, 611–631.

Fraley, C., and Raftery, A. E. (2007). Bayesian regularisation for normal mixture estimation and model-based clustering. Journal of Classification, 24, 155–181.

Fraley, C., Raftery, A. E., Murphy, T. B., and Scrucca, L. (2012). Mclust version 4 for R: Normal mixture modelling for model-based clustering, classification, and density estimation. Technical Report No.597. Accessed from, https://www.researchgate.net/profile/Thomas_Murphy7/publication/257428214_MCLUST_Version_4_for_R_Normal_Mixture_Modeling_for_Model-Based_Clustering_Classification_and_Density_Estimation/links/00b7d53c4d92c86041000000/MCLUST-Version-4-for-R-Normal-Mixture-Modeling-for-Model-Based-Clustering-Classification-and-Density-Estimation.pdf?origin=publication_detail

Fraley, C., Raftery, A. E. and Scrucca, L. (2020). mclust: Gaussian Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation. R package version 5.4.6. https://CRAN.R-project.org/package=mclust

Gordon A. D. (1987). A review of hierarchical classification. Journal of the Royal Statistical Society A, 150, 119–137.

Grün, B. (2018). Model-based clustering. In G. Celeux, S. Frühwirth-Schnatter, & C. P. Robert (Eds.), Handbook of Mixture Analysis (pp. 1–38). Chapman & Hall/CRC Press.

Hartigan, J. A. and Wong, M. A. (1979). [Algorithm AS 136] A k-means clustering algorithm, Applied Statistics, 28, 100–108.

Lloyd, S. (1957). Least squares quantization in PCM., Technical report, Bell Laboratories. Published in 1982 in IEEE Transactions on Information Theory, 28, 128-137.

Lomet. A., Govaert. G., and Grandvalet. Y. (2012) Integrated classification likelihood for model selection in block clustering. *Workshop statistical inference in complex/high-dimensional problems*, Jul 2012, Vienne, Austria. pp. 1-16. ⟨hal-00933256⟩

Melnykov, V. and Maitra, R. (2010). Finite mixture models and model-based clustering. *Statistics Publications*. 67.

Niraj, K., Lade, S., and Malviya, N. (2013). Clustering of datasets by using centroid based method. International Journal of Emerging Technology and Advanced Engineering, 3, 614–620.

OECD. (2013). PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy. OECD Publishing

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Scrucca, L. (2010). Dimension reduction for model-based clustering. Statistics and Computing, 20, 471-484.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. The R Journal, 8(1), 289–317.

Xue, J., Luo, Y., & Liang, F. (2017). Average (E)BIC-like criteria for Bayesian model selection[unpublished manuscript]. University of Florida. https://people.clas.ufl.edu/yeluo/files/ave4.pdf