7th International Conference on
Combinatorics, Cryptography, Computer Science and Computing
November. 16-17, 2022

Model-Based Clustering Performance Evaluation Using Some External Scores Measures

Rajaa Hasan Abbas[1]
University of Kufa, College of Education for Girls, Department of Mathematics, Najaf, Iraq

Samira Faisal Abushilah
University of Kufa, College of Education for Girls, Department of Mathematics, Najaf, Iraq

Abstract

An accurate statistical notion of clusters can be obtained by a density-based approach owing to the fact that a distribution is more informative than classical representations. The density-based clustering method is growing in popularity more recently. However, we seek to measure how well the partitioning is estimated by the density-based approach if we know the *true clusters* of the data. Therefore, in this paper some external scores measures are picked (Rand, F-M, Purity) to compare the success of the partition found by density-based algorithm under different models.

Keywords: Model-Based clustering, Rand index, F-M index, Purity index.

1 Introduction

More recently, the model-based clustering approach is growing in popularity for many reasons. First, the clusters which are formed are easy to understand and do not limit the shape of the clusters [2]. Second, no prior number of clusters is required. Third, this approach can deal with spatial databases, such as tomography and satellite images, and on large databases it has a good efficiency [5].

Model-based clustering has been used in a broad range of contexts including gene expression data [14], some applications in chemistry [6], image analysis [7], food science [11], social sciences [1], geochemistry [4].

In this paper some external scores measures are picked (Rand, F-M, Purity) to examine the performance of the model-based clustering to measure how well the partitioning is estimated by the density-based approach if we know the *true clusters* of the data.

2 Model-Based clustering(Mclust)

In 2002, Fraley and Raftery presented mixture models in hierarchical clustering [8]. The authors assumed that the best partitioning can be calculated by maximising the classification likelihood, which is given by

$$L_{CL}(\theta_1, \theta_2, ..., \theta_k; D_1, ..., D_m) = \prod_{i=1}^{m} f_{D_i}(x_i|\theta_i),$$

where

[1]speaker

- x_i is the data points (i$\in \{1, ..., m\}$).
- k is the number of clusters.
- D_i represents the clusters of x_i.
- θ_i are the parameters for the cluster D_i.
- f_{D_i} is the probability density function for the cluster D_i.

Fraley and Raftery in 2002 assumed that the density function f_{D_i} comes from multivariate Gaussian distribution with the following density function,

$$g(x_i; \mu_i, \Sigma_i) = \frac{exp(\frac{-1}{2}(x_i - \mu_i)^T \sum_i^{-1}(x_i - \mu_i))}{\sqrt{det(2\pi\Sigma_i)}},$$

where μ_i and Σ_i represent the mean and variance of the cluster D_i where these parameters can be estimated using EM algorithm with a fixed number of clusters. In order to determine which pairs we have to merge in each step in the Mclust method we have to look at the amount of increase in classification likelihood for each possible clusters by using the Bayesian information criteria (BIC). Model-based hierarchical clustering is available in the mclust package in R [9].

3 External scores

In order to evaluate the performance of the Mclust method, which have been illustrated above, we pick the following external scores: Rand index [12], FM index, and Purity index [13] [3].

- **Rand Index**

Let $L^{(1)} = \{D_1^{(1)}, \cdots, D_u^{(1)}\}$ and $L^{(2)} = \{D_1^{(2)}, \cdots, D_v^{(2)}\}$ be two partitions (see Table 1). The Rand measure, which is suggested by [12] is a measure of the agreement between two partitions, and this index is defined by

$$RI = \left[\binom{m}{2} + \sum_{i,j} n_{ij}^2 - \frac{1}{2}\left(\sum_{i=1}^{u} n_{i\cdot}^2 + \sum_{j=1}^{v} n_{\cdot j}^2\right)\right] \Big/ \binom{m}{2},$$

where n_{ij} represents the number of observations allocated in the cluster $D_i^{(1)}$ in $\mathrm{L}^{(1)}$ and to cluster $D_j^{(2)}$ in $L^{(2)}$. As we can see $RI \in [0, 1]$, the value 0 indicates that the two data clusterings do not agree on any pair of points, while the value 1 indicates that the two partitions are exactly the same.

- **Fowlkes and Mallows Index (FMI)**

Fowlkes and Mallows [10] introduced an external score to check the similarity between two partitions of a data points, and this index is given by

$$FMI = \frac{T_k}{\sqrt{P_k Q_k}}, FMI \in [0, 1]$$

where

$$T_k = \sum_{i=1}^{u}\sum_{j=1}^{v} n_{ij}^2 - m,\ P_k = \sum_{i=1}^{u}\left(\sum_{j=1}^{v} n_{ij}\right)^2 - m,\ Q_k = \sum_{j=1}^{v}\left(\sum_{i=1}^{u} n_{ij}\right)^2 - m$$

Table 1: The general form of contengency table between two partitions $L^{(1)} = \{D_1^{(1)}, \cdots, D_u^{(1)}\}$ and $L^{(2)} = \{D_1^{(2)}, \cdots, D_v^{(2)}\}$.

		$L^{(2)}$				sums
		$D_1^{(2)}$	$D_2^{(2)}$	$\cdots$	$D_v^{(2)}$	
$L^{(1)}$	$D_1^{(1)}$	n_{11}	n_{12}	$\cdots$	n_{1v}	$n_{1.}$
	$D_2^{(1)}$	n_{21}	n_{22}	$\cdots$	n_{2v}	$n_{2.}$
	.	.	.	.	.	
	.	.	.	.	.	
	.	.	.	.	.	
	$D_u^{(1)}$	n_{u1}	n_{u2}	$\cdots$	n_{uv}	$n_{u.}$
Sums		$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.v}$	

- **Purity Index**

The purity measure which is introduced by [13] to measure the agreement between two partitions $L^{(1)}$ and $L^{(2)}$ takes the average purity for each cluster D_i from the same partition, D_j, and the maximum number of elements clustered together will be defined as purity. Hence, the purity measure is defined by

$$\text{UI} = \frac{1}{m}\sum_{i=1}^{u} max_j |D_i^{(1)} \cap D_j^{(2)}|, \tag{1}$$

where UI $\in [-1, 1]$, if UI is close to one, this means the similarity between the clustering and the true clusters is high.

4 Evaluation

In this section, we evaluate the performance of model-based method under different external validity indices (Rnad, FMI, Putity). The observed (multivariate) data is assumed to have been generated from a finite mixture of component models. Where simulated data is used with different models, different sample sizes, using the R 3.1 software.

The purpose of the suggested algorithm is which external validity measure is able to get the high agreement between the results of model-based method and the true clusters. For this purpose, we generate groups of data from various models with different parameters and we measure the performance of clustering algorithms with the data that we have generated. We consider the following models:

1. Model 1: 5 clusters of size 50 observations are generated from $N_2\left(\mu_i, \begin{bmatrix} 1.25 & 0 \\ 0 & 1.25 \end{bmatrix}\right)$, where $\mu_i \in \left\{ \begin{bmatrix} 0.25 \\ 0.25 \end{bmatrix}, \begin{bmatrix} 4.5 \\ 4.5 \end{bmatrix}, \begin{bmatrix} 7 \\ 7 \end{bmatrix}, \begin{bmatrix} 11 \\ 11 \end{bmatrix}, \begin{bmatrix} 15 \\ 15 \end{bmatrix} \right\}, i = 1, ..., 5.$
2. Model 2: 8 clusters of size 60 observations are generated from $N_2(\mu_i, \Sigma_i)$, where

$$\mu_i \in \left\{ \begin{bmatrix} 8 \\ 8 \end{bmatrix}, \begin{bmatrix} 3 \\ 8 \end{bmatrix}, \begin{bmatrix} 8 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 5 \end{bmatrix}, \begin{bmatrix} 5 \\ 2 \end{bmatrix}, \begin{bmatrix} 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 8 \\ 5 \end{bmatrix} \right\}, i = 1, ..., 8$$

and

$$\Sigma_i \in \left\{ \begin{bmatrix} 0.75 & 0 \\ 0 & 0.75 \end{bmatrix}, \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}, \begin{bmatrix} 0.8 & 0 \\ 0 & 0.8 \end{bmatrix}, \begin{bmatrix} 0.4 & 0 \\ 0 & 0.4 \end{bmatrix}, \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}, \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}, \begin{bmatrix} 0.9 & 0 \\ 0 & 0.9 \end{bmatrix}, \begin{bmatrix} 0.25 & 0 \\ 0 & 0.25 \end{bmatrix} \right\}.$$

3. Model 3: 9 clusters are generated from $N_2(\mu_i, \Sigma_i)$, where

$$\mu_i \in \left\{ \begin{bmatrix} -2 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} -3 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ 4 \end{bmatrix}, \begin{bmatrix} 4 \\ -2 \end{bmatrix}, \begin{bmatrix} 0 \\ -3 \end{bmatrix}, \begin{bmatrix} -3 \\ 3 \end{bmatrix} \right\}, i = 1, ..., 9$$

and

$$\Sigma_i \in \left\{ \begin{bmatrix} 0.25 & 0 \\ 0 & 0.25 \end{bmatrix}, \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}, \begin{bmatrix} 0.7 & 0 \\ 0 & 0.7 \end{bmatrix}, \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix}, \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}, \begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix}, \begin{bmatrix} 0.4 & 0 \\ 0 & 0.4 \end{bmatrix}, \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix} \right\}.$$

4. Model 4: 10 clusters are generated from $N_2\left(\mu_i, \begin{bmatrix} 0.75 & 0 \\ 0 & 0.75 \end{bmatrix}\right)$ with sizes 40 for clusters $\{G_i, i = 1, ..., 5\}$ and 55 for clusters $\{G_i, i = 6, ..., 10\}$, where

$$\mu_i \in \left\{ \begin{bmatrix} 2 \\ 12 \end{bmatrix}, \begin{bmatrix} 12 \\ 8 \end{bmatrix}, \begin{bmatrix} 6 \\ 6 \end{bmatrix}, \begin{bmatrix} 13 \\ 3 \end{bmatrix}, \begin{bmatrix} 4 \\ 3 \end{bmatrix}, \begin{bmatrix} 7 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 8 \end{bmatrix}, \begin{bmatrix} 12 \\ 13 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 7 \\ 11 \end{bmatrix} \right\}, i = 1, 2, ..., 10.$$

We evaluate the performance of the Mclust method and some validity indices using the following algorithm.

Algorithm 1

1. From different models $N_2(\mu_i, \Sigma_i)$ as described above generate data points $\{(x_1, y_1), ..., (x_m, y_m)\}$, the data sets consists of k true clusters.
2. Use model-based clustering method to build the tree for the data sets.
3. Get the optimal number of clusters, q, using density-based clustering method.
4. Get the agreement between the true clusters and the partitions that we have computer from step (3) using external validity indices (Rand, FMI, Purity).

The results of Algorithm 1 are presented in the next section.

5 Results

The results of the simulation are presented in Table 2 and Figures 1 to 4.

Table 2 illustrates the rate of agreement between the true clusters and the results of the model-based clustering method using some external indices (Rand, FMI, Purity). While, Figures 1 to 4 (Top panels) show the true clusters for the models (Model 1, five groups), (Model 2, eight groups), (Model 3, nine groups), (Model 4, ten groups). Figures 1 to 4 (Middle panels and bottom panels) show BIC, classification, uncertainty, and density, respectively. In these results we can see the following:

- In Model 1 and Model 2 the Mclust method allows specification of the models and numbers of clusters, we note that model 1 has five clusters and model 3 has eight clusters by using the Model-Based method,which is equal to the number of true clusters (5 and 8 clusters).
- In model 3 and model 4 we can see that the Mclust method allows specification of the models and numbers of clusters, we note the model 2 has eight clusters and model 4 has nine clusters by using the Model-Based method,which differs from the true number of clusters adult 9 and 10 clusters.
- The percentage of similarity between the true clusters and the clusters resulting from the Model-based method ranges between $88\% - 98\%$.

Table 2: The rate of agreement between the true clusters and the partition that we have computed using Mclust method for four models using external indicies (Rand, FM, Purity).

Rate of agreement			k, q	Number of observations for each cluster	Log	
Rand	F-M	Purity		Clusters $[n_i]$	Likelihood	BIC
Model 1						
0.957	0.890	0.940	5, 5	50, 49, 49, 50, 50	-1201.8	-2486.4
Model 2						
0.977	0.909	0.955	8, 8	92, 56, 80, 58, 56, 78, 60, 80	-2257.5	-4711.1
Model 3						
0.976	0.911	0.921	9, 8	43, 73, 55, 44, 36, 36, 60, 98	-1684.3	-3557.7
Model 4						
0.973	0.879	0.905	10, 9	57, 38, 40, 45, 55, 40, 71, 74, 55	-2279.2	-4731.1

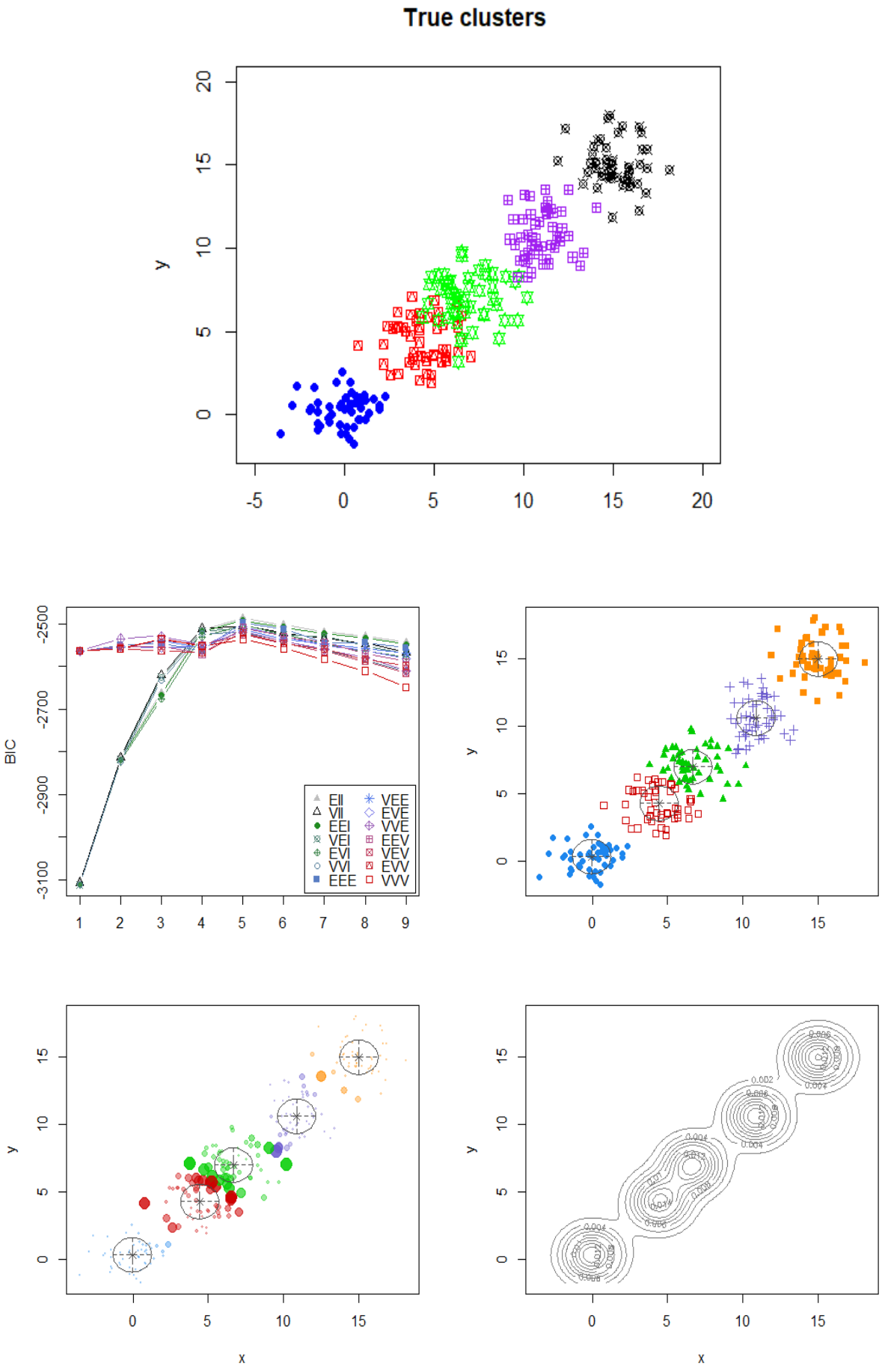

Figure 1: Top panel: true clusters for Model 1 (five groups). Middle panels: BIC, classification. Bottom panels: Uncertainty and density.

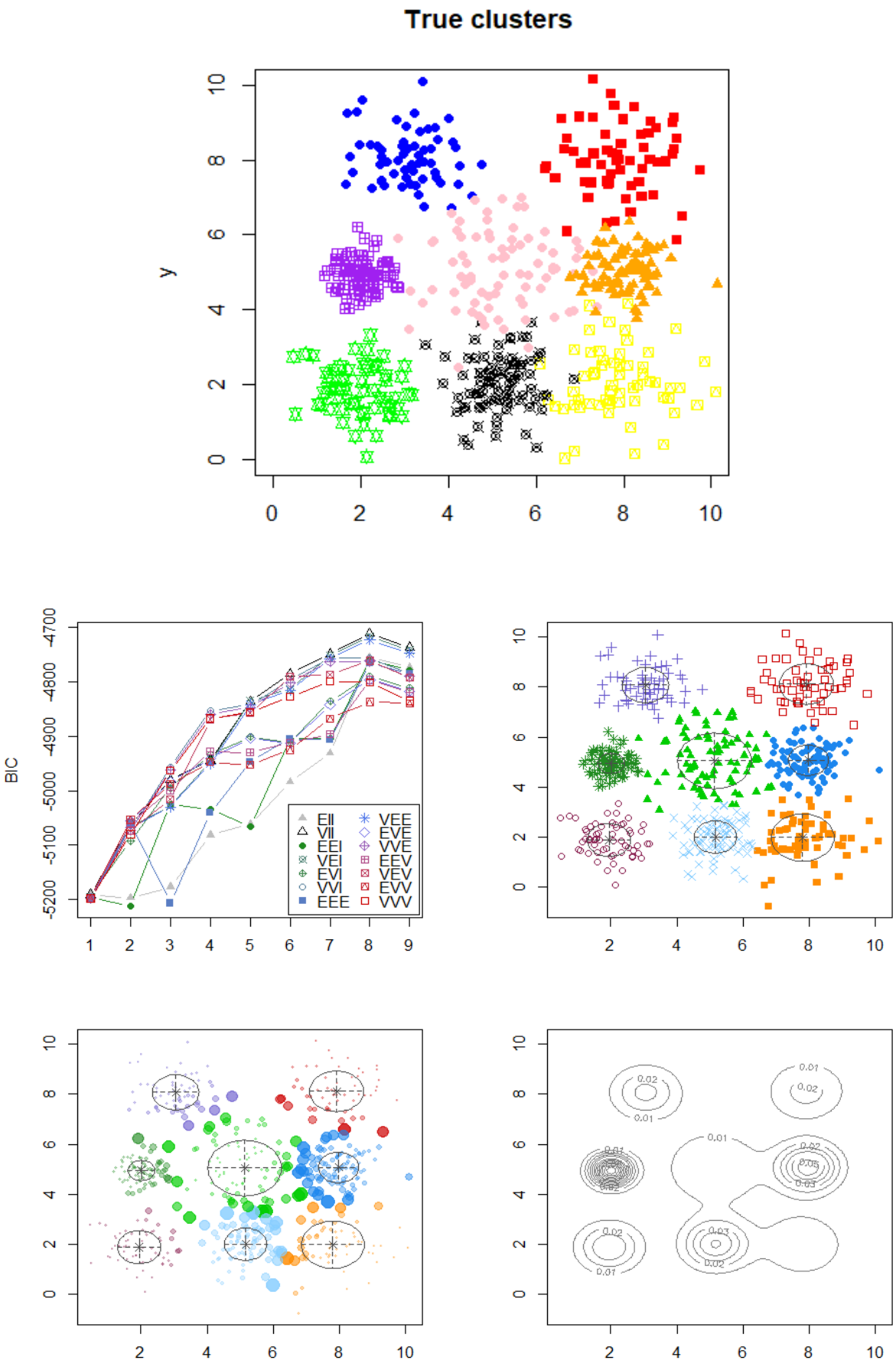

Figure 2: Top panel: true clusters for Model 2 (eight groups). Middle panels: BIC, classification. Bottom panels: uncertainty and density.

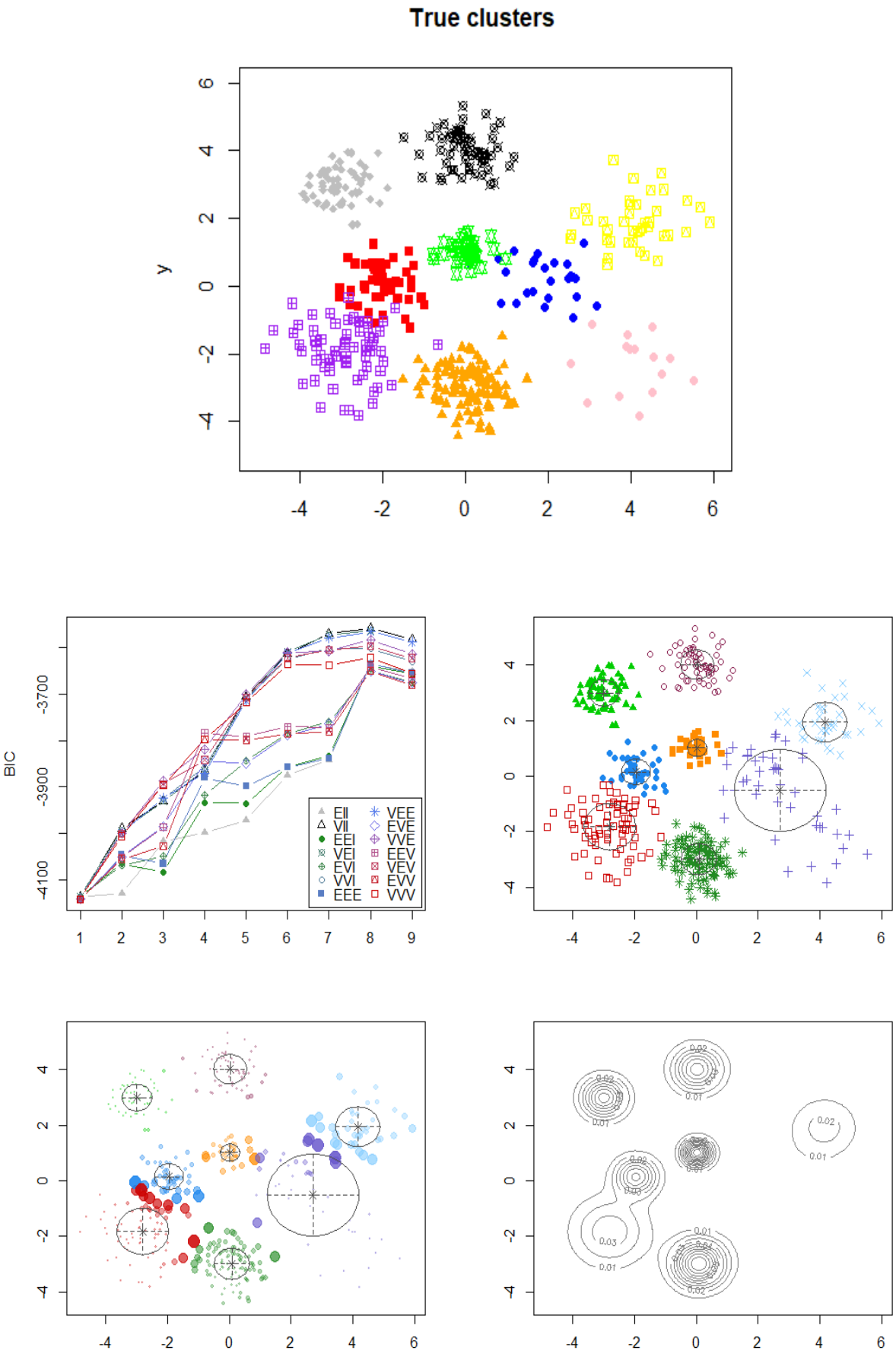

Figure 3: Top panel: true clusters for Model 3 (nine groups). Middle panels: BIC, classification. Bottom panels: uncertainty and density.

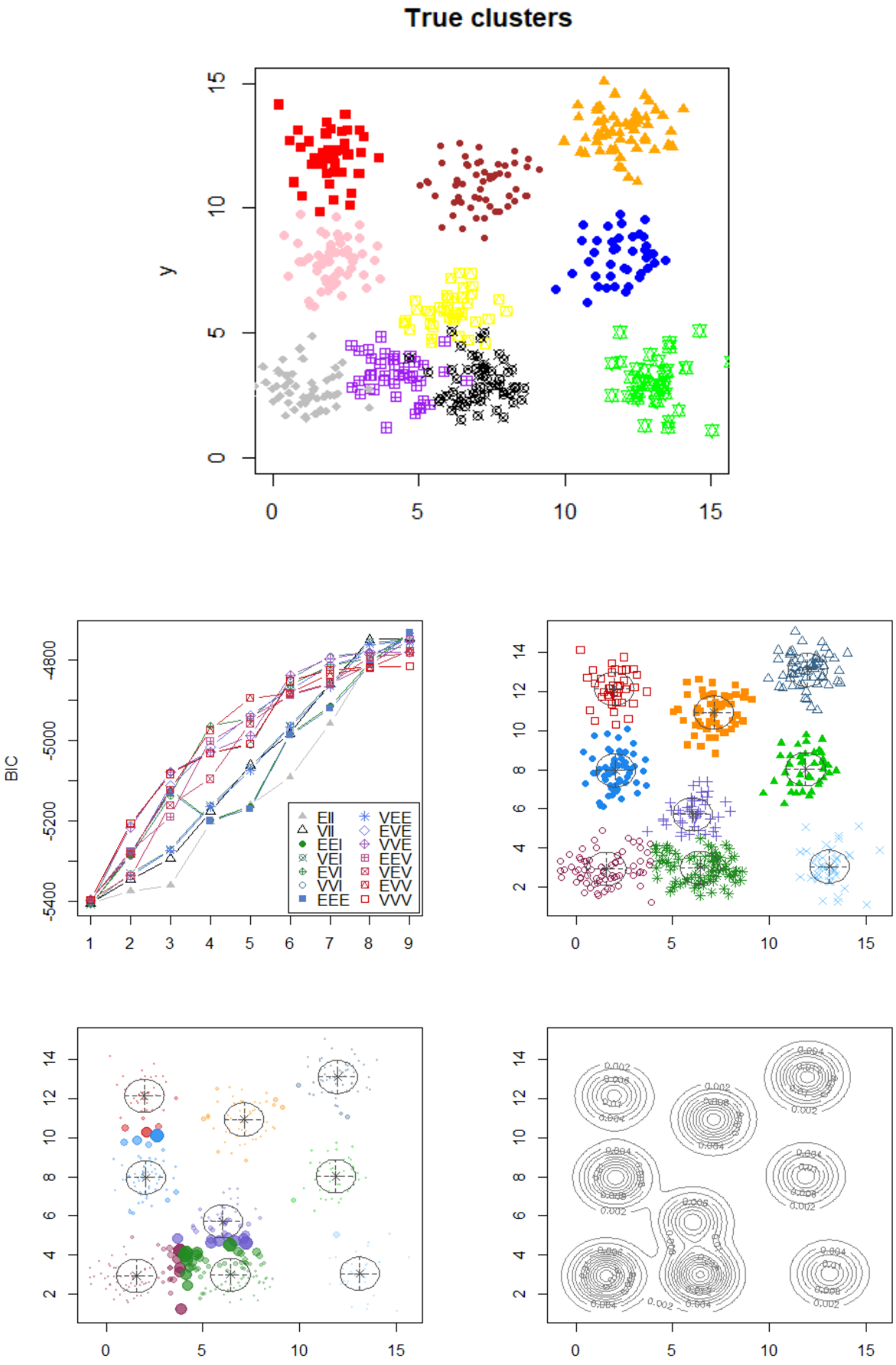

Figure 4: Top panel: true clusters for Model 4 (ten groups). Middle panels: BIC, classification. Bottom panels: uncertainty and density.

6 Conclusion

Different models have been used in this paper to evaluate the performance of model-based clustering method under different simulated models with different sample sizes by using R software 3.1. Also, some external indices (Rand, F-M, Purity) are used to get the rate of agreement between the true clusters of the data points and the partitions that we have computed using model-based clustering method. The results of the simulation indicate that under different models the results of clustering using model-based clustering method match the true clusters about more that 88%.

References

[1] Ahlquist, J. S., Breunig, C., 2012. *Model-based clustering and typologies in the social sciences*. Political Analysis 20 (1), 92-112.

[2] Azzalini, A., Torelli, N., 2007. *Clustering via nonparametric density estimation*. Statistics and Computing 17 (1), 71-80.

[3] Chalise, P., Raghavan, R., Fridley, B., 2016. *Intnmf: Integrative clustering of multiple genomic dataset*.

[4] Ellefsen, K. J., Smith, D. B., Horton, J. D., 2014. *A modified procedure for mixture-model clustering of regional geochemical data*. Applied Geochemistry 51, 315-326.

[5] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al., 1996. *A density-based algorithm for discovering clusters in large spatial databases with noise*. In: kdd. Vol. 96. pp. 226-231.

[6] Fraley, C., Raftery, A. E., 2006b. *Some applications of model-based clustering in chemistry*. R News 6 (3), 17-23.

[7] Fraley, C., Raftery, A. E., Scrucca, L., Murphy, T. B., Fop, M., Scrucca, M. L., 2012. *Package 'mclust'*

[8] Fraley, C., Raftery, A. E., 2002. *Model-based clustering, discriminant analysis, and density estimation*. Journal of the American statistical Association 97 (458), 611-631.

[9] Fraley, Raftery, Scrucca, Murphy, Fop, and Scrucca] Fraley, C., Raftery, A. E., Scrucca, L., Murphy, T. B., Fop, M., Scrucca, M. L., 2012. *Package 'mclust'*.

[10] Fowlkes, E. B., Mallows, C. L., 1983. *A method for comparing two hierarchical clusterings*. Journal of the American statistical association 78 (383), 553-569.

[11] Kozak, M., Scaman, C. H., 2008. *Unsupervised classification methods in food sciences: discussion and outlook*. Journal of the Science of Food and Agriculture 88 (7), 1115-1127.

[12] Rand, W. M., 1971. *Objective criteria for the evaluation of clustering methods*. Journal of the American Statistical association 66 (336), 846-850.

[13] Rendon, E., Abundez, I., Arizmendi, A., Quiroz, E. M., 2011. *Internal versus external cluster validation indexes*. International Journal of com- puters and communications 5 (1), 27-34.

[14] Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., Ruzzo, W. L., 2001. *Model-based clustering and data transformations for gene expression data*. Bioinformatics 17 (10), 977-987.

e-mail: rajaah.alabidy@uokufa.edu.iq
e-mail: sameerah.hathoot@uokufa.edu.iq