Croatian Operational Research Review
CRORR **10**(2019), 1-12

Investigation of the optimal number of clusters by the adaptive EM algorithm

Vedran Novoselac *

Mechanical Engineering Faculty in Slavonski Brod, J. J. Strossmayer University of Osijek
Trg Ivane Brlić Mažuranić 2, 35000 Slavonski Brod, Croatia
E-mail: ⟨Vedran.Novoselac@sfsb.hr⟩

Abstract. This paper considers the investigation of the optimal number of clusters for datasets that are modeled as the Gaussian mixture. For that purpose, the adaptive method that is based on a modified Expectation Maximization (EM) algorithm is developed. The modification is conducted within the hidden variable of the standard EM algorithm. Assuming that data are multivariate normally distributed, where each component of the Gaussian mixture corresponds to one cluster, the modification is provided by utilizing the fact that the Mahalanobis distance of samples follows a Chi-square distribution. Besides, the quantity measure is constructed in order to determine number of clusters. The proposed method is presented in several numerical examples.

Keywords: chi-square, clustering, EM algorithm, Gaussian mixture, Mahalanobis distance

Received: September 29, 2018; accepted: October 31, 2018; available online: July 4, 2019

DOI: 10.17535/crorr.2019.0001

1. Introduction

Cluster analysis is one of the most important tasks in data mining which has great application in several fields like image analysis, pattern recognition and statistics. In general, clustering is an unsupervised learning process, the task of which is to classify data points into groups or clusters. A wide variety of clustering algorithms are proposed for different applications where the problem of determining the number of clusters plays an important research problem. In most cases, the optimal number of clusters is an unknown parameter and thereby presents a challenging problem in cluster analysis, called the Cluster Validity (CV) problem [3, 10].

Clustering validation is a way of a providing the quality of different clustering algorithms, as well as a way of a comparing the two clustering results with different number of clusters. In that sense, cluster validity can be used for determining the optimal number of clusters in a dataset. In general, there exist three cluster validity criteria for evaluating the results of the clustering algorithms: external criteria, internal criteria and relative criteria. Both internal and external criteria are based on statistical testing and they have high computation demand, which is their main disadvantage. The relative criteria does not involve statistical testing. The basis of the relative criteria is a reference on obtained background knowledge about clustering results. The aim of the relative criteria is to choose the best clustering schema from the different results based on the Cluster Validity Indices (CVI). In literature there exist many different relative indices that can give a quantitative criteria for evaluating the quality of clustering results [3, 10, 11].

In this paper, the datasets that are modeled by the Gaussian mixture model are observed, and thus the new algorithm is developed based on modification of the well known Expectation

*Corresponding author.

Maximization (EM) algorithm [1]. The modification is conducted by the well known Mahalanobis distance, which has great application in data analysis [5, 6, 7, 9], within the hidden variable of the standard EM algorithm. The modification is conducted in sense of the outlier detection [8], which uses the fact that the squared Mahalanobis distance follows a Chi-square distribution for the data that are normally distributed [2, 4]. In that way, the adaptive EM algorithm that has flexible rejection procedure of data is developed. Furthermore, in that sense of the cluster validity indices, the new quantity measure is constructed. The proposed index is based on the percentage of data included in final clusters obtained by the adaptive EM. It is shown that the problem of determining the optimal number of clusters can be solved by finding a knee point of the proposed index [13].

The paper is organized as follows. In Section 2 the EM algorithm and its implementation for the Gaussian mixture model are introduced. In Section 3 the adaptive EM algorithm and its pseudo–code are presented. In Section 4 several illustrative examples are presented, and finally, in Section 5 the conclusion is given.

2. The EM for the Gaussian mixture model

Let us denote by $X = \{x_i \colon i \in I\} \subseteq \mathbf{R}^n$, $I = \{1, \ldots, m\}$, the dataset, and by $\pi_j \subseteq X$ clusters, where $J = \{1 \ldots, k\}$, $k \leq m$. Because the data are modeled as the Gaussian mixture, the well known EM algorithm is used [1, 10]. The Gaussian mixture models the data as a mixture of the multiple Gaussian distributions

$$P(x|\Theta) = \sum_{j \in J} w_j P(x|\theta_j), \quad \Theta = \{(w_j, \mu_j, \Sigma_j) \colon j \in J\}, \tag{1}$$

where w_j represents the a priori probability of belonging to a corresponding cluster π_j, and therefore $\sum_{j \in J} w_j = 1$. The parameter $\theta_j = (\mu_j, \Sigma_j)$ presents expectation $\mu_j \in \mathbb{R}^n$ and covariance matrix $\Sigma_j \in \mathbb{R}^{n \times n}$ of density function for the multivariate Gaussian distribution

$$P(x|\theta_j) = \frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} e^{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)} \tag{2}$$

The EM algorithm is an iterative procedure which computes the maximum likelihood estimate (MLE) of log-likelihood function

$$\ln \mathcal{L}(\Theta|X) = \ln P(X|\Theta) \tag{3}$$

In order to facilitate the ML estimation parameter Θ, the hidden variable Z is introduced. Then, instead of solving the log-likelihood of the observed data X, the log-likelihood function of the complete data (X, Z) is observed, i.e.

$$\ln \mathcal{L}(\Theta|X, Z) = \ln P(X, Z|\Theta) \tag{4}$$

The hidden variable Z is constructed to determine from which component observation originates. In that case Z is defined in a way that the entries $z_{ij} \in \{0, 1\}$ of the measurement vector $z_i \in \mathbb{R}^k$ are equal to one if and only if cluster π_j contains observation x_i. In that case $P(z_{ij} = 1) = w_j$ and $\sum_{j \in J} z_{ij} = 1$, which implies

$$P(Z = z_i) = \prod_{j \in J} w_j^{z_{ij}} \tag{5}$$

Now the complete log-likelihood function for the Gaussian mixture model can be carried out as

$$\ln \mathcal{L}(\Theta|X, Z) = \ln P(X, Z|\Theta) = \sum_{i \in I} \sum_{j \in J} z_{ij} (\ln w_j + \ln P(x_i|\theta_j)), \tag{6}$$

where the aim is to estimate parameters $\Theta = \{(w_j, \mu_j, \Sigma_j) \colon j \in J\}$ alternating the E-step and the M-step until convergence [12]. The process is described below:

E-step: The calculation of expectation $\mathcal{Q}(\Theta|\Theta^{(t)})$ is conducted, where $\Theta^{(t)} = \{(w_j^{(t)}, \mu_j^{(t)}, \Sigma_j^{(t)}) \colon j \in J\}$ present a current estimation of parameters:

$$\begin{aligned} \mathcal{Q}(\Theta|\Theta^{(t)}) &= \mathrm{E}_{Z|X,\Theta^{(t)}}[\ln \mathcal{L}(\Theta|X, Z)] \\ &= \sum_{i\in I}\sum_{j\in J} h_{ij}^{(t)}(\ln w_j + \ln P(x_i|\theta_j)) \end{aligned} \tag{7}$$

Parameter $h_{ij}^{(t)}$ presents the posteriori probabilities, i.e. the probability that observation x_i is generated by π_j, defined as follows:

$$h_{ij}^{(t)} = \frac{w_j^{(t)} P(x_i|\theta_j^{(t)})}{\sum_{l\in J} w_l^{(t)} P(x_i|\theta_l^{(t)})} \tag{8}$$

M-step: An optimum of $\mathcal{Q}(\Theta|\Theta^{(t)})$ must be calculated, which can be analytically solved from the equation $\nabla_\Theta \mathcal{Q}(\Theta|\Theta^{(t)}) = 0$. The optimal results can be carried out as:

$$w_j^{(t+1)} = \frac{1}{m}\sum_{i\in I} h_{ij}^{(t)}, \tag{9}$$

$$\mu_j^{(t+1)} = \frac{\sum_{i\in I} h_{ij}^{(t)} x_i}{\sum_{i\in I} h_{ij}^{(t)}}, \tag{10}$$

$$\Sigma_j^{(t+1)} = \frac{\sum_{i\in I} h_{ij}^{(t)} (x_i - \mu_j^{(t+1)})(x_i - \mu_j^{(t+1)})^T}{\sum_{i\in I} h_{ij}^{(t)}} \tag{11}$$

3. The adaptive EM algorithm

In this section the adaptive EM algorithm, which is based on the standard EM, is presented. The new algorithm is created as a modification of the standard EM [6], which is based on the Mahalanobis distance. A modification is conducted on the hidden variable Z, and thereby the condition is introduced which is based on the fact that the squared Mahalanobis distance follows χ_n^2 distribution [2, 4].

The squared Mahalanobis distance is defined as

$$d_M(x;\theta) = (x-\mu)^T \Sigma^{-1}(x-\mu), \quad \theta = (\mu, \Sigma), \tag{12}$$

where $\mu \in \mathbb{R}^n$ presents the arithmetic mean and $\Sigma \in \mathbb{R}^{n\times n}$ the sample covariance matrix of a sample $X \subseteq \mathbb{R}^n$. The squared Mahalanobis distance of samples approximates a Chi-squared distribution χ_n^2 with n degrees of freedom for normally distributed data. This condition is written as an inequality:

$$d_M(x;\theta) \leq \chi_n^2(p), \tag{13}$$

where $\chi_n^2(p)$, $p \in (0,1)$, present p-quantile value of Chi-squared distribution, and therefore reasonable will be $p \approx 0.05$, which prevents influence of those observations which are located

on the tail regions of the Gaussians of low probability. Finally, the proposed rejection can be defined as the Kronecker delta:

$$\delta(x;\Theta) = \begin{cases} 1, \min\limits_{j\in J} d_M(x;\theta_j) \leq \chi^2_n(p); \\ 0, \text{ else} \end{cases} \tag{14}$$

where $d_M(x;\theta_j)$ is generated by component j, i.e. cluster π_j. The modification now can be carried out as the restriction of the standard hidden variable, i.e.

$$\widehat{Z} = Z_{|\widehat{X}}, \tag{15}$$

where $\widehat{X} = \{x_i \colon i \in \widehat{I}\}$, $\widehat{I} = \{i \in I \colon \delta_i = 1\}$, $\delta_i = \delta(x_i;\Theta)$. Now the E-step and the M-step can be performed on $\ln \mathcal{L}(\Theta|\widehat{X}, Z)$, and the results can be easily derived as the optimization problem for the standard hidden variable. By including the initialization step $\Theta^{(t)}$ the optimization results can be carried out as:

$$w_j^{(t+1)} = \frac{1}{\widehat{m}^{(t)}} \sum_{i\in\widehat{I}^{(t)}} h_{ij}^{(t)}, \tag{16}$$

$$\mu_j^{(t+1)} = \frac{\sum\limits_{i\in\widehat{I}^{(t)}} h_{ij}^{(t)} x_i}{\sum\limits_{i\in\widehat{I}^{(t)}} h_{ij}^{(t)}}, \tag{17}$$

$$\Sigma_j^{(t+1)} = \frac{\sum\limits_{i\in\widehat{I}^{(t)}} h_{ij}^{(t)} (x_i - \mu_j^{(t+1)})(x_i - \mu_j^{(t+1)})^T}{\sum\limits_{i\in\widehat{I}^{(t)}} h_{ij}^{(t)}}, \tag{18}$$

where $\widehat{I}^{(t)} = \{i \in I \colon \delta_i^{(t)} = 1\}$, $\delta_i^{(t)} = \delta(x_i;\Theta^{(t)})$ and $\widehat{m}^{(t)} = |\widehat{I}^{(t)}|$. Below in *Algorithm 1* is presented a pseudo-code of the adaptive EM algorithm.

Algorithm 1 The adaptive EM algorithm

1: **Initialization step:** $\Theta = \{(w_j, \mu_j, \Sigma_j) \colon j \in J\}$, $p \in (0,1)$;
2: *loop*:
3: Calculate:

$$\widehat{I} \leftarrow \{i \in I \colon \delta_i = 1\};$$

4: **E-step:** For every $i \in \widehat{I}$ calculate cluster probability for ever $j \in J$:

$$h_{ij} \leftarrow \frac{w_j P(x_i|\theta_j)}{\sum\limits_{l \in J} w_l P(x_i|\theta_l)};$$

5: **M-step:** Calculation of the Gaussian mixture model parameters for every $j \in J$:

$$w_j \leftarrow \frac{1}{\widehat{m}} \sum_{i \in \widehat{I}} h_{ij};$$

$$\mu_j \leftarrow \frac{\sum\limits_{i \in \widehat{I}} h_{ij} x_i}{\sum\limits_{i \in \widehat{I}} h_{ij}};$$

$$\Sigma_j \leftarrow \frac{\sum\limits_{i \in \widehat{I}} h_{ij}(x_i - \mu_j)(x_i - \mu_j)^T}{\sum\limits_{i \in \widehat{I}} h_{ij}};$$

6: *iterate until* Θ *converges*;

In the next section it is shown on various numerical examples that the adaptive EM algorithm has convergence property. Furthermore, it is shown that the optimal number of the clusters can be effectively investigated by observing a percentage of non-rejected data, i.e. $\widehat{m}/m$.

4. Numerical examples

In this section, an experimental research on numerical examples modeled as the Gaussian mixture is considered. The research is conducted on two data types: without the outliers, and with the presence of outliers (i.e., in noisy environment). The noises are generated by the random vector, which is uniformly distributed over the pre-defined region.

The final result of the adaptive EM for a specific k is denoted by $\widehat{m}_k/m$. The result of $\widehat{m}_k/m$ is conducted by running the adaptive EM for many different initial parameters, and choosing among them the result which achieves maximum value of the proposed index (i.e. a percentage of non-rejected data). It is shown that the optimal number of clusters can be indicated by the knee point of $\widehat{m}_k/m$ via k. The knee point presents the point $T_k(k, \widehat{m}_k/m)$, $k = 2, 3, \ldots$, with the largest corresponding angle

$$\alpha_k = \arccos \frac{|v_1 \cdot v_2|}{\|v_1\| \|v_2\|}, \tag{19}$$

where $v_1 = T_k T_{k-1}$ and $v_2 = T_k T_{k+1}$ present line vectors. Considering that, the optimal number of clusters now can be written as

$$k^* = \underset{k=2,3,\ldots}{\operatorname{argmax}} \alpha_k \tag{20}$$

The initialization step is set as follows: the covariance matrices are set to pondered identity matrices, i.e. $\Sigma_j^{(0)} = \lambda_j \mathbf{I}$, $\lambda_j > 0$; the expectations $\mu_j^{(0)}$ are selected randomly over a dataset

X. In this sense, the n-dimensional Euclidean balls $d_M(x;\theta_j^{(0)}) \leq \chi_n^2(p)$ are introduced to present initial components; the component weights $w_j^{(0)}$ are set to be a ratio of an observed n-dimensional volume of the Euclidean ball and a sum of all initialization balls. The radii are selected to be less than the diameter of X, $\operatorname{diam} X = \max_{i,j\in I} \|x_i - x_j\|$, where the proposed method has shown that if the n-dimensional ball intersects a cluster, in most cases this leads to its detection. The p-quantil value of the $\chi_n^2(p)$ presents a crucial parameter of the flexibility of the adaptive EM, where it is discussed that $p \approx 0.05$ will be reasonable to be observed. For example, the proposed method acts the same as the standard EM if $p \to 0$, or if $p \in (0,1)$ is set to reject a large number of data, then the proposed method will cause continuous shrinking of the clusters, and eventually stop (because of the bad condition matrices). Consequently, numerical research is conducted with $p = 0.05$, where relevant results for certain models are achieved.

The running of the adaptive EM is regulated by the difference of the input parameters, where the Euclidean norm is observed, i.e.

$$\|\Theta^{(t+1)} - \Theta^{(t)}\| = \sqrt{\Delta w + \Delta\mu + \Delta\Sigma}, \tag{21}$$

where

$$\Delta w = \sum_{j\in J} \|w_j^{(t+1)} - w_j^{(t)}\|^2, \tag{22}$$

$$\Delta\mu = \sum_{j\in J} \|\mu_j^{(t+1)} - \mu_j^{(t)}\|^2, \tag{23}$$

$$\Delta\Sigma = \sum_{j\in J} \|\Sigma_j^{(t+1)} - \Sigma_j^{(t)}\|^2, \tag{24}$$

present the standard squared Euclidean norm for scalar, vector and matrix cases. A stoppage of the *Algorithm 1* is established if the difference does not exceed pre-defined $\varepsilon > 0$.

Example 1. *In* Figure 1 *the dataset with and without outliers is presented. The noises are uniformly distributed over the pre-defined rectangle.*

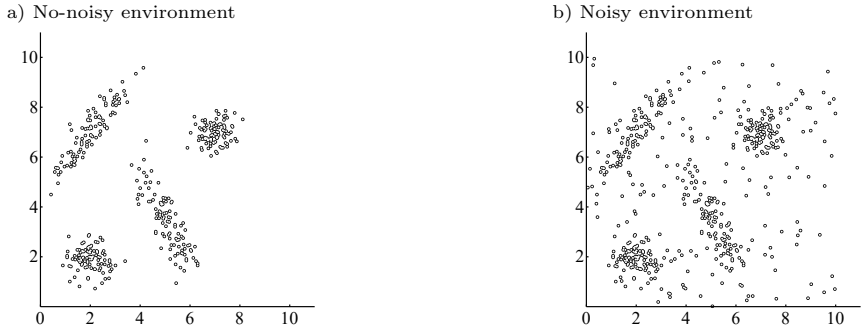

Figure 1: *Dataset*

In Figure 2 *the trends of the proposed validity criteria $\widehat{m}_k/m$ are presented. The results are obtained by running the adaptive EM a few hundred times for each observed k, where $p = 0.05$, $\varepsilon = 0.1$, and $\lambda_j = c$, $\forall j \in J$. The research is conducted with various $c \in \mathbb{R}$ in order to resolve the mentioned problem. The experimental research has shown that the trends of graphs possess the knee points. This is the situation when $\widehat{m}_k/m$ has a monotonous trend after the knee point, which means that the adaptive EM does not obtain remarkably better results for greater number of clusters. Thereby the knee point can be a good candidate for the optimal number of clusters.*

The Figures 2c),d) *presents that the largest angle* α_k *points to the point at* $k^* = 4$, *where* Figure 2c) *correspond to no-noisy case presented in* Figure 2a) *and* Figure 2d) *to noisy case presented in* Figure 2b).

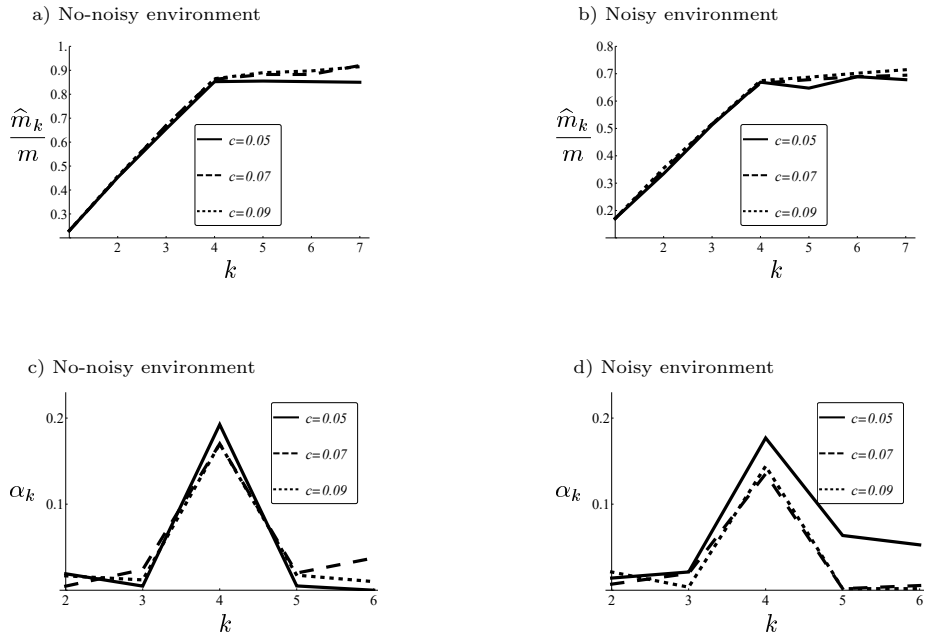

Figure 2: *Cluster validation*

In Figure 3 *the final results of the adaptive EM are presented, which correspond to the optimal number of clusters at* $k^* = 4$. *The red dashed lines present the contours* $d_M(x;\theta_j^{(t)}) = \chi^2_n(p)$ *of the adaptive EM final result, together with* $\mu_j^{(t)}$. *The blue dashed lines present the original contours* $d_M(x;\theta_j) = \chi^2_n(p)$ *and expectations* μ_j *of the component* j, *i.e. cluster* π_j.

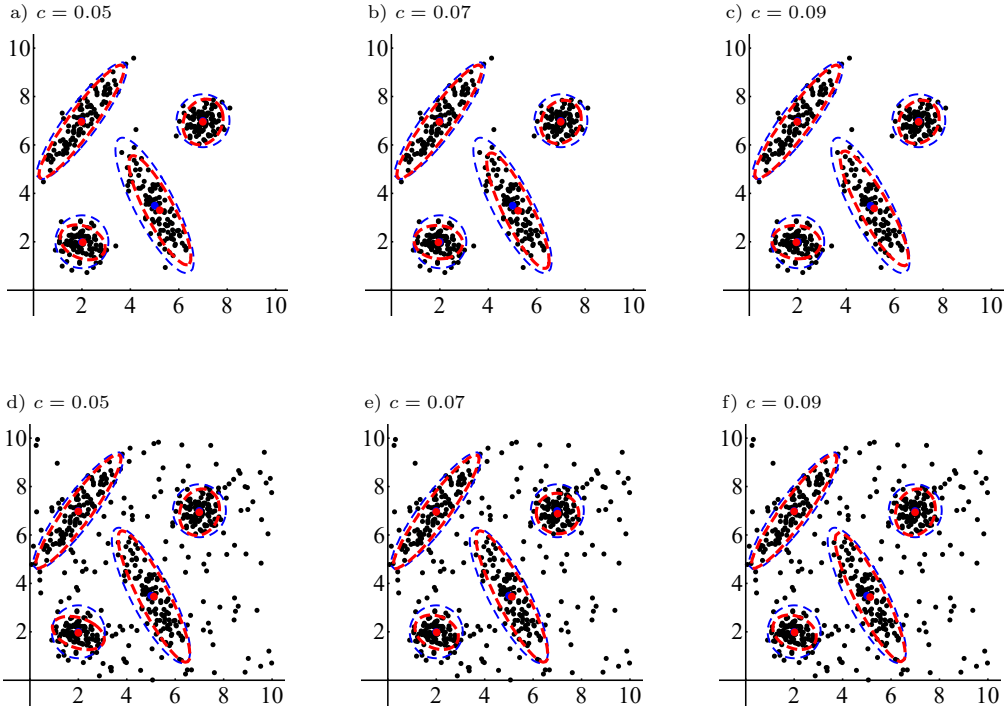

Figure 3: *Final results*

Example 2. *In* Figure 4 *the case of the overlapping clusters is presented, where many center-based indices do not achieve good results* [11].

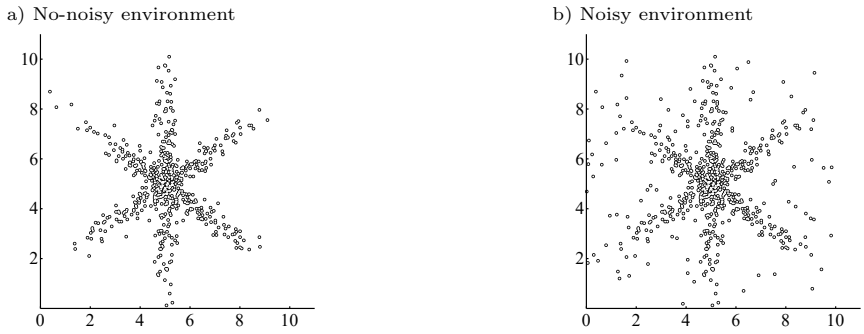

Figure 4: *Dataset*

The Figure 5 *present the trends of the proposed validity criteria. The experimental research has shown again that the trend of graphs possessses the knee point, which can indicate the optimal number of clusters. Besides that, one bad result was obtained, which has happened in a noisy case for* $c = 0.09$. *This is because the adaptive EM is very sensitive to input parameters, and thereby can converge to bad clustering results for a bad choosen initialization. This is the situation when the radius of the initialization Euclidean balls are too large, which may lead to results that include a large amount of outliers.*

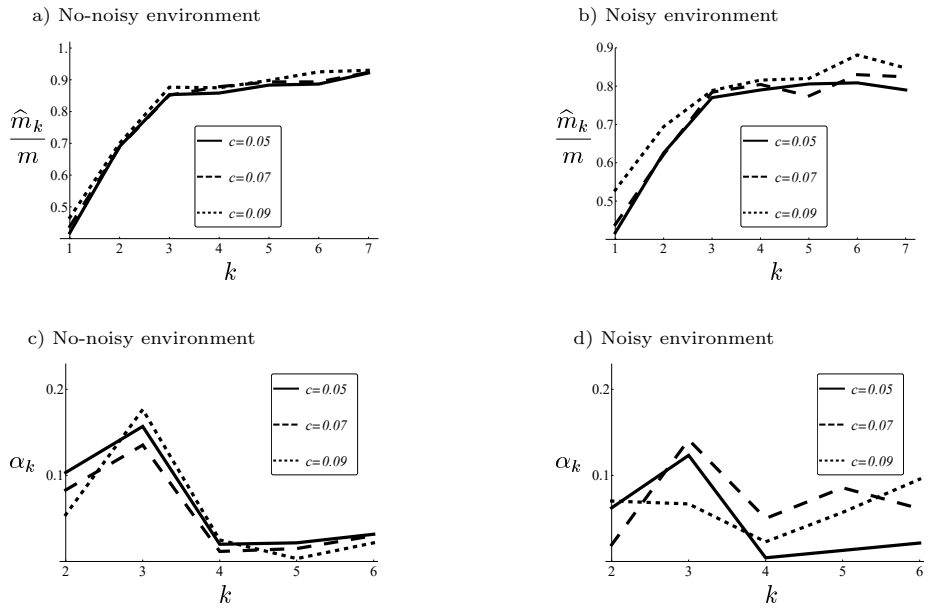

Figure 5: *Clustering validation*

In Figure 6 *the final results of the adaptive EM for* $k^* = 3$ *are presented.*

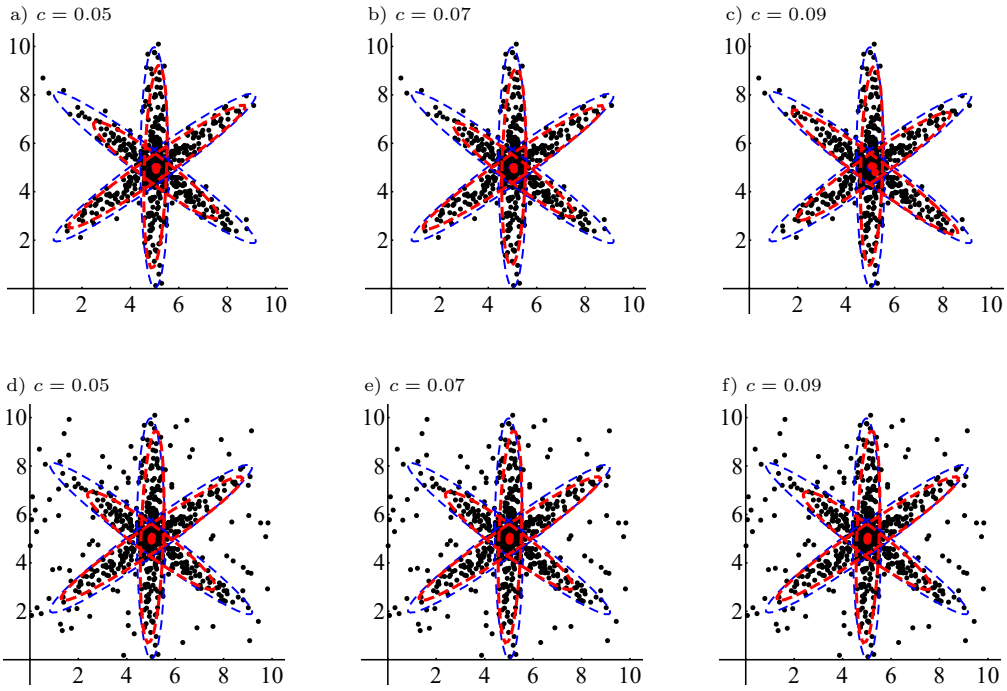

Figure 6: *Final results*

Example 3. *In* Figure 7 *the combination of overlapping and isolated clusters is presented.*

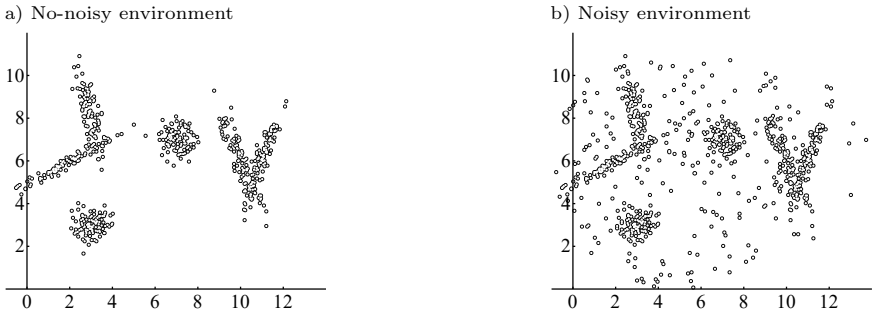

Figure 7: *Dataset*

The Figure 8 *shows again that the trend of graphs can be a good indicator of the optimal number of clusters. Because the adaptive EM is very sensitive to the initialization step, it can happen that the final clusters cover no-noisy data together with a large amount of outliers. This is the situation which is presented in* Figure 8b) *for* $c = 0.09$. *Because of that, the proposed method can converge to bad clustering results for a bad choosen initialization parameters, which means that the final results can lead to the wrong conclusion.*

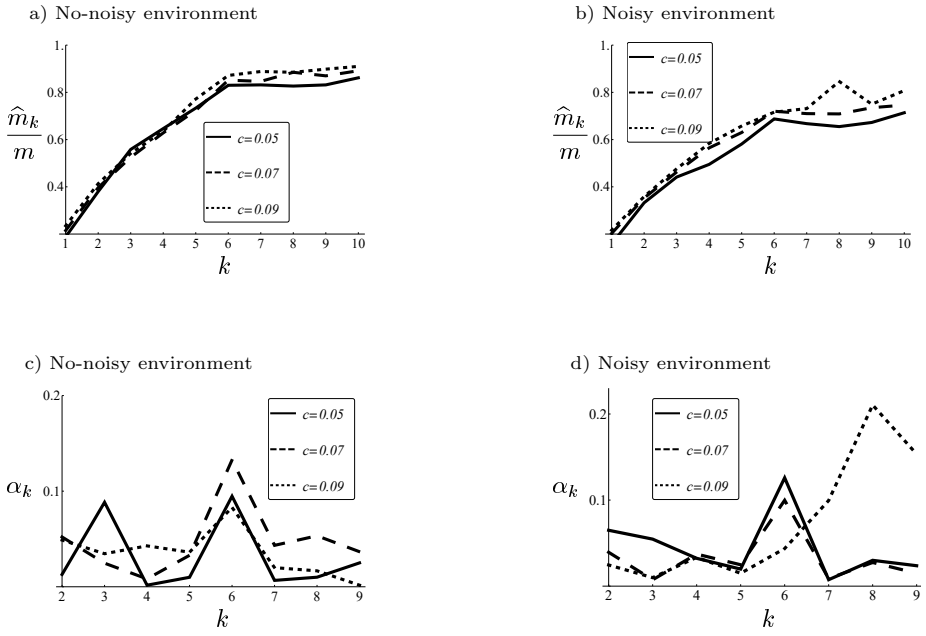

Figure 8: *Clustering validation*

In Figure 9 *the final results of the adaptive EM for* $k^* = 6$ *are presented.*

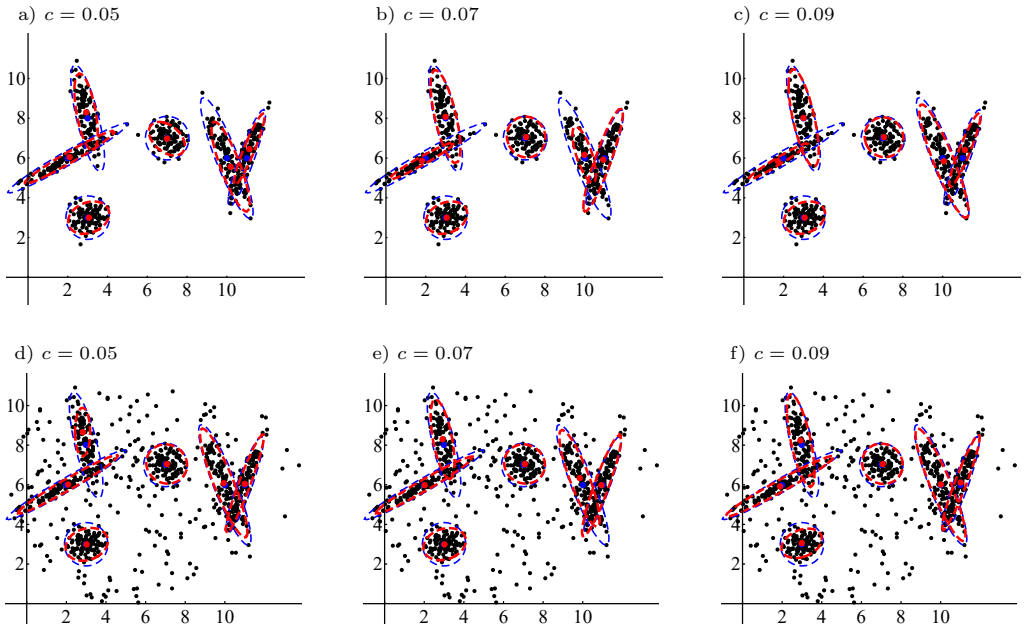

Figure 9: *Final results*

5. Conclusion

It is shown that the proposed method can resolve the problem of finding the optimal number of clusters, or in the other sense, it can detect the desired Gaussian mixture statistical model. Because the datasets are modeled as the Gaussian mixture, the statement that the squared Mahalanobis distance follows the χ^2_n distribution has been shown to be a good choice for modification of the EM algorithm. This statement is often used as an indicator whether a data point may be an outlier, or have a multivariate normal distribution [2, 4], and thereby the restriction on the hidden variable by the defined Kronecker delta presents an excellent choice. Furthermore, the proposed clustering measurement, i.e. percentage of the no-noisy data, has been shown to be an excellent choice for the investigation of the optimal number of clusters, where it is well known that most validity indices are not reliable for the mentioned problem [11]. Among all features, the convergence property of the adaptive EM algorithm is established for every numerical example presented in the paper, where it is shown that relevant results have been achieved.

References

[1] Dempster, A. P., Laird, N. M. and Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistics Society*, 39(1), 1–38. http://web.mit.edu/6.435/www/Dempster77.pdf

[2] Filzmoser, P., Garrett, R. G. and Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. *Computers and Geosciences*, 31(5), 579–587. doi: 10.1016/j.cageo.2004.11.013

[3] Gan, G., Ma, C. and Wu, J. (2007). *Data Clustering: Theory, Algorithms and Applications*. ASA–SIAM Series on Statistics and Applied Mathematics, Philadelphia. doi: 10.1137/1.9780898718348

[4] Johnson, R. A. and Wichern, D. W. (2007). *Applied multivariate statistical analysis* (6th Ed). New Jersey: Pearson Prentice Hall.

[5] Marošević, T. and Scitovski, R. (2015). Multiple ellipse fitting by center-based clustering. *Croatian Operational Research Review*, 6(1), 43–53. doi: 10.17535/crorr.2015.0004

[6] Novoselac, V. and Pavić, Z. (2018). Cluster Detection in the Noisy Environment by Using the Modified EM Algorithm. *Croatian Operational Research Review*, 9(2), 223–234. doi: 10.17535/crorr.2018.0017

[7] Novoselac, V. and Pavić, Z. (2014). Outlier detection in experimental data using a modified expectation-maximization algorithm. *Proceedings of 6th International Scientific and Expert Conference of the International TEAM Society*. Faculty of Mechanical Engineering and Automation, 112–115.

[8] Rousseeuw, P. J. and Leroy, A. M. (2003). *Robust Regression and Outlier Detection*. New York: Wiley.

[9] Scitovski, R. and Scitovski, S. (2013). A fast partitioning algorithm and its application to earthquake investigation. *Computers and Geosciences*, 59, 124–131. doi: 10.1016/j.cageo.2013.06.010

[10] Theodoridis, S. and Koutroumbas, K. (2009). *Pattern Recognition*. Burlington: Academic Press.

[11] Vendramin, L., Campello, R. J. G. B., Hruschka, E. R. (2009). On the Comparison of Relative Clustering Validity Criteria. *Proceedings of the SIAM International Conference on Data Mining*, 733–744. doi: 10.1137/1.9781611972795.63

[12] Wu, J. C. F. (1983). On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1), 95–103. doi: 10.1214/aos/1176346060

[13] Zhao, Q., Xu, M. and Frnti, P. (2008). Knee Point Detection on Bayesian Information Criterion. *20th IEEE International Conference on Tools with Artificial Intelligence*, 431–438. doi: 10.1109/ictai.2008.154