Croatian Operational Research Review
CRORR **9**(2018), 223–234

Cluster Detection in Noisy Environment by Using the Modified EM Algorithm

Vedran Novoselac[1,*] **and Zlatko Pavić**[2]

[1,2] *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, Zlatko.Pavic}@sfsb.hr⟩

Abstract. The paper studies the problem of cluster detection in noisy environment. The solution of this problem is based on the well known Expectation Maximization (EM) algorithm. By utilizing the Mahalanobis distance, and modifying the hidden variable, the rejection procedure is constructed so that it omits data from calculation of the current iteration step. Thus we construct the adaptive framework for solving the above problem. Several numerical examples are presented to illustrate the proposed algorithm.

Key words: clustering, EM, Mahalanobis distance, least squares, least absolute deviation, Davies-Bouldin index

Received: May 29, 2018; accepted: July 20, 2018; available online: December 13, 2018

DOI: 10.17535/crorr.2018.0017

1. Introduction

Clustering is a widely used exploratory data analysis tool that has been successfully applied to data analysis, image processing, pattern recognition, engineering [2, 4, 6, 7, 8, 15, 17, 18], and many other fields. In this paper, we focus on the detection of clusters in a noisy environment based on the well-known EM algorithm [2, 3, 9, 11, 18]. This implies data sets $X = \{x_i \colon i \in I\} \subset \mathbb{R}^n$, $I = \{1, \ldots, m\}$ with the presence of outliers or large database sets [2, 11, 13, 17].

Consequently, the main aim is to detect clusters $\{\pi_j \colon j \in J\}$, $J = \{1 \ldots, k\}$, $k \leq m$, where $\pi_j = \{x_i \colon i \in I_j \subseteq I\}$. As it was earlier mentioned, some data do not belong to any cluster, i.e. they are noisy, and thus they disrupt the clustering process causing degeneration on the final clusters structure. In this situation, the clustering problem becomes even more complicated and requires an effective rejection procedure. The rejection procedure restricts the clustering process on the selected data, resulting in an adaptive process which detects the specific statistical model. Consequently, set $\widetilde{X} = \{x_i \colon i \in \widetilde{I}\} \subseteq X$ is extracted, where $\widetilde{I} = \bigcup_{j \in J} I_j$.

For that purpose we propose a rejection procedure within the EM algorithm by modifying the hidden variable. The aim is to disregard noisy data from further calculation in the current clustering step. In the sense of the Gaussian mixture model, the Mahalanobis distance is used, which is widely applied in application in data clustering analysis [8, 11, 16]. The Mahalanobis distance is used to determine data dispersion within cluster π_j by weighted mean and weighted median of data [13, 14, 19]. In this sense, the problem of dispersion is solved as the LS (Least Squares) and the LAD (Least Absolute Deviation) problems which have a wide variety of applications, such as image processing, clustering, data analysis, outliers detection, pattern recognition etc. [1, 6, 12, 14, 17, 19]. To accomplish a better clustering quality, the dispersion

*Corresponding author.

of each cluster is pondered with a rejection parameter $\alpha > 0$, where data is considered noisy if it exceeds each threshold.

In order to determine the ddesired rejection parameter $\alpha > 0$, calculations on the modified EM are conducted on various numerical examples. To achieve this, a clustering quality is considered. Many different and useful clustering validity measures exist to achieve this purpose. Because there is a range of new measures to choose from, it is not easy to choose a specific one. In our case, a cluster quality measure is provided by observation in the well known Davies-Bouldin index [4, 15, 20]. Besides, the percentage of rejected data is also taken into the observation because it gives good information about the noise ratio.

The paper is organized as follows: in *Section 2* the EM algorithm is briefly introduced, as well as its implementation for the Gaussian mixture model which is presented in *Subsection 2.1.* In *Section 3* the modified EM algorithm and its pseudocode are presented. In *Section 4* we present several illustrative examples, and finally, in *Section 5* we give the conclusion.

2. Clustering with EM algorithm

The EM algorithm is an iterative procedure aiming to compute the maximum likelihood estimate (MLE) of log-likelihood

$$\ln \mathcal{L}(\Theta|X) = \ln P(X|\Theta) \tag{1}$$

where X is some random variable. In order to facilitate the ML estimation parameter Θ for which the observed data are the most likely, the so-called hidden random 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.

$$\ln \mathcal{L}(\Theta|X, Z) = \ln P(X, Z|\Theta). \tag{2}$$

In the presence of hidden data Z, in order to estimate model parameters Θ for which the observed data are the most likely, the EM algorithm iteratively applies the following two steps:

Expectation step (E-step): Calculate the expected value of the complete data log-likelihood function under the current estimate of the parameters $\Theta^{(t)}$:

$$\mathcal{Q}(\Theta|\Theta^{(t)}) = \mathrm{E}_{Z|X,\Theta^{(t)}}[\ln \mathcal{L}(\Theta|X, Z)]. \tag{3}$$

Maximization step (M-step): Find the parameter that maximizes this expectation:

$$\Theta^{(t+1)} = \arg\max_{\Theta} \mathcal{Q}(\Theta|\Theta^{(t)}). \tag{4}$$

Each iteration consists of an E-step which finds the distribution for the unobserved variables Z. In the M-step, the log-likelihood function is maximized under the assumption that the hidden data Z are known. This whole procedure is repeated until some stopping criteria, e.g. until the difference of change between the parameter updates becomes very small using a norm or until convergence of the observed log-likelihood function (1) is established.

Convergence is assured since the algorithm is guaranteed to increase the log-likelihood at each iteration and converge to a global or local maximum of the log-likelihood function [21]. Its convergence also depends on the initial parameters and on the model. However, even if a local maximum solution is reached, it may still capture satisfactorily the clustering structure. Once the algorithm has converged, parameters are assigned to clusters according to the final estimates where data x_i is appointed to cluster π_j if

$$P(\pi_j|x_i) > P(\pi_l|x_i), \quad j, l \in J, j \neq l, \tag{5}$$

where $P(\pi_j|x_i)$ is the probability of x_i belonging to a cluster π_j, and $P(\pi_l|x_i)$ to a cluster π_l, respectively.

2.1. The EM algorithm for the Gaussian mixture model

The Gaussian mixture model is a powerful model for data clustering. It models the data as a mixture of multiple Gaussian distributions where each Gaussian component corresponds to one cluster. The EM algorithm optimizes the Gaussian mixture model

$$P(x|\Theta) = \sum_{j\in J} w_j P(x|\theta_j), \tag{6}$$

where w_j represents the a priori probability of belonging to a corresponding cluster π_j, what directly implies $\sum_{j\in J} w_j = 1$. The parameter $\theta_j = (\mu_j, \Sigma_j)$ is presented with expectation $\mu_j \in \mathbb{R}^n$ and covariance matrix $\Sigma_j \in \mathbb{R}^{n\times n}$ of density function for the multivariate normal (Gaussian) distribution of dimension n, i.e.

$$P(x|\theta_j) = \frac{1}{(2\pi)^{n/2}|\Sigma_j|^{1/2}} e^{-\frac{1}{2}(x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)}. \tag{7}$$

In this case the hidden variable Z determines the component from which the observation originates. A good choice for a hidden variable Z, where $z_i \in \mathbb{R}^k$ is a measurement vector whose entry z_{ij} is equal to one, if and only if, component j (i.e. cluster π_j) contains observation x_i. In that case $P(z_{ij} = 1) = w_j$ and $\sum_{j\in J} z_{ij} = 1$ and we can write

$$P(Z = z_i) = \prod_{j\in J} w_j^{z_{ij}}. \tag{8}$$

Now the complete log-likelihood function can be carried out as

$$\ln \mathcal{L}(\Theta|X, Z) = \ln P(X, Z|\Theta) = \sum_{i\in I}\sum_{j\in J} z_{ij}(\ln w_j + \ln P(x_i|\theta_j)), \tag{9}$$

where the aim is to estimate parameters $\Theta = \{(w_j, \mu_j, \Sigma_j)\colon j \in J\}$ of the Gaussian mixture model alternating E-step and M-step until convergence.

E-step: Thus, the E-step results with the calculation of expectation $\mathcal{Q}(\Theta|\Theta^{(t)})$ where $\Theta^{(t)} = \{(w_j^{(t)}, \mu_j^{(t)}, \Sigma_j^{(t)})\colon j \in J\}$ (and thus $\theta_j^{(t)} = (\mu_j^{(t)}, \Sigma_j^{(t)})$) are current estimation of parameters, i.e.

$$\begin{aligned}\mathcal{Q}(\Theta|\Theta^{(t)}) &= \mathrm{E}_{Z|X,\Theta^{(t)}}[\ln \mathcal{L}(\Theta|X, Z)] \\ &= \sum_{i\in I}\sum_{j\in J} h_{ij}^{(t)}(\ln w_j + \ln P(x_i|\theta_j)).\end{aligned} \tag{10}$$

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

$$h_{ij}^{(t)} = \frac{w_j^{(t)} P(x_i|\theta_j^{(t)})}{\sum_{l\in J} w_l^{(t)} P(x_i|\theta_l^{(t)})}. \tag{11}$$

M-step: In the M-step, an optimum of $\mathcal{Q}(\Theta|\Theta^{(t)})$ must be carried out. It can be analytically solved from the equation $\nabla_\Theta \mathcal{Q}(\Theta|\Theta^{(t)}) = 0$. Easily, the optimal results can be carried

out as:

$$w_j^{(t+1)} = \frac{1}{m}\sum_{i\in I} h_{ij}^{(t)}, \tag{12}$$

$$\mu_j^{(t+1)} = \frac{\sum\limits_{i\in I} h_{ij}^{(t)} x_i}{\sum\limits_{i\in I} h_{ij}^{(t)}}, \tag{13}$$

$$\Sigma_j^{(t+1)} = \frac{\sum\limits_{i\in I} h_{ij}^{(t)}(x_i-\mu_j^{(t+1)})(x_i-\mu_j^{(t+1)})^T}{\sum\limits_{i\in I} h_{ij}^{(t)}}. \tag{14}$$

3. Modified EM algorithm

Considering the problem of finding clusters in data set X in presence of noisy data, it is possible that the standard EM algorithm can not accomplish the desired results. That is because the standard EM processes all data which may affect the clustering structure and consequently alter the final results. This fact requires the data omission in the current EM steps in order to preserve the cluster structure. This is carried out by modifying the hidden variable Z to $\widetilde{Z}$, where $\tilde{z}_i = \tilde{\delta}_i z_i$. A noisy data is determined by $\tilde{\delta}_i \in \mathbb{R}$ which is defined as

$$\tilde{\delta}_i = \begin{cases} 1, \ \|\delta_i\| \geq 1; \\ 0, \ \text{else}, \end{cases} \tag{15}$$

where $\delta_i \in \mathbb{R}^k$ presents a vector whose components δ_{ij} are the Kronecker deltas defined as

$$\delta_{ij} = \begin{cases} 1, \ d_M(x_i;\theta_j) \leq \alpha\sigma_j; \\ 0, \ \text{else}. \end{cases} \tag{16}$$

As it is defined in (16), observation x_i is detected as noisy if $d_M(x_i;\theta_j)$ exceeds the predefined pondered value σ_j, which presents data dispersion of the observed cluster π_j. Parameter $\alpha > 0$ presents a fine regulation factor of data rejection. The function $d_M \colon \mathbb{R}^n \to \mathbb{R}$ presents the Mahalanobis distance defined as

$$d_M(x_i;\theta_j) = \sqrt{(x_i-\mu_j)^T\Sigma_j^{-1}(x_i-\mu_j)}. \tag{17}$$

Dispersion σ_j is derived directly from corresponding cluster π_j as the minimization problem

$$\sigma_j = \underset{a\in\mathbb{R}}{\operatorname{argmin}} \sum_{i\in I_j} P(x_i|\theta_j)(a - d_M(x_i;\theta_j))^2, \tag{18}$$

where $I_j = \{i \in I \colon \delta_{ij} = 1\}$. In this situation, problem (18) is solved in the sense of ℓ_2 norm and is known as the LS problem where the optimal solution is called a weighted mean of data [13]. If the problem of dispersion is observed in the sense of ℓ_1 norm, i.e.

$$\sigma_j = \underset{a\in\mathbb{R}}{\operatorname{argmin}} \sum_{i\in I_j} P(x_i|\theta_j)|a - d_M(x_i;\theta_j)|, \tag{19}$$

then problem (19) is known as the LAD problem where the optimal solution is called a weighted median of data [13, 14, 19]. In both cases, weights are set to be a value of probability density

function, which gives a greater impact on the dispersion of those data which are better grouped. Finally, the probability of hidden variable $\widetilde{Z}$ is defined as

$$P(\widetilde{Z} = \tilde{z}_i) = \tilde{\delta}_i \prod_{j \in J} w_j^{z_{ij}}. \tag{20}$$

Consequently, some data do not belong to any cluster, and thus they are rejected and considered as noise. In this case $\tilde{\delta}_i = 0$, which directly indicates that the probability of belonging to any cluster is equal to zero.

To execute the E-step and the M-step, the complete log-likelihood function $\ln \mathcal{L}(\Theta|X, \widetilde{Z})$ must be derived. In this situation, we redefine probability (20) to the expression with the same operation defined as

$$P(Z = z_i) = \prod_{j \in J} w_j^{z_{ij}}, \quad i \in \widetilde{I} = \{i \in I \colon \tilde{\delta}_i = 1\}. \tag{21}$$

Consequently, $\ln \mathcal{L}(\Theta|X, \widetilde{Z})$ can be rewritten to

$$\ln \mathcal{L}(\Theta|\widetilde{X}, Z) = \sum_{j \in J} \sum_{i \in \widetilde{I}} z_{ij} (\ln w_j + \ln P(x_i|\theta_j)). \tag{22}$$

This implies the same optimization procedure as for the standard EM, with initialization step $\Theta^{(t)} = \{(w_j^{(t)}, \mu_j^{(t)}, \Sigma_j^{(t)}) \colon j \in J\}$, together with $\sigma_j^{(t)} > 0$, $j \in J$, and the pre-defined rejection parameter $\alpha > 0$. Now $\widetilde{I}^{(t)} = \{i \in I \colon \tilde{\delta}_i^{(t)} = 1\}$ can be easily conducted, where

$$\delta_{ij}^{(t)} = \begin{cases} 1, \, d_M(x_i; \theta_j^{(t)}) \leq \alpha \sigma_j^{(t)}; \\ 0, \, \text{else}, \end{cases} \tag{23}$$

and hence

$$\sigma_j^{(t)} = \operatorname*{argmin}_{a \in \mathbb{R}} \sum_{i \in I_j^{(t)}} P(x_i|\theta_j^{(t)})(a - d_M(x_i; \theta_j^{(t)}))^2, \tag{24}$$

or

$$\sigma_j^{(t)} = \operatorname*{argmin}_{a \in \mathbb{R}} \sum_{i \in I_j^{(t)}} P(x_i|\theta_j^{(t)})|a - d_M(x_i; \theta_j^{(t)})|, \tag{25}$$

where $I_j^{(t)} = \{i \in I \colon \delta_{ij}^{(t)} = 1\}$. Finally, it can be written that

$$\tilde{\delta}_i^{(t)} = \begin{cases} 1, \, \|\delta_i^{(t)}\| \geq 1; \\ 0, \, \text{else}. \end{cases} \tag{26}$$

The results of the M-step now can be easily conducted as:

$$w_j^{(t+1)} = \frac{1}{\widetilde{m}^{(t)}} \sum_{i \in \widetilde{I}^{(t)}} h_{ij}^{(t)}, \tag{27}$$

$$\mu_j^{(t+1)} = \frac{\sum\limits_{i \in \widetilde{I}^{(t)}} h_{ij}^{(t)} x_i}{\sum\limits_{i \in \widetilde{I}^{(t)}} h_{ij}^{(t)}}, \tag{28}$$

$$\Sigma_j^{(t+1)} = \frac{\sum\limits_{i \in \widetilde{I}^{(t)}} h_{ij}^{(t)} (x_i - \mu_j^{(t+1)})(x_i - \mu_j^{(t+1)})^T}{\sum\limits_{i \in \widetilde{I}^{(t)}} h_{ij}^{(t)}}, \tag{29}$$

where $\widetilde{m}^{(t)} = |\widetilde{I}^{(t)}|$.

Because of the noisy environment, the initialization step presents a hard task, where the goal is to refine the initial mixture model parameter to a better fit. A standard practice for the execution of *Algorithm 1* is to call a running for many different initial parameter values and choose the mixture model with the best quality. Nevertheless, the initial parameters should be reasonably determined. For example, initial component covariance matrices are set to the identity matrix, i.e. $\Sigma_j^{(0)} = \mathbf{I}$. Initial expectations $\mu_j^{(0)}$ can be efficiently selected by visual insight or randomly generated over the data set X. In this sense the n-dimensional balls $d_M(x, \theta_j^{(0)}) \leq r_j$ are introduced to present initial components defined by $\delta_{ij}^{(0)}$. The radii r_j are selected to be less than the diameter of X, i.e. $r_j < \operatorname{diam} X$, $\operatorname{diam} X = \max_{i,j \in I} \|x_i - x_j\|$, where the proposed method has shown good properties if the n-dimensional ball intersects cluster π_j, which in most cases leads to its detection. Furthermore, the determination of r_j can be effectively done by visual insight, or by further research (e.g. aim is to accomplish $r_j \leq \operatorname{diam} \pi_j$). Now data dispersions $\sigma_j^{(0)}$ can be easily calculated by (24) or (25), as well as the component weights

$$w_j^{(0)} = \frac{r_j^n}{\sum\limits_{l \in J} r_l^n}, \tag{30}$$

which are obtained as the ratio of the observed n-dimensional volume of a Euclidean ball, with a sum of all initialization balls. The determination of the optimal number of clusters is not taken into the observation wherever it is supposed that it is known. A good insight into the problem can be achieved by observing different numbers of clusters and choosing number k with the best cluster quality measurement [10]. A pseudo-code for the modified EM is presented in *Algorithm 1*.

By including the initialization step, the problem of the cluster detection becomes very sensitive and multiple local maxima may occur. The stoppage criteria based on the difference between parameter values is constructed, i.e.

$$\|\Theta^{(t+1)} - \Theta^{(t)}\| = \sqrt{\Delta^2 w^{(t+1)} + \Delta^2 \mu^{(t+1)} + \Delta^2 \Sigma^{(t+1)}} \tag{31}$$

where

$$\Delta^2 w^{(t+1)} = \sum_{j \in J} \|w_j^{(t+1)} - w_j^{(t)}\|^2, \tag{32}$$

$$\Delta^2 \mu^{(t+1)} = \sum_{j \in J} \|\mu_j^{(t+1)} - \mu_j^{(t)}\|^2, \tag{33}$$

$$\Delta^2 \Sigma^{(t+1)} = \sum_{j \in J} \|\Sigma_j^{(t+1)} - \Sigma_j^{(t)}\|^2, \tag{34}$$

present the standard squared Euclidean norm for scalar, vector and matrix cases. If the difference does not exceed pre-defined $\varepsilon > 0$, then the stoppage of *Algorithm 1* is achieved. The proposed method has convergence property, where a problem occurs for situations when initialization component has no data, or rejection parameter $\alpha > 0$ is too small, which causes continuous cluster shrinking.

Algorithm 1 The modified EM algorithm for the Gaussian mixture

1: $\Theta^{(0)} = \{(w_j^{(0)}, \mu_j^{(0)}, \Sigma_j^{(0)})\colon j \in J\}$, $r_j > 0$, $\sigma_j^{(0)} > 0$, $\forall j \in J$, $\alpha > 0$, $\varepsilon > 0$, $t = 0$;

2: *loop*:

3: Calculate:

$$\widetilde{I}^{(t)} = \{i \in I\colon \tilde{\delta}_i^{(t)} = 1\};$$

4: **E-step:** For every $i \in \widetilde{I}^{(t)}$ and every $j \in J$ calculate cluster probability:

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

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

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

$$\mu_j^{(t+1)} = \frac{\sum\limits_{i \in \widetilde{I}^{(t)}} h_{ij}^{(t)} x_i}{\sum\limits_{i \in \widetilde{I}^{(t)}} h_{ij}^{(t)}};$$

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

6: Iterate until:

if $\|\Theta^{(t+1)} - \Theta^{(t)}\| \leq \varepsilon$ **then** STOP;
else go to *loop* **and** $t = t + 1$;

Remark 1. *If parameter $\alpha > 0$ satisfies next inequality for each iteration step of the modified EM, i.e.*

$$\alpha \geq \max_{i \in I, j \in J} \frac{d_M(x_i; \theta_j^{(t)})}{\sigma_j^{(t)}}, \quad \forall t, \tag{35}$$

then $\widetilde{I}^{(t)} = I$, $\forall t$, i.e. all data are considered as no-noisy. This implies that the modified EM acts the same as the standard EM for $\alpha \gg 0$.

Remark 2. *The squared Mahalanobis distance approximates a Chi-squared distribution χ_n^2 with n degrees of freedom what can indicate whether a data point may be an outlier or have a multivariate normal distribution [5]. In this case observation x_i belongs to cluster π_j if it satisfies*

$$d_M^2(x_i, \theta_j) \leq \chi_n^2(p), \tag{36}$$

where $p \in (0, 1)$ determines the p-quantile value, and therefore $p \approx 0.05$ will be reasonable, preventing the influence of those observations which are located on the tail regions of the Gaussians of low probability. This fact leads to the approximation of the rejection parameter $\alpha > 0$, which can be conducted from (16) and (36) as $\alpha \approx \frac{1}{\sigma_j} \chi_n(p)$.

4. Numerical examples

This section provides a few numerical examples. Data sets are distributed by the Gaussian mixture model, whereas noises are added randomly over a pre-defined region or distributed by a statistical model. In order to observe the clustering efficiency of the modified EM, the well-known Davies-Bouldin (DB) index is observed, which is defined as

$$\mathrm{DB} = \frac{1}{k}\sum_{j\in J} R_j, \tag{37}$$

where

$$R_j = \max_{\substack{l\in J\\ j\neq l}} R_{jl}, \tag{38}$$

$$R_{jl} = \frac{r_j + r_l}{D_{jl}}, \tag{39}$$

$$D_{jl} = d(\mu_j, \mu_l), \tag{40}$$

$$r_j = \frac{1}{m_j}\sum_{i\in I_j} d(x_i, \mu_j). \tag{41}$$

In (40) and (41) $d\colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ presents the Euclidean distance measurement, while m_j in (41) presents a cardinal number of π_j, i.e. $m_j = |\pi_j|$. The lower DB index indicates better clustering properties. The percentage of the rejected data $\tilde{\pi} = X\backslash\widetilde{X}$, i.e. noise ratio

$$\widetilde{w} = \frac{m - \widetilde{m}}{m} \tag{42}$$

is also taken into the observation, where $\widetilde{m} = |\widetilde{X}|$. The stoppage criteria iteration rate is also observed for the threshold $\varepsilon > 0$.

Example 1. *The data set X is generated by the Gaussian distributions*

$$X_j \sim \mathcal{N}(\mu_j, \Sigma_j), \quad j = 1, 2, 3,$$

while noisy data are generated by a random vector, uniformly distributed over a pre-defined rectangle $\Omega = [a, b] \times [c, d]$, i.e.

$$X_4 \sim \mathcal{U}(\Omega).$$

The number of data generated by the corresponding random vectors are $m_j = 50$, $j = 1, 2, 3$, and $m_4 = 100$, i.e. $m = m_1 + m_2 + m_3 + m_4 = 250$.

In Figure 1a) the initialization step is presented together with data set X. Contours of $d_M(x; \theta_j^{(0)}) = r_j$, $\forall j \in J$, are presented with red-dashed lines, together with their corresponding expectations $\mu_j^{(0)}$, also marked red. Figures 1b) and 1c) present final results of the modified EM in the sense of the LS($\alpha = 3$) and the LAD($\alpha = 3$) problems. Now red-dashed lines represent ellipses $d_M(x; \theta_j^{(t)}) = \alpha\sigma_j^{(t)}$, $\forall j \in J$, with their corresponding expectation $\mu_j^{(t)}$. The blue-dashed lines present the original contours $d_M(x; \theta_j) = \alpha\sigma_j^{(t)}$ and expectations μ_j, $\forall j \in J$, as a visual insight into the clustering quality. In Figures 1d), 1e), and 1f) trends are presented via α of the DB index, noise ratio $\widetilde{w}$, and the iteration rate t of the stoppage criteria, where each initialization step is set to be as it is presented in Figure 1a). The solid graph presents the modified EM in the sense of the LS, while dashed in the sense of the LAD problem. It can be seen from Figure 1d) that both graphs have breakdown points, indicating cluster distortion. At the same time, a great amount of data is associated with the clusters, which can be seen from

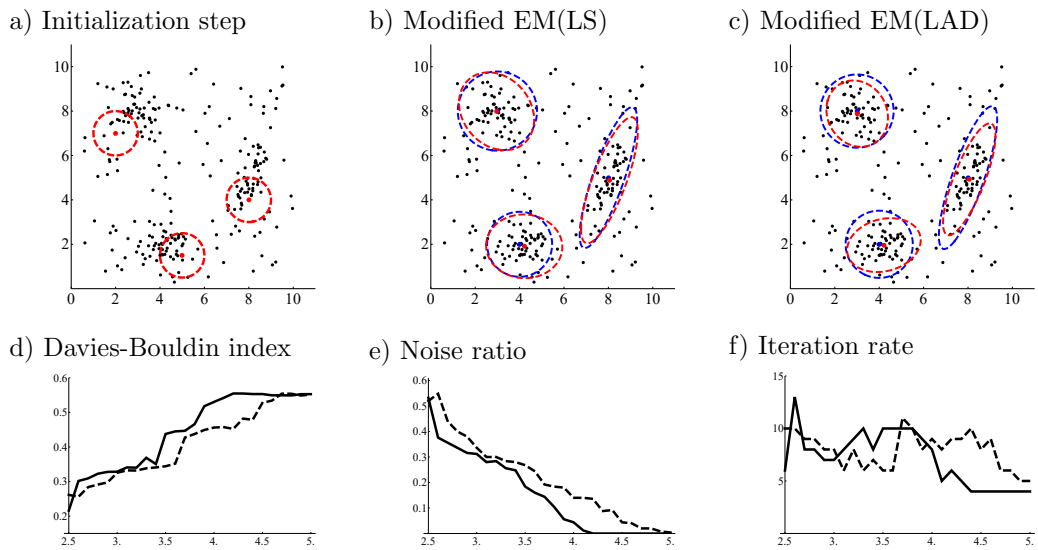

Figure 1: *Cluster detection with the modified EM.*

the graphs presented in Figure 1e). Finally, this fact suggests that some data are noise, and can point to a rejection parameter α. In our case $\alpha = 3$ points to a good cluster quality and stable noise ratio at the same time. In Figure 1f) the iteration rate of stoppage criteria for $\varepsilon = 0.1$ is presented.

Example 2. *The data set X is generated by the Gaussian distributions*

$$X_j \sim \mathcal{N}(\mu_j, \Sigma_j), \quad j = 1, 2, 3, 4,$$

while noisy environment is generated by the random vector

$$X_5 \sim \mathcal{N}(\varphi(\rho), \Sigma_5), \quad \rho \sim \mathcal{U}([a, b]),$$

where $\varphi\colon \mathbb{R} \to \mathbb{R}^2$ presents a curve. In this example we obsereved $\varphi(u) = s_0 + (\beta_1 \sin(\gamma_1 u), \beta_2 \sin(\gamma_2 u))$, $u \in [a, b]$, and some $s_0 \in \mathbb{R}^2$. The number of data generated by corresponding random vectors are $m_j = 50$, $j = 1, 2$, and $m_j = 150$, $j = 3, 4$, and $m_5 = 450$, i.e. $m = m_1 + m_2 + m_3 + m_4 + m_5 = 950$.

Analogously to Example 1, in Figure 2a) the initialization step is presented together with data set X with the same initial covariance matrix. Figures 2b) and 2c) present final results where $\alpha = 3$ is observed for both situations. In Figures 2d), 2e) and 2f) trends are presented via α of DB index, noise ratio $\widetilde{w}$, and iteration rate t for $\varepsilon = 0.1$. The graphs in the LAD sense show that for an $\alpha \approx 2.5$ the method has no results. In this situation the cluster shrinks, which leads to bad condition matrices, causing the stoppage of Algorithm 1.

Example 3. *The data set X is generated by the Gaussian distributions*

$$X_j \sim \mathcal{N}(\mu_j, \Sigma_j), \quad j = 1, 2, 3, 4, 5, 6,$$

whereas noisy environment is generated by the random vector

$$X_7 \sim \mathcal{U}(\Omega),$$

where $\Omega = \{x \in \mathbb{R}^2\colon (d(x, s_1) < \beta_1) \,||\, (\beta_2 < d(x, s_2) < \beta_1)\}$, $0 < \beta_2 < \beta_1$, and some $s_1, s_2 \in \mathbb{R}^2$ such that $d(s_1, s_2) < \beta_1$. The number of data generated by the random vectors are $m_j = 90$, $j = 1, 2, 3, 4, 5, 6$, and $m_7 = 400$, i.e. $m = 960$.

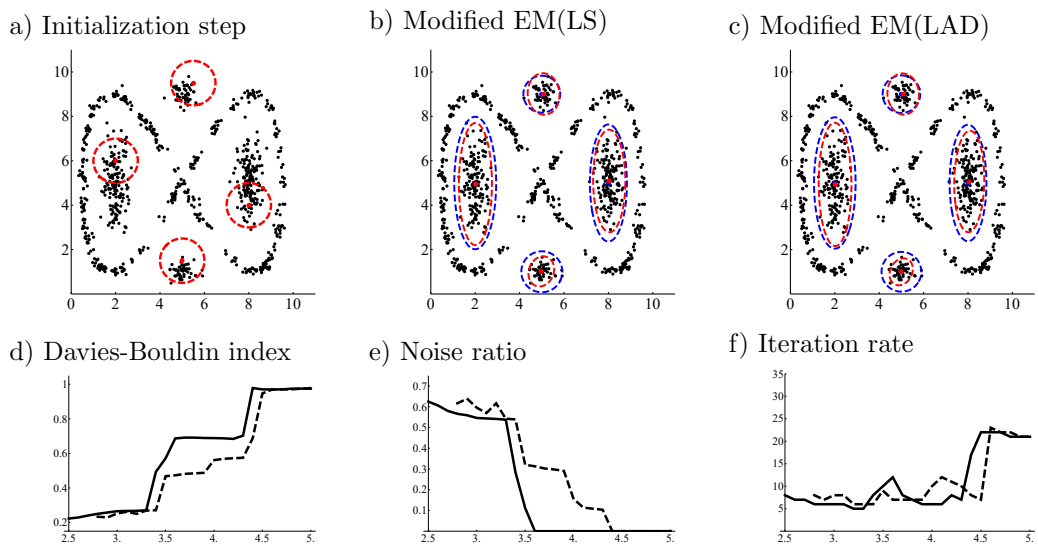

Figure 2: *Cluster detection with the modified EM.*

Analogously to Example 1, Figure 3a) presents the initialization step together with data set X. *Figures 3b) and 3c) present final results, where* $\alpha = 3$ *is observed for both situations. Now DB index presented in Figure 3d) is not reliable because some clusters are intersected. However, the noise ratio, presented in Figure 3e), shows the real situation from which the rejection parameter can be easily determined, i.e.* $\alpha = 3$. *In Figure 3f) the iteration rate of stoppage criteria for threshold* $\varepsilon = 0.1$ *is presented.*

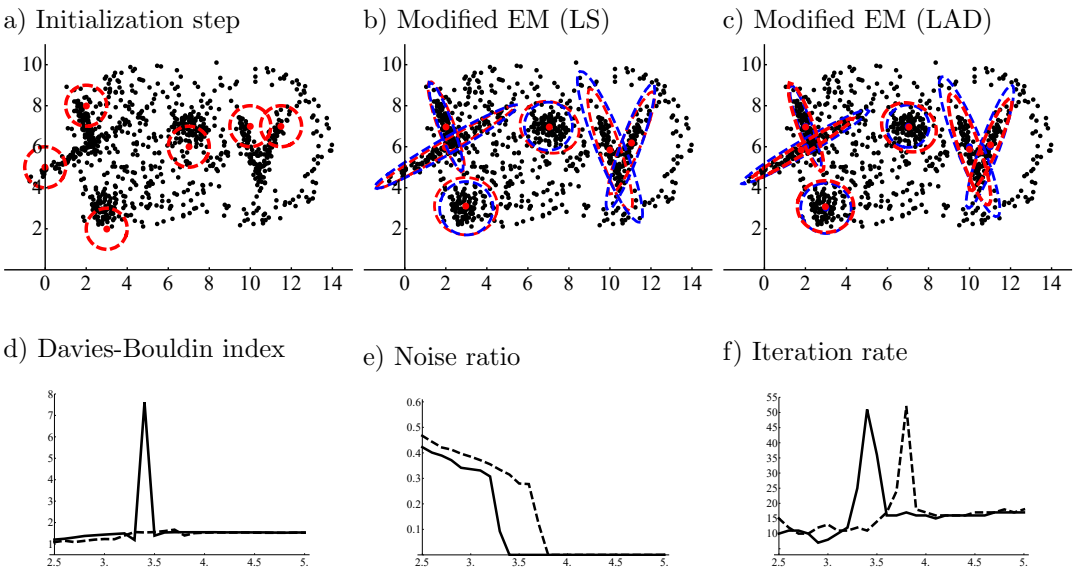

Figure 3: *Cluster detection with the modified EM.*

5. Conclusion

Cluster detection in a noise environment with implemented modification within the EM algorithm proved to be a good choice. It has been shown that the rejection procedure based on the Mahalanobis distance preserves the clusters structure from noise data and effectively extracts the requested mixture model. The noise ratio has proven to be a good choice alongside the Davies-Bouldin index, where results showed a stability of the indicated parameter $\alpha = 3$ for the LS and the LAD modifications, respectively. Besides, for those rejection parameters results show very similar characteristics for both modifications, where *Algorithm 1* in the LAD sense proves to be more robust on the outliers [13], which can be seen from the graphs displaying the noise ratio trends. The iteration rate results for the criterion $\varepsilon = 0.1$ of the parameter $\alpha = 3$ are also satisfactory and stable, where the convergence of *Algorithm 1* is established for every properly chosen initialization step.

References

[1] Bloomfield, P. and Steiger, W. (1983). Least Absolute Deviations: Theory, Applications and Algorithms, Birkhauser, Boston.

[2] Bradley, P. S., Fayyad, U. M. and Reina, C. A. (1999). Scaling EM (Expectation-Maximization) Clustering to Large Databases, Microsoft Research.

[3] 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. Series B (Methodological), 39(1), 1–38.

[4] Gan, G., Ma, C. and Wu, J. (2007). Data Clustering, Theory, Algorithms and Applications, SIAM, Philadelphia.

[5] Johnson, R. A. and Wichern, D. W. (2007). Applied multivariate statistical analysis (6th ed.). Upper Saddle River, New Jersey: Pearson Prentice Hall.

[6] Krpić, Z., Martinović, G. and Vazler, I. (2010). Data clustering: Applications in engineering, Croatian Operational Research Review, 1(1), 180–189.

[7] Marošević, T. (2014). Data clustering for circle detection. Croatian Operational Research Review, 5(1), 15–24. doi:10.17535/crorr.2014.0025

[8] 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

[9] McLachlan, G. J. and Krishnan, T. (2008). The EM Algorithm and Extensions, John Wiley & Sons, Inc., Hoboken, New Jersey.

[10] Novoselac, V. and Pavić, Z. (2016). Optimal number of clusters provided by k-means and E-M algorithm, Proceedings of 8th International Scientific and Expert Conference of the International TEAM Society, Trnava: Faculty of Materials Science and Technology in Trnava, 286–291.

[11] 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, Kecskemét: Faculty of Mechanical Engineering and Automation, 112–115.

[12] Pitas, I. (2000). Digital Image Processing Algorithms and Applications, John Wiley & Sons.

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

[14] Sabo, K. and Scitovski, R. (2008). The best least absolute deviations line-properties and two efficient methods for its derivation, ANZIAM Journal. 50(2), 185–198.

[15] 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

[16] Scitovski, S. and Šarlija, N. (2014). Cluster analysis in retail segmentation for credit scoring. Croatian Operational Research Review, 5(2), 235–245. doi:10.17535/crorr.2014.0010

[17] Taler, P. and Sabo, K. (2014). Color image segmentation based on intensity and hue clustering - a comparison of LS and LAD approaches. Croatian Operational Research Review, 5(2), 375–385. doi:10.17535/crorr.2014.0020

[18] Theodoridis, S. and Koutroumbas, K. (2009). Pattern Recognition. Acad. Press, Burlington.
[19] Vazler, I., Sabo, K. and Scitovski, R. (2012). Weighted median of the data in solving least absolute deviations problems, Communications in Statistics-Theory and Methods, 41(8), 1455–1465.
[20] Vendramin, L., Campello, R. J. G. B. and Hruschka, E. R. (2009). On the Comparison of Relative Clustering Validity Criteria, Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA. SIAM, 733–744.
[21] Wu, C. F. J. (1983). On the Convergence Properties of the EM Algorithm, The Annals of Statistics, 11(1), 95–103.