IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 3, 2021, accepted November 17, 2021, date of publication November 23, 2021,
date of current version December 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3130066

A Novel Approach for Gaussian Mixture Model Clustering Based on Soft Computing Method

MARUF GOGEBAKAN
Department of Maritime Business and Administration, Maritime Faculty, Bandirma Onyedi Eylül University, 10200 Bandırma/Balıkesir, Turkey
e-mail: mgogebakan@bandirma.edu.tr

ABSTRACT Determining the number of clusters in a data set is a significant and difficult problem in cluster analysis. In this study, a new model-based clustering approach is proposed for the estimation of the number of clusters. In the proposed method, the number of components in each variable is determined by using univariate Gaussian mixture models. The number of alternative cluster centres and mixture models was determined according to the number of components in heterogeneous variables. In this study, appropriate Gaussian mixture models were determined with the help of ''mixture model soft computing method'' for the first time. Vector arrays showing the number and addresses of clusters in appropriate Gaussian mixture models were created, and according to the parameter estimations of these models that fit the arrays, the best model was obtained through information criteria. The clustering success achieved with the proposed mixture model soft computing method was compared with the results of Gaussian mixture model clustering methods namely, mclust, clustvarsel, varselLCM, selvarMix and vscc model selection methods in R package. All respective methods analyse and determine the number of clustering for the data sets, synthetic-1, synthetic-2, Iris, and Landsat Satellite Image data sets, respectively and evaluate the correct classification rate. The results revealed that the proposed method shows better results for the determination of number of clustering as well as correct classification rate. The novelty of the study is that a new model-based dimension reduction method is proposed for the estimation of the number of clusters. A deterministic clustering approach is proposed for clustering and classification success on reduced data.

INDEX TERMS Model-based clustering, variable selection, mixture model soft computing method, appropriate Gaussian mixture models, information criteria, components of heterogeneous variable.

I. INTRODUCTION

Model-based clustering is widely used in cluster analysis for clustering data from the mixture of Gaussian distributions. McLachlan and Rathnayake, Bozdogan, Scrucca and Raftery, and McNicholas are some of those who use the mixture of multivariate Gaussian distributions in cluster analysis [1]–[4].

In model selection studies for the perspective and strategies of mixture models, Celeux *et al.* proposed using cluster analysis based on mixture models to determine the number of components (g) in the finite mixture models [5]. In multivariate data, components in the heterogeneous variable are used to determine the number and the location of clusters in the mixture model [6]. Each sub-group (component) in the variables corresponds to at least one cluster in the mixture model [7]. In model-based clustering, mixture models are created according to the number of components in the variables or subsets of variables. When the number of components in variables is for $g < 2$, it is called a homogeneous variable, and because this variable does not have any effects in creating a subset, it is excluded from calculations [8]. Galimberti and Soffritti obtained multiple cluster structures in mixture models, depending on the number and location of subgroups in variables [9]. Galimberti *et al.* presented the mixture components of heterogeneous variables as variable sub-vectors in their study, and they defined how the components in variables affect clustering in model-based clustering. In this study, it was explained that each sub-vector in the variables has information on at least one set [10]. Akogul and Erisoglu proposed a model-based clustering method that uses Analytic Hierarchy Process (AHP) to reveal clustering in the data set. The proposed AHP method was used to determine the best model among the conditions based on certain criteria [11]. A variable/feature selection approach,

The associate editor coordinating the review of this manuscript and approving it for publication was Alberto Cano.

which is based on Bayesian factors, was used to select the best model among the subsets that will occur in model-based clustering. The most appropriate model is determined among the candidate sub-clusters according to the assumption based on the Bayesian Information Criterion (BIC) difference [12]. It is very important in clustering analysis to prevent information loss within the variables that are reduced while variable selection occurs. There are subspace learning feature selection methods that improve learning performance by using the local discriminant information and geometry information found in the original data. [13], [14]. In multivariate data, mixture models were created based on the numbers and volume of the components in the variables [15]. Fop and Murphy generalized variable selection methods according to the related and unrelated variables in model-based clustering. They named the variables that did not affect cluster formation in model-based clustering and did not contain useful information in terms of group membership as "Redundant variables". In their paper, the methods based on mclust namely, mclust, clustvarsel, varselLCM, selvarMix and vscc were applied on a synthetic data and compared results [16].

The choice of assignment algorithms is important when assigning observations to components in variables. The chosen distance function is important when assigning the observations to the components in the variables. Therefore, an adapted similarity measure is used in the cluster analysis [17]–[20]. Components in heterogeneous variables are assigned observations based on their means. Since the volume of each component in the variables is different, k-means algorithms assign a different number of observations to the components according to their means. Different observation numbers assigned to the components of the variables provide suitable solutions for EVI-VVV types from parsimonious models with a different covariance matrix structure. Covariance matrices, which are obtained from different number and size components of variables, significantly affect the number and location of clusters in mixture model clustering [21]. Finite mixture models in a grid structure based on the number of components are obtained from multivariate Gaussian mixture distributions. Among the mixture models obtained according to the determined cases, the best models are selected based on information criteria [22].

In this study, a new model-based approach is proposed for cluster number estimation of multivariate data based on Gaussian mixture models (GMM). The algorithmic method, developed based on soft computing, consists of variable/feature selection, creation of mixture models and selection of the best model. The proposed method was applied on two synthetic datasets and two real datasets namely, Iris (UCI), Landsat satellite Image dataset. The results obtained from the application were compared with the well-known methods namely mclust, clustvarsel, varselLCM, selvarMix and vscc. All obtained results show that the proposed clustering algorithm outperforms existing approaches.

The contributions of this paper are as follows:

(1) The variable/feature selection method was developed with univariate mixed models for the data set.

(2) By defining the grid structured mixed models based on the component numbers in the variables, the model numbers in the search space were obtained.

(3) Appropriate-GMMs were obtained according to the number of components falling into the variables in the reduced data. Vector representations were defined for A-GMMs and the parameters of the models were calculated from linear models.

(4) Information criteria for finite mixed models are calculated and the best model is obtained based on information criteria.

The study is organized as follows. In Section 2, MMSCM and model-based clustering stages are explained for the proposed number of clusters estimation approach. In Section 3.1, all steps are explained on the synthetic-1 and fifteen-variable synthetic-2 data sets, which are simple and comprehensive, respectively, to facilitate the understanding of the optimum cluster number estimation method with MMSCM. In Section 3.2, the recommended method (MMSCM) was applied on two real data sets and the results are compared with well-known methods of GMM based clustering. In Section 4, the results of the study are discussed and compared and the success of the method is presented.

II. MATERIALS AND METHODS

A. THE MODEL-BASED CLUSTERING

Grid structured models are created with the components of each variable in multivariate data. The number of AGMMs among mixture models with grid structure is determined by using "MMSCM" model-based clustering. The number and volume of components of the variables in the mixture model reveal the number and structure of clusters in multivariate data. An algorithmic clustering method, which consists of five steps, is proposed for the estimation and clustering of the cluster number. Model-based clustering assumes that a data set consists of several clusters with different distributions. All variables in the data set are modelled by the mixture of these distributions. The model-based clustering assumes a set of n observations with p-dimensions, such that an observed random sample is expressed as $x = \left(x_1^T, \ldots, x_n^T\right)^T$ [23]. The probability density function of finite mixture distributions are as follows;

$$f\left(x_j; \Psi\right) = \sum\nolimits_{i=1}^{g} \pi_i f_i\left(x_j; \theta_i\right) \tag{1}$$

where $f_i\left(x_j; \theta_i\right)$ are probability density functions of the components and π_i indicates the mixing weight (volume of clusters in the mixture model) in cases of $0 < \pi_i < 1$ and $\sum\nolimits_{i=1}^{g} \pi_i = 1 (i = 1, \ldots, g)$. The parameter vector $\Psi = (\pi, \theta)$ contains all of the parameters of the mixture models. Here, $\theta = \left(\theta_1, \ldots, \theta_g\right)$ denotes unknown parameters of the probability density function of the ith components in the mixture models. In Equation (1), the number of components or clusters is represented by g.

The mixture density function of the multivariate normal distribution is given as;

$$f\left(x_j;\Psi\right)=\sum\nolimits_{i=1}^{g}\pi_i\Phi_i\left(x_j;\mu_i,\Sigma_i\right) \tag{2}$$

where $\Phi_i\left(x_j;\mu_i,\Sigma_i\right)$ are assumed to be multivariate Gaussian densities of the form

$$\Phi_i\left(x_j;\mu_i,\Sigma_i\right)=\frac{1}{(2\pi)^{\frac{P}{2}}\left|\Sigma_i\right|^{\frac{1}{2}}}e^{\left\{-\frac{1}{2}(x_j-\mu_i)^T\Sigma_i^{-1}(x_j-\mu_i)\right\}} \tag{3}$$

where μ_i and Σ_i for $i=1,\ldots,k$ indicates the mean vector (locations of clusters in mixture model) and the covariance matrix (shapes of clusters in the mixture model), respectively [24]. All unknown parameters of the model are shown as $\Psi=\left(\pi_1,\ldots,\pi_{g-1},\xi^T\right)^T$, where ξ obtained from compound vectors of $\mu=\left(\mu_1,\ldots,\mu_g\right)$ and $\Sigma=\left(\Sigma_1,\ldots,\Sigma_g\right)$.

B. DETERMINATION OF NUMBER OF COMPONENTS IN VARIABLES

In finite mixture models, the correct determination of the number of components in each variable provides the correct calculation of the number of clusters of mixture models [25]. Some well-known clustering algorithms such as GMM, K-means, K-Nearest Neighborhood (K-NN), Support Vector Machines (SVM), Decision Trees (DT), etc., are used to determine the number of components in mixture models. In the proposed method, U-GMMs were used as the unsupervised clustering method to determine the number of components in the variables.

The number of components in U-GMMs corresponds to a component in each variable. U-GMM is shown as;

$$f\left(x;\theta\right)=\sum\nolimits_{i=1}^{g}\pi_i f_i\left(x_j;\mu_i;\sigma_i\right) \tag{4}$$

where $f\left(x;\theta\right)$ denotes density function of univariate Gaussian mixture distributions, g denotes the components of Gaussian mixture distributions, π_i denotes mixing weights, and $f_i\left(x_j;\mu_i;\sigma_i\right)$ denotes component probability density function. The component probability density function is shown as;

$$f_i\left(x_j;\mu_i;\sigma_i\right)=\frac{1}{\sqrt{2\pi}\sigma_i}exp\left\{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2\right\} \tag{5}$$

where μ_i denotes the mean and σ_i denotes the standard deviations of Gaussian distribution. Log-likelihood (logL) and BIC [26] values obtained from U-GMMs are used to determine the components in the variables. Expectation and Maximization (EM) algorithms are used to estimate the parameters of π, μ and, σ in U-GMMs. Parameters are estimated with the EM algorithm to determine the optimum component numbers in mixture models. The likelihood value is calculated by using estimated parameters. The BIC value is calculated depending on the likelihood. The numbers of components in each variable are determined according to the information criteria. The mixing weights and covariance matrices in the mixture model are indirectly affected by the number of observations in the components. The covariance matrix structure for multivariate GMMs corresponding to the clusters of components of the GMMs in the grid structure is shown as follows;

$$\Sigma_i=\begin{bmatrix}\sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 & \cdots & \rho_{1p}\sigma_1\sigma_p\\ \rho_{2,1}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2p}\sigma_2\sigma_p\\ \vdots & & \ddots & \vdots\\ \rho_{p,1}\sigma_p\sigma_1 & \rho_{n,2}\sigma_p\sigma_2 & \cdots & \sigma_p^2\end{bmatrix} \tag{6}$$

for $i=1,\ldots,g$, where g represents the number of components and p represents the number of dimensions of the data set. $\sigma_1,\ldots,\sigma_p$ represent standard deviations of p-dimensions.

Correlations between components of the variables are defined as $\rho_{1,2}=Corr\left(X_1,X_2\right),\ldots,\rho_{p,p-1}=Corr\left(X_p,X_{p-1}\right)$.

This type of covariance matrix is used due to the existence of different sizes of components in variables. The geometric standard spectral decomposition of a covariance matrix can be interpreted as follows;

$$\Sigma_k=\lambda_k D_k A_k D_k^T \tag{7}$$

where the scalar constant λ_k denotes volume, the orthogonal matrix of eigenvectors D_k denotes orientation, and the diagonal matrix A_k denotes the shape of the covariance matrix, respectively, with the form $Diag\left\{\alpha_{1k},\ldots,\alpha_{pk}\right\}$ where $\alpha_{1k}\geq\alpha_{2k}\geq\ldots\geq\alpha_{pk}\geq 0$. Utilising this decomposition of the covariance matrix Σ_k, geometric characteristics of the distributions can be imposed and a suitable model can be generated, where, k and p denote the number of components and dimensions of mixture models, respectively. For more detailed descriptions of parsimonious covariance matrices family and mixture model types, see [27].

The mixture weights (π_i) of the models obtained from the covariance matrix shown in (7) are calculated from the number of elements in the component. Probability weights in the mixture model are the most important parameters for determining the number and structure of the cluster. Unsupervised clustering algorithms such as GMM and k-means can be used to assign observations to the components from U-GMM. The mean of the observations is used to determine the cluster centres of the mixture models. While determining the components to which the observations belong, their distance from the cluster centre is used. The k-means assignment algorithm is used to assign different numbers of observations to the components according to their distance from the cluster centres. In this study, " mclust [28], clustvarsel [29], varselLCM [30], selvarMix [31] and vscc [32], " packages of R software were used to determine the number of components on model-based clustering in the synthetic-1, synthetic-2, Iris, and LSI data sets. MATLAB® 2018b software was used to determine the model selection method. The best model can be obtained by using statistical information criteria for model selection after fitting the models to the data set with the Likelihood estimation method. Variable selection and assignment of observations to components are shown in Algorithm 1 as follows.

Algorithm 1 Variable/Feature Selection and Assignment of Observations to Components

Input: $\aleph^d \to \mathbb{R}$; k_i: components of X_i (variables), d: dimension of dataset

Output: Determination of Homogeneous ($k_i = 1$), and Heterogeneous ($k_i \geq 2$) variables. Assigning all observations to variable components with k-means algorithm.

1. For k=1,…,4, U-GMM $f(x;\theta) = \sum_{i=1}^{g} \pi_i f_i \left(x_j; \mu_i; \sigma_i\right)$ (4) is applied to each variable X_i.

2. In U-GMMs, the number of components (k_i) is determined based on log-L and BIC values.

3. Homogeneous variants ($k_i = 1$) are eliminated, the algorithm continues with heterogeneous variables ($k_i \geq 2$).

4. The k_i component numbers in the variables are considered preliminary information for the k-means algorithm. According to the known k numbers, the observations are assigned to the components they belong to.

C. MIXTURE MODEL SOFT COMPUTING METHOD BASED ON THE COMPONENTS OF VARIABLE

The minimum and maximum numbers of clusters in the mixture model are denoted as C$_{min}$ and C$_{max}$ and are defined as follows;

$$C_{min} = max\left\{k_s\right\} \tag{8}$$

$$C_{max} = \prod_{s=1}^{p} k_s, \quad s = 1, \ldots, p \tag{9}$$

where p represents the dimension of data and k_s represents components of the heterogeneous variables. The number of GMM based on the components in heterogeneous variables represented by M$_{Total}$ can be calculated as follows;

$$\mathrm{M}_{Total} = 2^{\prod_{s=1}^{p} k_s} - 1 \tag{10}$$

where the term "-1" in Equation (10) represents the null model.

Theorem 1: The number of cluster ways to form k clusters in variables with m components where $k \leq m$ and $k \neq 0$ is given by

$$\mathrm{s}(m, k) = \sum_{i=0}^{m} (-1)^i \binom{m}{i} (k-i)^m \tag{11}$$

Proof: Odell and Duran (1974, p.26) [33].

Definition 1: A function is defined as $f{:}D(f) \to R(f)$ between the cluster centres corresponding to the components of the variables, and the AGMMs obtained by the orientations of the cluster centres. This function defines a "one-to-one and onto" relationship between the components of variables, and the number of AGMMs. Where, $D(f) = \left[max\left\{k_s\right\}, \prod_{s=1}^{p} k_s\right], \forall k_s \in \left\{C_{min}, C_{max}\right\}$ is the domain of the function corresponding to the number of components in the variables. The number of AGMMs are obtained as the range set of the function, $R(f)$, as follows;

$$R(f) = \left[\sum_{i_1,\ldots i_k=0}^{j_1,\ldots,j_k} (-1)^{\sum_{r=1}^{k} i_r} \binom{j_1}{i_1}\binom{j_2}{i_2}\cdots\binom{j_n}{i_n} \times \binom{(j_1-i_1)(j_2-i_2)\cdots(j_k-i_k)}{c}\right] \tag{12}$$

where j_k and i_k correspond to the components in multivariable data for p dimensions and clusters in mixture models, respectively. The number c indicates the minimum and maximum clusters that can occur in mixture models.

D. STRUCTURE OF GRID-BASED POSSIBLE MIXTURE MODELS

In this study, a novel clustering method is proposed to calculate the number of mixture models in the grid structure based on the components in heterogeneous variables according to the soft computing method. Mean, covariance matrix, and probability weights were calculated from the population for each component of the variables that make up mixture models in multivariate data. Each model that corresponds to the appropriate model is defined as;

$$f^u = \left(x, \mu^u, \Sigma^u\right) = \sum_{i=1}^{g} \pi_i^u f_i \left(x, \mu_i^u, \Sigma_i^u\right) \tag{13}$$

for $u = 1, \ldots, k$, where $\pi^u = \frac{\pi_i}{\sum\limits_{s=1}^{g} \pi_s}$ are mixing proportions,

$$\mu_i^u = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

are mean vectors, and

$$\Sigma_i^u = \begin{bmatrix} \sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 & \cdots & \rho_{1p}\sigma_1\sigma_p \\ \rho_{2,1}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2p}\sigma_2\sigma_p \\ \vdots & & \ddots & \vdots \\ \rho_{p,1}\sigma_p\sigma_1 & \rho_{p,2}\sigma_p\sigma_2 & \cdots & \sigma_p^2 \end{bmatrix}$$

are variance-covariance matrices for component Gaussian density functions for $i = 1, \ldots, g$. Each possible mixture model corresponds to a vector representation. The vector representation of the model, for example, "10110100", corresponds to each mixture model in determining the mixture models in the grid structure according to the soft computing method. The number of clusters in the models is shown in the structure blocks in the GMM with $o(H)$, which is called the degree of subset, as follows;

$$o(H) = \sum_{i=1}^{g} c_i, \tag{14}$$

where c_i represents "0" and "1" in elements of vector arrays.

Another structure block corresponding to the orientation of the clusters in the GMM is the length of the subset and is denoted as $\delta(H)$. The distance between the specific first

```
**Algorithm 2** Grid Structure Mixture Models and Vector Representations
**Input:** The variables and components ($k_i \in X_i$) that will create the dimensions of the models in the grid structure are placed in the grid-based model.
**Output:** Grid structure mixture models and vector representations
**1.** For $\forall k_i \in X_i$, the variables and components selected in the data set are determined in the grid-based model.
**2.** With $C_{min} = max\{k_i\}$ and $C_{max} = \prod_{i=1}^{p} k_i$, min and max ranges are obtained for the number of clusters in grid structured mixture models.
**3.** The number of mixture models in grid structure is calculated with $M_{Total} = 2^{\prod_{s=1}^{p} k_s} - 1$.
**4.** Vector representations consisting of "0" and "1" digits are created for mixture models in the grid structure.
**5.** The number of components in each mixture model is obtained by $o(H) = \sum_{i=1}^{g} k_i$.
```

cluster and the positions of the last cluster in the subset are shown as;

$$\delta(H) = arg_{i,j}\left\|c_i - c_j\right\|, \quad i, j = 1, \ldots, C_{max} \tag{15}$$

where c_i and c_j represent the first and the second cluster centre of vector, respectively.

The number of clusters in the vector representation of mixture models is equal to the number obtained by the degree of the subset ($o(H)$) in GMM. Besides, the location of the "1" in the vector sequence of the mixture model is shown with the location determined by the length of the subset ($\delta(H)$) in the block structure of GMMs.

The introduced concept of vector representation is that the structure blocks represent the clusters in the GMM model. While the cluster corresponding to each component is represented by "1", the null cluster in components is indicated by "0". A vector representation corresponds to each appropriate model obtained from mixture models. The creation of grid structured mixture models and their vector representations are shown in Algorithm 2 as follows.

E. INFORMATION CRITERIA FOR APPROPRIATE MIXTURE MODELS IN GRID STRUCTURE

logL functions of GMMs in grid structure are calculated as follows;

$$logL(\pi, \mu, \Sigma) = \sum_{j=1}^{n} log\left(\sum_{i=1}^{k} \pi_i f_i\left(x_j, \mu_i, \Sigma_i\right)\right) \tag{16}$$

BIC is calculated as follows depending on the logL function, the number of independent parameters d, and the number of observations n;

$$BIC = 2logL(\pi, \mu, \Sigma) - dlogn \tag{17}$$

where n and d represent the number of observations and the number of free parameters in the model, respectively. The model that maximises BIC is selected.

Based on the variables of the data set, the mixture clustering algorithm, which determines the number of clusters appropriate for the data structure from the components and the structure of the clusters, was developed by applying several methods step by step as stated in the sections above. The determination of appropriate mixture models by MMSCM and the best model selection are shown in Algorithm 3 as follows.

```
**Algorithm 3** Determination of AGMMs With MMSCM, Parameter Estimation and Best Model Selection
**Input:** Grid structure mixture models and their vector representations
**Output:** Parameter estimations of AGMMs, and selection of the best mixed model.
**1.** Among $M_{Total}$ mixture models, AGMMs are determined by (12) based on MMSCM according to the locations of variable components.
**2.** Parameter estimation of mixture models is obtained by EM algorithm.
**3.** The information criteria of the mixed models (16) and (17) are calculated using vector sequences for $f(x_j; \Psi) = \sum_{i=1}^{g} \pi_i \Phi_i(x_j; \mu_i, \Sigma_i)$ and AGMMs.
**4.** The best model is selected from AGGMs based on information criteria.
**5.** The algorithm terminates.
```

In summary, U-GMMs were used to determine the number of components in the variables. While the k-means algorithm was used to assign observations to the components in the variables, the soft computing method in the resulting models was solved with the GMM. In the last step of the clustering algorithm, the best model was obtained by using the vector representation of GMMs for model-based clustering. In Figure 1, the proposed approach is described to determine the number of clusters of a data set in the mixture model soft computing-based clustering.

The proposed method will be applied to the synthetic-1, synthetic-2, Iris [34] and 3D LSI data sets [35].

F. MIXTURE MODEL SOFT COMPUTING METHODS FOR OPTIMISATION

In this section, an effective optimization algorithm for MMSCM is introduced by determining the objective function. The proof of convergence of the algorithm is also presented. For time complexity, we define information complexity. The objective function basically consists of two parts: variable selection with univariate normal mixture models and the number of estimations with soft computing method.

A computer consists of Intel(R) Core(TM) i7-8700 CPU@3.20GHz and 8-GB RAM, Intel UHD Graphics 630 running on Windows 10 with a 64-bit R XX compiler was

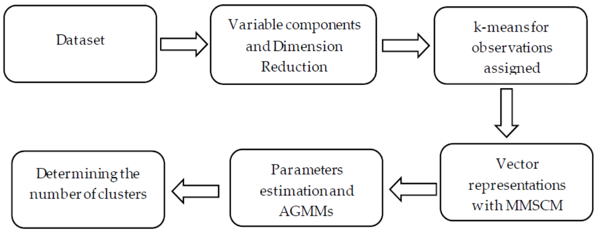
Dataset
Variable components and Dimension Reduction
k-means for observations assigned
Vector representations with MMSCM
Parameters estimation and AGMMs
Determining the number of clusters

FIGURE 1. The proposed approach for determining the number of clusters.

used for this study. Each step of the study was done in the R sotfware development environment.

1) OBJECTIVE FUNCTION

The aforementioned univariate mixed models and the variable selection method reduce the variables that have no effect on clustering. Therefore, it preserves the information on the variables in the data set, the geometric structure and volume of the clusters to be formed. In the reduced data, information complexity is minimal since there are no iterative processes when calculating the number of clusters and mixed models of grid structured models based on variable components.

The objective of the proposed method is to determine the best mixture model among the reduced total mixture models with the minimum number of variables. To achieve that, the proposed method minimizes not only the number of variables but also number of total mixture models.

2) CONVERGENCE ANALYSIS

This section discusses the convergence of GMM with MMSCM. The purpose of the proposed clustering is to separate the $\mathfrak{R}^{dxn}$ data set into clusters different from each other as $2 \leq k \leq n$.

The problem can be solved using a two-stage deterministic method to identify components in a data set with multivariate normal mixture distributions. In the first step, the dimension reduction takes place by selecting the variable ($dxn \geq lxn$). In the last step, the mixed models obtained from the variables are updated to minimize the problem from convergence to the best model selection.

3) TIME COMPLEXITY

In this section, the time complexity is analyzed to present the effectiveness of the MMSCM algorithm. In MMSCM, run time is mainly spent in U-GMM based variable selections, determining vector representations of AGMMs, and calculating information criteria for each grid based mixture models.

n: number of observations, d: number of variables (dimension), and m: the number of components (clusters) in the model, while the information complexity of the Em algorithm is $\mathcal{O}\left(nd^2\right)$ [36] in multivariate models, for variable selection in univariate models information complexity is $\mathcal{O}\left(nd\right)$.The information complexity of calculating model numbers to obtain vector representations of AGMMs is $\mathcal{O}\left(md\right)$. Finally, the information complexity is $\mathcal{O}\left(d^2n + dmn\right)$ to obtain the information criteria for each AGMMs.

TABLE 1. log-L and BIC values for all variables in the data set.

Datasets	Samples	Variables	Clusters
Synthetic-1	600	3	4
Synthetic-1	400	15	3
Iris	150	4	3
LSI	39600	3	5

III. RESULTS AND DISCUSSION

A. APPLICATION OF THE PROPOSED MIXTURE MODEL CLUSTERING ON THE SYNTHETIC DATA SETS

In this section, the proposed method for the estimation of the number of clusters is applied on the synthetic-1 data set produced to explain the simple and clear steps of the study. To measure the performance of the proposed method, it was applied on Synthetic-2 dataset with more variables (15 variables). Results, which are gathered from the analysis of Synthetic-1 and Synthetic-2 data sets, were compared with the results of mclust, clustvarsel, varselLCM, selvarMix and vscc methods. In Table 1, the number of variables/features, number of observations and number of components/clusters of the data sets used are given.

1) APPLICATION OF THE PROPOSED CLUSTER ESTIMATION METHOD ON SYNTHETIC-1 DATASET

In this section, the principles of the proposed MMSCM are explained on the synthetic-1 data set. In order to determine the number of clusters with univariate approaches, a multivariate synthetic-1 data set was produced by simulation.

The synthetic-1 data set was generated from the mixture of Gaussian distributions using mean vectors and covariance matrices, with three variables and four clusters. It is designed to have 1, 2, and 3 components in the variables, respectively, to demonstrate the availability of different numbers of components in the variables and a different number of observations in each component. While creating the synthetic-1 data set, the parameters that make up the variables are given as follows:

The mean and standard deviation values for variable X_1 are $\mu_1 = [29.55]$ and $\sigma_1 = [4.97]$. Mean and covariance matrices for variables X_2 *and* X_3 are

$$\mu_2 = \begin{bmatrix} 14.43 \\ 44.69 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 25.12 & 1.104 \\ 1.104 & 26.43 \end{bmatrix}$$

and

$$\mu_3 = \begin{bmatrix} 9.64 \\ 50.29 \\ 77.19 \end{bmatrix}, \quad \Sigma_3 = \begin{bmatrix} 23.294 & 1.358 & -1.266 \\ 1.358 & 24.94 & 0.335 \\ -1.266 & 0.335 & 23.227 \end{bmatrix}$$

respectively.

According to the information criteria obtained from the U-GMMs, there are 1, 2, and 3 components in variables $X_1, X_2,$ *and* X_3 respectively. logL and BIC values obtained from U-GMMs to determine the components in variables X_1, X_2 *and* X_3 are given in Table 2 below.

TABLE 2. log-L and BIC values for all variables in the data set.

Variable	X_1		X_2		X_3	
#Comp.	logL	BIC	logL	BIC	logL	BIC
1	-1813.60	-3640.10	-2510.70	-5034.20	-2833.10	-5679.00
2	-1813.60	-3659.20	-2236.50	-4505.00	-2607.90	-5247.70
3	-1813.40	-3678.00	-2236.30	-4523.90	-2454.00	-4959.10
4	-1813.40	-3697.10	-2234.60	-4539.70	-2453.10	-4977.50

TABLE 3. Components and size for X_1, X_2 and X_3 variables of the synthetic-1 data set.

Variables	X_1	X_2		X_3		
Components	X_1	X_{21}	X_{22}	X_{31}	X_{32}	X_{33}
#observations	600	329	271	179	231	190
Total	600	600		600		

The components in the synthetic-1 data set and the number of observations per component are given in Table 3.

The cluster number ranges, $C_{min} = max\{k_s\} = max\{1, 2, 3\} = 3$ and $C_{max} = \prod_{s=1}^{p} k_s = 1.2.3 = 6$, of the mixture model to be created in the grid structure were calculated according to the components in the variables.

Thus, the minimum and maximum clusters in the synthetic-1 data set resulted as 3 and 6 respectively. Variable components and cluster centres are illustrated in Figure 2.

In the synthetic-1 data set, the variable X_1 is called as "redundant variable" because it has a homogeneous structure, and the variable selection is made so that the reduced data set consists of variables X_2 *and* X_3. Total mixture models for clusters obtained from variable components of the synthetic-1 data set. M_{Total} was computed as $M_{Total} = 2^6 - 1 = 63$ for $C_{max} = 6$. The cluster numbers, the number of total models, and the number of appropriate models of the synthetic-1 data are shown in Table 4.

For the three-dimensional synthetic-1 dataset, the mean vector and covariance matrix structure are in the form of

$$\mu_i^u = \begin{bmatrix} \mu_1^u \\ \mu_2^u \\ \mu_3^u \end{bmatrix}$$

and

$$\Sigma_i^u = \begin{bmatrix} \left(\sigma_1^u\right)^2 & \rho_{1,2}^u \sigma_1^u \sigma_2^u & \rho_{1,2}^u \sigma_1^u \sigma_3^u \\ \rho_{2,1}^u \sigma_2^u \sigma_1^u & \left(\sigma_2^u\right)^2 & \rho_{2,3}^u \sigma_2^u \sigma_3^u \\ \rho_{3,1}^u \sigma_3^u \sigma_1^u & \rho_{3,2}^u \sigma_3^u \sigma_2^u & \left(\sigma_3^u\right)^2 \end{bmatrix},$$

where $i = 1, \ldots, k$ respectively. Here σ_1, σ_2 *and* σ_3 corresponds to the components of X_1, X_2 *and* X_3, respectively.

The vector representation of appropriate mixture models corresponds to 25 appropriate mixture models. The general form of the mixture model having vector representation by k component and number of components were expressed as follows:

$$f^u = \left(x, \mu^u, \Sigma^u\right) = \sum_{i=1}^{g} \pi_i^u f_i\left(x, \mu_i^u, \Sigma_i^u\right), \quad \text{for } u = 1, 2, 3$$

where $\pi^u = \frac{\pi_i}{\sum\limits_{s=1}^{3,4,5,6} \pi_s}$ are mixing weights, for $i = 1, \ldots, k$ and $k = 1, \ldots, 25$.

After dimension reduction, the corresponding cluster centres of the components in variables X_2 *and* X_3 and grid structures of appropriate models were obtained.

Component density function is the probability density function for bivariate Gaussian distribution.

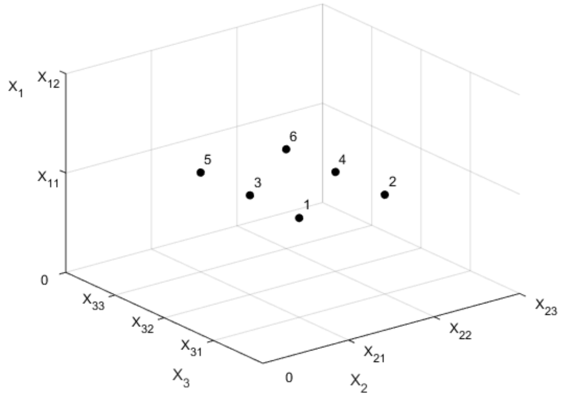

FIGURE 2. Variable components in X_1, X_2 and X_3 forms and clusters in synthetic-1 data set.

TABLE 4. The number of clusters, Alternatives mixture models, and possible mixture models.

#Clusters	#Alternatives GMMs	#AGMMs
1	6	0
2	15	0
3	20	6
4	15	12
5	6	6
6	1	1
Total	63	25

TABLE 5. logL and BIC for the best mixture model.

#AGMMs	logL	BIC	Vector Rep
1	-7363	-14726	101001
2	-10216	-20431	100110
3	-14305	-28790	011010
4	-13580	-27159	010110
5	-10546	-21092	011001
6	-8486	-16972	100101
7	-11211	-22422	111010
8	-8542	-17086	101110
9	-7334.7	-15069	101011
10	-8654	-17309	110101
11	-9500	-19000	011101
12	-11425	-22851	010111
13	-6274	-12549	101101
14	-7488	-14976	011011
15	-8118	-16237	110110
16	-10381	-20763	100111
17	-11770	-23539	011110
18	-10720	-21441	111001
19	-8664	-17329	011111
20	-6397	-12795	101111
21	-7622	-15245	110111
22	-8243	-16486	111011
23	-6402	-12804	111101
24	-9630	-19260	111110
25	-6508	-13016	111111

The logL and BIC of the best model obtained from the parameters calculated based on the components in the variables of the synthetic-1 data set are shown in Table 4.

The best mixture model that fits the 3D synthetic-1 data set, which is determined by model-based clustering among 25 appropriate models, was the four-component mixture model. Information criteria and vector representation of the model obtained from mixture density functions for determining the best model are shown in Table 5. According to the centres presented in Figure 2, the structure blocks of the best model was obtained as $o\,(H) = 4$ and $\delta\,(H) = \{1, 3, 4, 6\}$.

The results obtained from MMSCM are compared to the results of mclust and mclust based model selection methods for synthetic-1 data set in Table 6.

The illustration of BIC values and number of components for synthetic-1 data set is shown in Figure 3.

The GMM results were compared with the proposed MMSCM to estimate the number of clusters and the correct classification ratio on the synthetic-1 data set. According to the general CCR values, the recommended soft computing clustering method is approximately 15% more successful than mclust based methods. For the full data in the synthetic data set, the scatter plot obtained from mclust is shown in Figure 4.

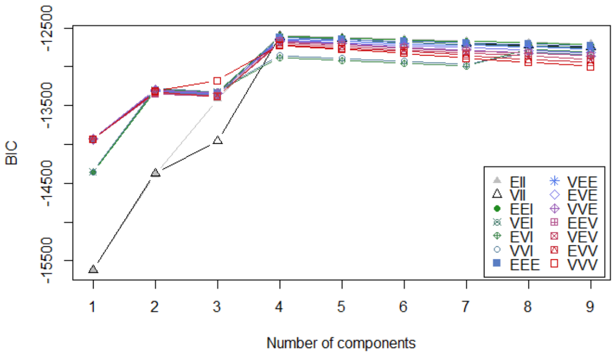

FIGURE 3. The number of components in multivariate the synthetic-1 data set according to different covariance structures.

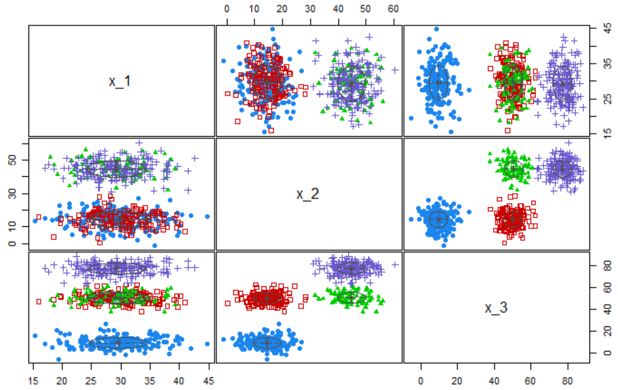

FIGURE 4. Scatter plot created by variables X_1, X_2 and X_3 in the data reduced by variable selection.

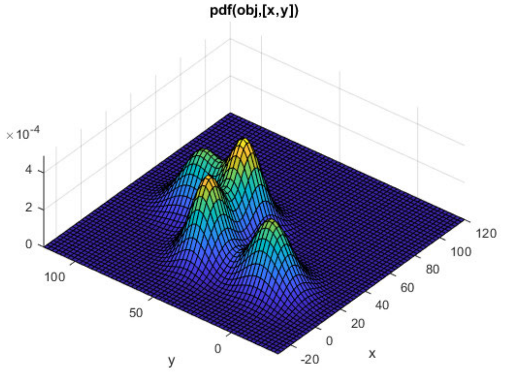

FIGURE 5. 3D surface graph of the best mixture model with four components.

The 3D surface graph of the model, which has "101101" vector representation with MMSCM, is illustrated in Figure 5.

According to the classification success values obtained from Table 6 for the Synthetic-1 dataset, MMSCM achieved

TABLE 6. Results of comparison MMSCM and mclust based methods on the synthetic-1 data set.

	log-L	BIC	#cluster	Variables	Cov. model	CCR%	ARI%
MMSCM	-6274	-12549	4	X2-X3	VVV	100	-
Mclust	-6245	-12592	4	X1-X2-X3	EII	85	78
Clustvarsel	-4433	-8942	4	X2-X3	EII	85	78
VarselLCM	-		4	X2-X3	-	84	78
selvarMix	-	-12622	4	X1-X2-X3	-	85	78
vscc	-	-1664	4	X2-X3	EEI	85	78

an average of 15% higher success than the mclust based methods. In the graph of the number of components in Figure 3, it was seen that the best covariance model was the full type.

2) NUMBER OF COMPONENT ESTIMATION AND BEST MODEL SELECTION FOR SYNTHETIC-2 DATA SET WITH MMSCM

The synthetic-2 data set was generated by simulation with the proposed MMSCM for dimension reduction and selection of the best model. In the data set with 15 variables, the 1st and 2nd variables were produced as 2 components and the other variables were homogeneous. The parameter values used while generating the variables in the data set are as follows;

The mean vector and covariance matrix parameter values for the variables X_1 and X_2 are

$$\mu_1 = \begin{bmatrix} 21.3 \\ 57.1 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} 84.38 & -1.67 \\ -1.67 & 119.88 \end{bmatrix}$$

and

$$\mu_2 = \begin{bmatrix} 21.4 \\ 61.1 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 92.42 & 5.39 \\ 5.39 & 80.81 \end{bmatrix},$$

respectively. The mean and standard deviation values for other homogeneous variables are

$$\mu_{3,4,\ldots,15} = \begin{bmatrix} 27.5, 32.3, 18.8, 23.5, 30.3, 56.8, 64.3, \\ 20.7, 44.3, 39.1, 36.2, 28.1, 50.1 \end{bmatrix}$$

and

$$\sigma_{3,4,\ldots,15} = \begin{bmatrix} 9.9, 9.6, 9.8, 10.03, 10.1, 10.08, 10.3, 10.29, \\ 10.06, 9.52, 10, 9.76, 9.68 \end{bmatrix}.$$

The component numbers of each variable in the 15-variable data set are revealed with U-GMM. Homogeneous and heterogeneous variables are determined and redundant variables are excluded for the number of cluster estimations. U-GMM values for 15 variables are shown in Table 7.

According to the logL and BIC values presented in Table 6, X_1 and X_2 are heterogeneous two-component variables in the 15-variable data set. Others are homogeneous variables with a single component. The number of observations assigned to the components of the variables is shown in Table 8.

The number of components was determined with U-GMM for the variables and dimension reduction was made by eliminating homogeneous variables. X_1 and X_2 are used as heterogeneous variables to estimate the number of clusters, and other homogeneous variables are eliminated as redundant.

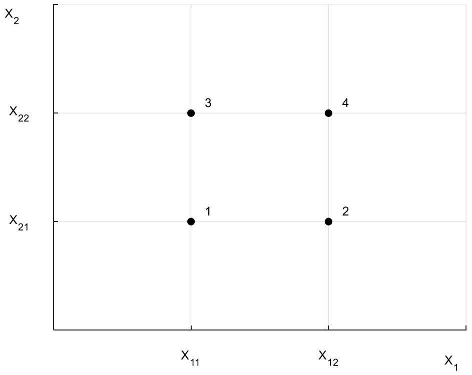

FIGURE 6. Components of variables X_1 and X_2 and their corresponding cluster centres.

$max\{2, 2\} = 2$ and $C_{max} = 2.2 = 4$ are obtained for alternative cluster numbers in the mixture model based on the component numbers of the variables. The components of the variables and possible cluster centres in the reduced data set are shown in Figure 6.

Mixture models obtained from the components of the variables X_1 and X_2 in the reduced data set and the logL, BIC values, and vector sequences of the models are shown in Table 9.

The model with 3 components and "1110" vector array representation according to logL and BIC values was determined as the best mixture model among the AGMMs.

The results obtained from MMSCM are compared to the results of mclust and mclust based model selection methods for synthetic-2 data set in Table 10.

Although the classification successes are the same, mclust and vscc methods could not determine the redundant variables.

The number of components and BIC value graph obtained through mclust according to different covariance types are shown in Figure 7.

The comparison of the performance of GMM with the correct classification rate of the synthetic-2 data set according to the clusters and locations obtained by MMSCM is given in Table 9. The surface plot of the best mixture model obtained by MMSCM is shown in Figure 8.

TABLE 7. The number of components, logL and BIC values in variables with U-GMM for 15-variable data set.

Var	# comp	logL	BIC	Var	# comp	logL	BIC
X_1	1	-1753.4	3510.8	X_9	1	-1500.6	3013.1
	2	-1704.7	3439.4		2	-1500.5	3030.9
	3	-1702.3	3452.5		3	-1500.5	3048.9
X_2	1	-1775.8	3563.7	X_{10}	1	-1499.8	3011.7
	2	-1693	3416		2	-1499.9	3029.7
	3	-1693	3433.9		3	-1499.8	3047.6
X_3	1	2975.9	2983.8	X_{11}	1	-1490.6	2993.2
	2	2981.5	3001.5		2	-1490.6	3011.2
	3	2986.7	3018.6		3	-1490.3	3028.6
X_4	1	-1471.2	2956.3	X_{12}	1	-1468.8	2949.5
	2	-1471.7	2973.4		2	-1466.3	2962.6
	3	-1471.8	2991.5		3	-1460	2980
X_5	1	-1482.6	2977.3	X_{13}	1	-1488.2	2988.4
	2	-1482.6	2995.3		2	-1487.1	3004.2
	3	-1480	3009		3	-1487	3021.9
X_6	1	2982.8	2990.7	X_{14}	1	-1478.6	2969.2
	2	2986.9	3006.9		2	-1478.6	2987.2
	3	2989.2	3021.2		3	-1475.4	2997
X_7	1	-1495	3001.9	X_{15}	1	-1475.3	2962.5
	2	-1495	3019.9		2	-1476.6	2979.2
	3	-1494	3035.9		3	-1471.5	2991.
X_8	1	-1491.4	2994.8				
	2	-1490.7	3011.3				
	3	-1490.4	3028.7				

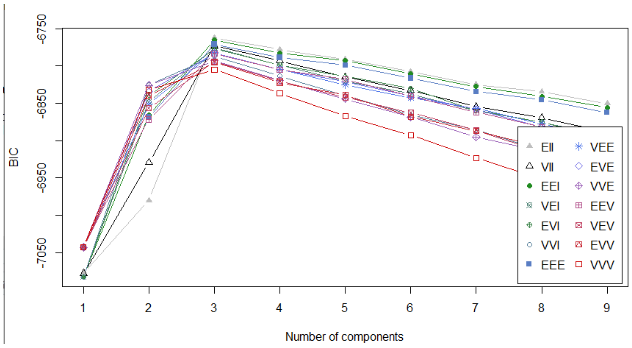

FIGURE 7. The number of components and BIC values according to different covariance types with mclust.

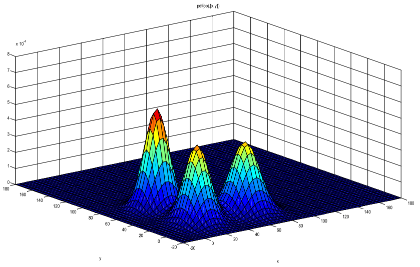

FIGURE 8. Cluster centres and surface plot corresponding to components of variables X_1 and X_2 in the best mixture model.

According to the results in Table 10 for the Synthetic-2 data set, the success of the proposed method and the mclust based methods are the same. While MMSCM used VVV full covariance for the number of component estimation, other methods used EVI covariance type as seen in Figure 7.

B. TESTING OF THE PROPOSED MIXTURE MODEL CLUSTERING FOR THE REAL DATA SETS

While the principles of the proposed method for model-based clustering in the grid structure on the synthetic-1 and synthetic-2 data sets were explained in the previous section, the clustering algorithm of the proposed method was applied on Iris and LSI real data sets one after another in this section.

TABLE 8. In the 15-variable data set, the number of observations falling on the components of variables X_1 and X_2 and the number of observations of other homogeneous variables.

Variables	X_1		X_2		$X_3, X_4, \ldots, X_{15}$
Components	$X_{1.1}$	$X_{1.2}$	$X_{2.1}$	$X_{2.2}$	-
# observations	120	280	150	250	400

1) APPLICATION OF THE PROPOSED CLUSTERING METHOD FOR THE IRIS DATA SET

The steps of the proposed method (MMSCM) is explained on the Iris data set (UCI Machine Learning Repository), which is widely used for clustering and classification analysis. Iris data set, presented by Fisher in 1936 and widely used in clustering, is a multivariate data set with three clusters (Setosa, Versicolor, and Virginica), four variables (Sepal length, Sepal width, petal length, and petal width), and 150 observations.

In the finite mixture models, it has always been a difficult problem to correctly determine the number of components in the variables. In the proposed method, the number of components of the heterogeneous variables in the Iris data set is determined with the U-GMM. logL and BIC values were obtained from U-GMMs to determine the components in variables X_1, X_2, X_3 *and* X_4 (Sepal length, Sepal width, petal length, and petal width) and are given in Table 11.

Applying U-GMMs to the 4 variables in the Iris data set show that there was no (homogeneous) component in the Sepal width variable (X_2) according to the findings obtained from the values in Table 6. Thus, the variable X_2 was determined as a ''redundant variable'' and a variable selection was made. While determining the number of clusters is based on the number of components in the variables, according to MMSCM assumptions, homogeneous variables do not affect the number of clusters and clustering. Variable selection was made by removing homogeneous variables, so dimension reduction was applied to the data. Sepal length (X_1), Petal length (X_3) and Petal width (X_4) variables is used to determine the clusters and also the number of clusters in the Iris data set.

Components of variables X_1, X_2, X_3 *and* X_4 of the Iris data set and assigned observations are given in Table 12.

The range is calculated as $C_{min} = max\{k_s\} = max\{2, 1, 2, 2\} = 2$, and $C_{max} = \prod_{s=1}^{p} k_s = 2\times1\times2\times2 = 8$ for the number of clusters, which are determined based on the components in heterogeneous variables. Thus, $C_{min} \leq k \leq C_{max}$ cluster intervals were obtained according to the component numbers $s = 1, 2$, $s = 1$, $s = 1, 2$, and $s = 1, 2$ for variables X_1, X_2, X_3, and X_4 in the Iris data set.

The grid structure model consisting of the components in variables X_1, X_3 *and* X_4, the clusters in the model, and the components forming each cluster are shown in Figure 9.

The multivariate data set is converted into mixture models in grid structure according to the number of components in the variables and the cluster centres that may occur. Components of variables that fit cluster centres are shown in Figure 9.

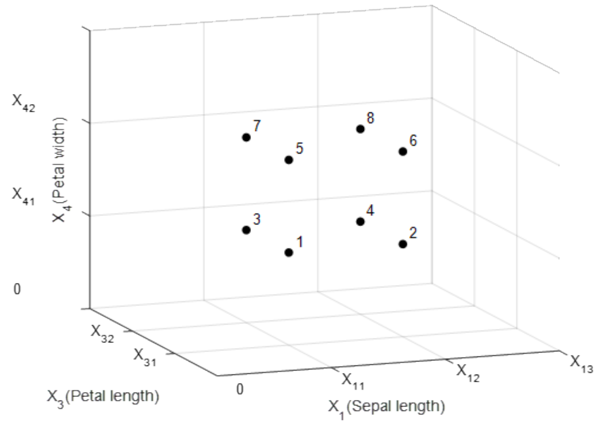

FIGURE 9. Components of variables X_1, X_3 and X_4 in reduced Iris data set.

2) SOFT COMPUTING FOR APPROPRIATE MIXTURE MODELS AMONG TOTAL MIXTURE MODELS

According to the maximum number of clusters from Equation (9) obtained from the components in the variables, the total number of mixture models from Equation (10) is obtained as $M_{Total} = 2^8 - 1 = 511$. The null model without any set is deducted from the calculation.

Under the assumptions of the proposed clustering algorithm, the number of models, which can occur according to the components of each variable, and mixture models that fit the assumptions from these models are calculated as shown in (12) with the MMSCM.

In the soft computing method, where $j_1 = 2, j_3 = 2$, *and* $j_4 = 2$ correspond to components variables X_1, X_3 *and* X_4, i_1, i_3, *and* i_4 indices are used to denote the number of clusters and c ranges from 2 to 8 to show the number of clusters in the mixture model.

The number of clusters, the number of models, and the number of possible mixture models are given in Table 13.

3) VECTOR REPRESENTATION AND INFORMATION CRITERIA OF BEST MIXTURE MODEL FITTING TO IRIS DATA IN GRID STRUCTURE

In this section, the soft computing method, which was expressed in Section 2, was proved that it could obtain the number and locations of the cluster in the Iris data set.

According to the information criteria, the best model was chosen from the appropriate mixture of multivariate Gaussian densities. The logL and BIC values of the best mixture models for the Iris data set are given in Table 14. The three cluster-centred mixture model was seen as the best model that fits the data among the mixture models in the Iris data set.

For the Iris data set, the number of components' plots for complete and reduced data obtained from mclust are shown in Figure 10 and Figure 11, respectively.

As it can be seen on the Figures 10 and 11, mclust based methods result 2 components for complete data. Since

TABLE 9. AGMM numbers, logL, BIC values and vector array representations according to the component numbers of the variables.

Models	#components	#AGMMs	logL	BIC	Vector Representation
1	2	2	-3979.01	7958.01	1001
2			-4544.6	9089.10	0110
3	3	4	-4091.83	8183.61	0111
4			-4103.81	8207.54	1101
5			-3466.5	6933.0	1110
6			-4050.22	8100.51	1011
7	4	1	-3550.91	7101.92	1111
Total	-	7	-	-	-

TABLE 10. Results of comparison MMSCM and mclust based methods on the synthetic-2 data set.

	log-L	BIC	#cluster	Variables	Cov. model	CCR%	ARI%
MMSCM	-3466	-6933	3	X1-X2	VVV	98	-
Mclust	-22665	-45618	3	X1,…,X15	EII	98	93
Clustvarsel	-3354	-6762	3	X1-X2	EII	98.25	94
VarselLCM	-	-	3	X1-X2	-	98	93
selvarMix	-	-45503	3	X1-X2	-	98	93
vscc	-	-17010	3	X1,…,X15	EEI	98	93

TABLE 11. -logL and BIC values for all variables in Iris data set.

Variables	Sepal length		Sepal width		Petal length		Peal width	
#Comp.	logL	BIC	logL	BIC	logL	BIC	logL	BIC
1	-184.03	-378.10	-86.99	-184.00	-297.51	-605.04	-171.79	-353.61
2	-177.95	-380.96	-85.28	-195.62	-200.53	-426.12	-106.47	-237.99
3	-177.28	-394.65	-83.71	-207.51	-200.53	-441.15	-101.82	-243.72
4	-173.76	-402.64	-83.18	-221.48	-199.95	-455.02	-100.87	-256.85

they cannot reduce the number of variables, real number of components could be achieved. On the other hand, the variables, which are determined by the proposed MMSCM, are applied on mclust based methods, the real number of component, 3, is obtained.

The 3D scatter plot of the best mixture model obtained from the proposed clustering method is shown in Figure 12.

The vector representation of the best model, which fits the data set among the mixture models from the AGMMs, is "10010010". The cluster centres in the best-mixture model are the 1st (Setosa), 4th (Versicolor) and 7th (Virginica) centres as shown in Figure 9. According to the centres in Figure 9, the structure blocks of the best model are obtained as $o(H) = 3$ and $\delta(H) = \{1, 4, 7\}$. 1st center $(X_{11}, X_{31}\ and\ X_{41})$, 4th centre $(X_{12}, X_{32}\ and\ X_{41})$ and 7th centre $(X_{11}, X_{32}\ and\ X_{42})$ are composed of components.

TABLE 12. Variable components and their observations for Sepal length (X_1), Sepal width (X_2), Petal length (X_1) and Petal width (X_1) of Iris data set.

Variable	X_1		X_2	X_3		X_4	
Components	X_{11}	X_{12}	X_2	X_{31}	X_{32}	X_{41}	X_{42}
#Observations	67	83	150	56	94	57	93
Total	150		150	150		150	

The results obtained from MMSCM are compared to the results of mclust and mclust based model selection methods for Iris data set in Table 15.

TABLE 13. The number of clusters, total mixture models, possible mixture models, and free parameters.

#Clusters	#Alternatives GMMs	#AGMMs
1	8	-
2	28	4
3	56	32
4	70	64
5	56	56
6	28	28
7	8	8
8	1	1
Total	255	193

TABLE 14. log L and BIC values for the best mixture model for Iris data set.

#Component	log L	BIC	Vector Rep.
3	-175.93	-499.12	10010010

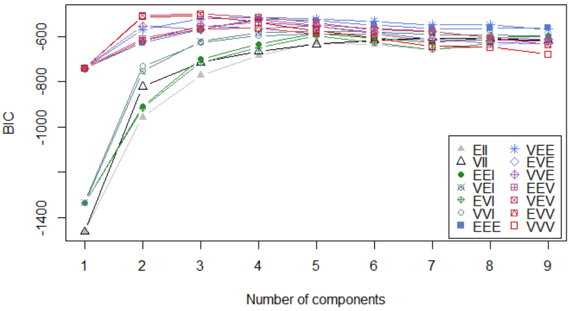

FIGURE 10. The number of components (2) and covariance types for complete data with 4 variables.

For the Iris dataset, MMSCM's variable selection and component count estimation is 22% more successful than other methods. While Mclust-based methods could not select variables, the estimation of the number of clusters was wrong and low success rate was obtained. However, when the variable selection determined by MMSCM was applied to the compared methods, they were able to correctly estimate the number of clusters.

According to the CCRs calculated for the mclust based methods and MMSCM and shown in Table 14, a higher success rate with 32%, in other words, better model fit was achieved as a result of components determined by MMSCM and dimension reduction, while mclust based methods directed a wrong number of components as 2.

4) APPLICATION OF THE PROPOSED MIXTURE MODEL CLUSTERING FOR THE LANDSAT SATELLITE IMAGE DATA SET

The proposed clustering method was applied to the remote sensed LSI data. Pixel values in the 3rd, 4th, and 5th bands were used among seven variables of LSI data. There are 5 components in the three-variable LSI dataset: Wheat, Potato, Vegetable Garden, Citrus and Bare Soil [24]. The components of each variable in the LSI data set corresponded to those in the U-GMMs are given in Table 16.

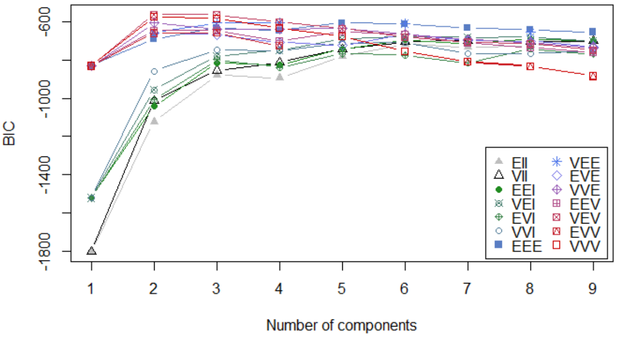

FIGURE 11. The number of components (3) and covariance types for reduced data with 3 variables after the variable selection.

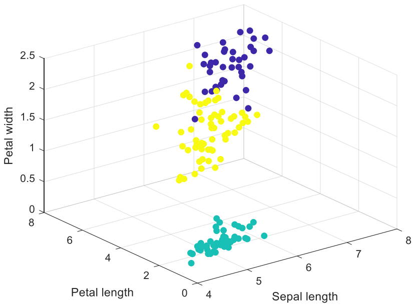

FIGURE 12. The 3D scatter plot of the best mixture model fitting the Iris data set.

There were three components in variables X_1, X_2 *and* X_3 according to the information criteria obtained from the U-GMMs. The components in variables of the LSI data set and the number of observations per component are given in Table 17.

The cluster number ranges, $C_{min} = max\{3, 3, 3\} = 3$ and $C_{max} = k_1 k_2 k_3 = 3.3.3 = 27$, of the mixture model to be created in the grid structure were calculated according to the components in the variables. Thus, the minimum and maximum clusters in the LSI data set were 3 and 27, respectively. Variable components and cluster centres are illustrated in Figure 13.

Total mixture models for clusters were obtained from variable components of LSI data set. M_{Total} can be computed as $M_{Total} = 2^{27} - 1 = 134217727$ for $C_{max} = 27$.

The clusters, grid-based total mixture models, and appropriate mixture models for components are given in Table 18.

The vector representation of appropriate mixture models corresponds to 131.964.460 appropriate models.

The best mixture model, which fits the 3D LSI data set and was determined by MMSCM among 131.964.460 appropriate models, is the five-component mixture model. Information criteria and vector representation of the model obtained from mixture density functions for determining the best model are shown in Table 19. The structure blocks of

TABLE 15. Results of comparison MMSCM and mclust based methods on the Iris data set.

	log-L	BIC	#cluster	Variables	Cov. model	CCR%	ARI%
MMSCM	-175	-499	3	X1-X3-X4	VVV	98	-
Mclust	-215	-561	2	X1,…,X4	VEV	66	56
Clustvarsel	-215	-561	2	X1,…,X4	VEV	66	56
VarselLCM	-	-	2	X1,…,X4	-	66	56
selvarMix	-	-574	2	X1,…,X4	-	66	56
vscc	-	-790	2	X1,…,X4	VVV	66	56

TABLE 16. The number of free parameters, logL and BIC values for all variables in the data set.

Variable		X_1		X_2		X_3	
#Comp.	#Param.	logL	BIC	logL	BIC	logL	BIC
1	9	-13947.0	-27897.0	-16418.0	-32838.0	-16664.0	-33331.0
2	19	-13767.0	-27538.0	-15912.0	-31829.0	-16169.0	-32344.0
3	19	-13567.0	-27242.0	-15858.0	-31724.0	-16126.0	-32261.0
4	39	-13617.0	-27246.0	-15857.0	-31725.0	-16134.0	-32339.0
5	49	-13617.0	-27249.0	-15857.0	-31728.0	-16138.0	-32352.0

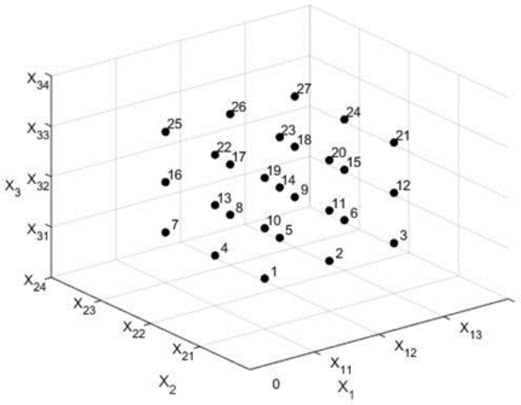

FIGURE 13. Variable components and cluster centres of LSI data set.

TABLE 17. Variable components and size for variables X_1, X_2 and X_3 of LSI data.

Var.	X_1			X_2			X_3		
Comp.	X_{11}	X_{12}	X_{13}	X_{21}	X_{22}	X_{23}	X_{31}	X_{32}	X_{33}
#observations	4.291	17.241	18.068	10.279	12.697	16.624	17.929	4.999	16.672
Total	39.600			39.600			39.600		

TABLE 18. Clusters, total mixture models, and appropriate mixture models.

# Cluster	# Alternatives Models	# Appropriate Models	# Cluster	# Alternatives Models	# Appropriate Models
1	27	0	15	17383860	17376516
2	351	0	16	13037895	13036518
3	2925	36	17	8436285	8436123
4	17550	1890	18	4686825	4686816
5	80730	24300	19	2220075	2220075
6	296010	153828	20	888030	888030
7	888030	623106	21	296010	296010
8	2220075	1839672	22	80730	80730
9	4686825	4255194	23	17550	17550
10	8436285	8044245	24	2925	2925
11	13037895	12751803	25	351	351
12	17383860	17216811	26	27	27
13	20058300	19981143	27	1	1
14	20058300	20030760	Total	134.217.728	131.964.460

the best model are shown in Figure 13 as $o(H) = 5$ and $\delta(H) = \{1, 6, 13, 18, 20\}$.

The results obtained from MMSCM are compared to the results of mclust and mclust based model selection methods for LSI data set in Table 19.

Runtime analysis of MMSCM, the four datasets mentioned above are listed in Table 21. It can be seen from Table 21 that the processing time is longer when the size of the dataset is large. Table 20 also shows the acceleration and high-quality clustering results of MMSCM.

TABLE 19. logL and BIC for the best mixture model.

logL	BIC	Vector Representation
-392715.4	-785660.7	10000100000001000010100000000

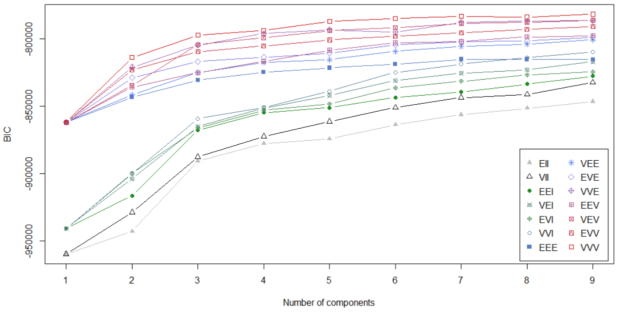

FIGURE 14. The number of components and BIC values according to different covariance types with mclust for LSI data set.

With respect to the values shown in Table 20, mclust based methods determine 9 components for VVV covariance types. Moreover, low CCR percentages for mclust based methods yield that the number of components are determined wrongly. The BIC values as well as number of components with respect to the different covariance types are presented in Figure 14.

It could observed on Figure 14 that the optimum number of components for VVV covariance type on LSI data set is 5. The results of the 5-component model obtained from mclust method are as follows: BIC = 786830.9, CCR = 62.9%, and ARI = 0.44. The scatter plot of the components and the positions of the components are shown in Figure 15.

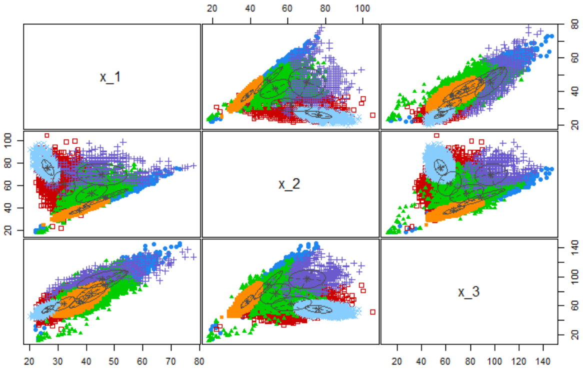

FIGURE 15. Cluster centres and scatter plots in three-variable and 5-component LSI data set.

The correct classification table for LSI data set is given in Table 22.

where W: Wheat, P: Potato, VG: Vegetable Garden, C: Citrus, BS: Bare Soil, RT: Row Total, CT: Column Total. The overall CCR is 0.95 for the proposed method from Table 22.

The image data of 5 clusters and their colours separated as a result of the clustering method proposed in the LSI data set are shown in Figure 16.

For LSI dataset, MMSCM achieved more than 50% success compared to mclust based methods. Existing methods of cluster estimation are quite unsuccessful in big data with a small number of variables but a large number of observations. With the univariate grid structure mixture models approach, MMSCM has shown a very high success not only in the number of clusters and CCR, but also in time complexity.

TABLE 20. Results of comparison MMSCM and mclust based methods on the LSI data set.

	log-L	BIC	#cluster	Variables	Cov. model	CCR%	ARI%
MMSCM	-392715	-785660	5	X1-X2-X3	VVV	95	-
Mclust	-390314	-781571	9	X1-X2-X3	VVV	46	35
Clustvarsel	-390317	-781579	9	X1-X2-X3	VVV	48	35
VarselLCM	-	-	9	X1-X2-X3	VVV	66	46
selvarMix	-	-781457	9	X1-X2-X3	VVV	51	32
vscc	-	-177803	9	X1-X2-X3	VVV	49	35

TABLE 21. Running time of different algorithms (seconds).

Datasets	mclust	clustvarsel	varselLCM	selvarMix	vscc	MMSCM
Synthetetic-1	1.8381	6	80.3940	9.3431	5.2669	3.254
Synthetetic-2	16.3263	35.1709	281.1840	32.7872	19.5008	17.575
Iris	0.7011	11.6139	7.7393	2.942	2.5911	2.372
LSI	2.8693	29885.04	11270.16	2001.078	716.634	14.336

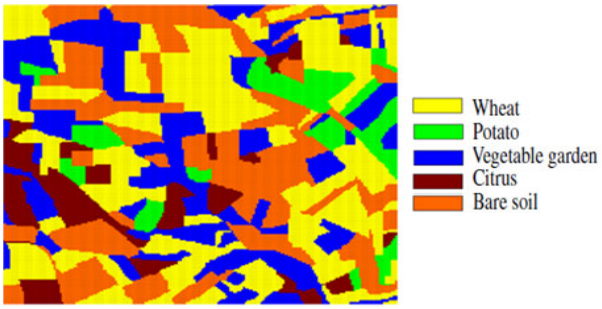

FIGURE 16. Five components and their colours separated as a result in the LSI data set.

TABLE 22. Results of the MMSCM with the correct classification ratio matrix of the LSI data set.

	W	P	VG	C	BS	RT	CCR for MMSCM
W	12989	235	33	0	273	13530	0.94
P	502	3724	79	24	312	4641	0.93
VG	0	27	11194	0	86	11307	0.98
C	188	0	0	1379	0	1567	0.98
BS	191	0	151	0	8213	8555	0.92
CT	13870	3986	11457	1403	8884	39600	0.95

IV. CONCLUSION

In this study, based on the components of heterogeneous variables in the data according to the mixture model soft computing method, a novel method was proposed for determining the clustering in GMM. The developed clustering algorithm was applied to the synthetic-1, synthetic-2, Iris and LSI data set, and it was observed that the variable cluster and number of clusters made an accurate clustering compared to the studies in the literature.

The dimension reduction method proposed by MMSCM has been important preliminary information in accurately estimating the number of components for mclust based methods.

In conclusion, the proposed MMSCM yields %15 better CCR results for synthetic-1 data set. Moreover, for synthetic-2 data set, MMSCM is better than mclust and vscc methods with respect to the variable selection. Furthermore, the proposed method gives better results for number of cluster estimation and variable selection as well as higher CCR regarding to Iris data set. Finally, for the LSI data set, MMSCM not only estimates number of cluster better, but also results a higher CCR value. It is clear that the proposed MMSCM performs better on the variable selection and the number of cluster estimation in comparison with the mclust based models.

The disadvantage of the proposed algorithm is that the number of elements of the search space increases exponentially when the number of variables increases and there are too many components in each variable. Determining AGMMs and obtaining vector representations among mixture models in the search space increases the time complexity. In order to overcome this problem, solutions for parallel computation can be studied in the proposed algorithm.

In addition, the proposed method gives faster and more accurate results than existing methods in terms of variable selection and cluster number estimation in big data.

In future studies, it is aimed to combine the mixture model soft computing method with deep learning methods in estimating the number of clusters of big data within the framework of variable selection. In addition, it is expected to combine the proposed method with multi-criteria decision-making methods and test it in different application areas. Besides, the variable selection, number of cluster estimation and classification success approaches of the proposed method should be developed as a package in R software.

REFERENCES

[1] G. J. McLachlan and S. Rathnayake, "On the number of components in a Gaussian mixture model," *Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery*, vol. 4, no. 5, pp. 341–355, 2014.

[2] H. Bozdogan, "Choosing the number of clusters, subset selection of variables, and outlier detection in the standard mixture-model cluster analysis," in *New Approaches in Classification and Data Analysis*. Berlin, Germany: Springer, 1994, pp. 169–177.

[3] T. L. J. Ng and T. B. Murphy, "Model-based clustering of count processes," *J. Classification*, vol. 38, no. 2, pp. 188–211, Jul. 2021.

[4] P. D. McNicholas, "Model-based clustering," *J. Classification*, vol. 33, no. 3, pp. 331–373, 2016.

[5] G. Celeux, S. Frühwirth-Schnatter, and C. P. Robert, "Model selection for mixture models-perspectives and strategies," in *Handbook of Mixture Analysis*. Boca Raton, FL, USA: CRC Press, 2019, pp. 117–154.

[6] X. Ye, J. Zhao, and Y. Chen, "A nonparametric model for multi-manifold clustering with mixture of Gaussians and graph consistency," *Entropy*, vol. 20, no. 11, p. 830, Oct. 2018.

[7] A. E. Raftery and N. Dean, "Variable selection for model-based clustering," *J. Amer. Stat. Assoc.*, vol. 101, no. 473, pp. 168–178, Mar. 2006.

[8] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *J. Amer. Stat. Assoc.*, vol. 97, no. 458, pp. 611–631, Jun. 2002.

[9] G. Galimberti and G. Soffritti, "Model-based methods to identify multiple cluster structures in a data set," *Comput. Statist. Data Anal.*, vol. 52, no. 1, pp. 520–536, Sep. 2007.

[10] G. Galimberti, A. Manisi, and G. Soffritti, "Modelling the role of variables in model-based cluster analysis," *Statist. Comput.*, vol. 28, no. 1, pp. 145–169, Jan. 2018.

[11] S. Akogul and M. Erisoglu, "An approach for determining the number of clusters in a model-based cluster analysis," *Entropy*, vol. 19, no. 9, p. 452, Aug. 2017.

[12] C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery, *Model-Based Clustering and Classification for Data Science: With Applications in R*, vol. 50. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[13] R. Shang, Y. Meng, W. Wang, F. Shang, and L. Jiao, "Local discriminative based sparse subspace learning for feature selection," *Pattern Recognit.*, vol. 92, pp. 219–230, Aug. 2019.

[14] R. Shang, W. Wang, R. Stolkin, and L. Jiao, "Non-negative spectral learning and sparse regression-based dual-graph regularized feature selection," *IEEE Trans. Cybern.*, vol. 48, no. 2, pp. 793–806, Feb. 2017.

[15] C. Maugis, G. Celeux, and M.-L. Martin-Magniette, "Variable selection for clustering with Gaussian mixture models," *Biometrics*, vol. 65, no. 3, pp. 701–709, 2009.

[16] M. Fop and T. B. Murphy, "Variable selection methods for model-based clustering," *Statist. Surveys*, vol. 12, pp. 18–65, Jan. 2018.

[17] K. Kumar Sharma, A. Seal, E. Herrera-Viedma, and O. Krejcar, "An enhanced spectral clustering algorithm with S-distance," *Symmetry*, vol. 13, no. 4, p. 596, Apr. 2021.

[18] K. K. Sharma and A. Seal, "Clustering analysis using an adaptive fused distance," *Eng. Appl. Artif. Intell.*, vol. 96, Nov. 2020, Art. no. 103928.

[19] K. K. Sharma, A. Seal, A. Yazidi, A. Selamat, and O. Krejcar, "Clustering uncertain data objects using Jeffreys-divergence and maximum bipartite matching based similarity measure," *IEEE Access*, vol. 9, pp. 79505–79519, 2021.

[20] K. K. Sharma and A. Seal, "Spectral embedded generalized mean based k-nearest neighbors clustering with S-distance," *Expert Syst. Appl.*, vol. 169, May 2021, Art. no. 114326.

[21] Y. Wei and P. D. McNicholas, "Mixture model averaging for clustering," *Adv. Data Anal. Classification*, vol. 9, no. 2, pp. 197–217, Jun. 2015, doi: 10.1007/s11634-014-0182-6.

[22] M. Gogebakan and H. Erol, "A new semi-supervised classification method based on mixture model clustering for classification of multispectral data," *J. Indian Soc. Remote Sens.*, vol. 46, no. 8, pp. 1323–1331, Aug. 2018.

[23] V. Melnykov and R. Maitra, "Finite mixture models and model-based clustering," *Stat. Survey*, vol. 4, pp. 80–116, Jan. 2010.

[24] C. Bouveyron and C. Brunet-Saumard, "Model-based clustering of high-dimensional data: A review," *Comput. Statist. Data Anal.*, vol. 71, pp. 52–78, Mar. 2014.

[25] J. D. Banfield and A. E. Raftery, "Model-based Gaussian and non-Gaussian clustering," *Biometrics*, vol. 49, no. 3, pp. 803–821, 1993.

[26] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.

[27] G. Celeux and G. Govaert, "Gaussian parsimonious clustering models," *Pattern Recognit.*, vol. 28, no. 5, pp. 781–793, May 1995.

[28] C. Fraley and A. E. Raftery, "Enhanced model-based clustering, density estimation, and discriminant analysis software: Mclust," *J. Classification*, vol. 20, no. 2, pp. 263–286, 2003.

[29] L. Scrucca and A. E. Raftery, "clustvarsel: A package implementing variable selection for Gaussian model-based clustering in R," *J. Stat. Softw.*, vol. 84, no. 1, pp. 1–28, 2018.

[30] M. Marbac and M. Sedki. *VarSelLCM: Variable Selection for Model-Based Clustering of Continuous, Count, Categorical or Mixed-Type Data Set With Missing Values R Package Version 2.0.1*. Accessed: 2017. [Online]. Available: https://CRAN.R-project.org/package=VarSelLCM

[31] M. Sedki, G. Celeux, and C. Maugis-Rabusseau. *SelvarMix: Regularization for Variable Selection in Model-Based Clustering and Discriminant Analysis R Package Version 1.2.1*. Accessed: 2017. [Online]. Available: https://CRAN.R-project.org/package=SelvarMix

[32] J. L. Andrews and P. D. McNicholas. *VSCC: Variable Selection for Clustering and Classification R Package Version 0.2*. Accessed: 2013. [Online]. Available: https://cran.r-project.org/package=vscc

[33] P. L. Odell and B. S. Duran, *Cluster Analysis: A Survey*. New York, NY, USA: Springer, 1974.

[34] *UCI Machine Learning Repository*. Accessed: Mar. 6, 2021. [Online]. Available: http://archive.ics.uci.edu/ml

[35] H. Erol and F. Akdeniz, "A multispectral classification algorithm for classifying parcels in an agricultural region," *Int. J. Remote Sens.*, vol. 17, no. 17, pp. 3357–3371, Nov. 1996.

[36] J. J. Verbeek, N. Vlassis, and B. Kröse, "Efficient greedy learning of Gaussian mixture models," *Neural Comput.*, vol. 15, no. 2, pp. 469–485, 2003.

MARUF GOGEBAKAN was born in Gaziantep, Turkey, in 1981. He received the Ph.D. degree in mathematics/statistics from Erciyes University, in 2017.

From 2011 to 2017, he worked as a Research Assistant with the Applied Mathematics Department, Abdullah Gul University. Since 2019, he has been working as an Assistant Professor with the Department of Maritime Business Management, Bandirma Onyedi Eylül University. There are six articles and three book chapters on statistics and data analysis. His research interests include applied statistics, classification and clustering analysis, big data, and multivariate statistics applications.