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

Maruf Gogebakan

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 perspectives 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, 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, clustvassel, varselLMC, 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, clustvassel, varselLMC, 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 \( x = (x_1^T, ..., x_n^T)^T \) [23]. The probability density function of finite mixture distributions are as follows:

\[
 f(x_j; \Psi) = \sum_{i=1}^g \pi_i f(x_j; \theta_i) \tag{1}
\]

where \( f(x_j; \theta_i) \) are probability density functions of the components and \( \pi_i \) indicates the mixing weight (volume of clusters in the mixture model) in cases of \( 0 < \pi_i < 1 \) and \( \sum_{i=1}^g \pi_i = 1 \) \((i = 1, ..., g)\). The parameter vector \( \Psi = (\pi, \theta) \) contains all of the parameters of the mixture models. Here, \( \theta = (\theta_1, ..., \theta_g) \) denotes unknown parameters of the probability density function of the \( i \)-th component 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:

\[
 f(x_j; \Psi) = \sum_{i=1}^g \pi_i \Phi(x_j; \mu_i, \Sigma_i) \tag{2}
\]

where \( \Phi(x_j; \mu_i, \Sigma_i) \) are assumed to be multivariate Gaussian densities of the form

\[
 \Phi_i(x_j; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{p/2} |\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x_j-\mu_i)^T \Sigma_i^{-1}(x_j-\mu_i)} \tag{3}
\]

where \( \mu_i \) and \( \Sigma_i \) for \( i = 1, ..., k \) 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 \( \Psi = (\pi_1, ..., \pi_g, \xi) \), where \( \xi \) obtained from compound vectors of \( \mu = (\mu_1, ..., \mu_g) \) and \( \Sigma = (\Sigma_1, ..., \Sigma_g) \).

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:

\[
f(x; \theta) = \sum_{i=1}^{g} \pi_i f_i(x; \mu_i; \sigma_i)
\]

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

\[
f_i(x; \mu_i; \sigma_i) = \frac{1}{\sqrt{2\pi} \sigma_i} \exp \left\{ -\frac{1}{2} \left( \frac{x - \mu_i}{\sigma_i} \right)^2 \right\}
\]

where \( \mu_i \) denotes the mean and \( \sigma_i \) 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 \( \pi, \mu, \) and \( \sigma \) 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:

\[
\Sigma_i = \begin{bmatrix}
\sigma_1^2 & \rho_{12} \sigma_1 \sigma_2 & \cdots & \rho_{1p} \sigma_1 \sigma_p \\
\rho_{21} \sigma_1 \sigma_2 & \sigma_2^2 & \cdots & \rho_{2p} \sigma_2 \sigma_p \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{p1} \sigma_1 \sigma_p & \rho_{p2} \sigma_1 \sigma_p & \cdots & \sigma_p^2
\end{bmatrix}
\]

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 \( \rho_{ij} = Corr(X_{i1}, X_{j2}), \ldots, \rho_{pp, p-1} = Corr(X_{p1}, X_{p, p-1}) \).

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:

\[
\Sigma_k = \lambda_k D_k A_k D_k^T
\]

where the scalar constant \( \lambda_k \) 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 \( \text{Diag}(a_{1k}, \ldots, a_{pk}) \) where \( a_{1k} \geq a_{2k} \geq \cdots \geq a_{pk} \geq 0 \). Utilising this decomposition of the covariance matrix \( \Sigma_k \), 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 \( (\pi_i) \) 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, "mlclust" [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**: \( \mathbb{R}^d \rightarrow \mathbb{R}^d \): \( k_i \) : components of \( X_1 \) (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 \( k=1, \ldots, 4 \), U-GMM \( f(x; \theta) = \sum_{i=1}^{g} \pi_i f_i(x; \mu_i; \sigma_i) \) (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 \( C_{\text{min}} \) and \( C_{\text{max}} \) and are defined as follows:

\[
C_{\text{min}} = \max(k_s)
\]

\[
C_{\text{max}} = \prod_{s=1}^{p} k_s \quad , s = 1, \ldots, p
\]

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 \( M_{\text{Total}} \) can be calculated as follows:
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

\[
s(m, k) = \sum_{i=0}^{m} (-1)^i \binom{m}{i} (k - i)^m
\]

(11)

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

**Definition 1.** A function is defined as \( f: D(f) \rightarrow R(f) \) 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, \( D(f) = \{\max\{k\}, \prod_{i=1}^{k} k_i\}, \forall k_i \in \{C_{\min}, C_{\max}\} \) 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;

\[
R(f) = \left[\sum_{i=1}^{k} j_i \prod_{i=1}^{k} \left(\frac{j_i}{i!}\right) \right] \times \left[\prod_{i=1}^{k} \left(\frac{j_i}{i!}\right) \right] \times \left[\prod_{i=1}^{k} \left(\frac{j_i}{i!}\right) \right]
\]

(12)

where \( j_k \) and \( k_k \) 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;

\[
f^u = (x, \mu^u, \Sigma^u) = \sum_{i=1}^{g} \pi_i f(x, \mu_i^u, \Sigma_i^u)
\]

(13)

for \( u = 1, \ldots, k \), where \( \pi^u = \sum_{i=1}^{g} \pi_i \) are mixing proportions,

\[
\mu_i^u = \left[\begin{array}{c}
\mu_1 \\
\mu_2 \\
\vdots \\
\mu_p 
\end{array}\right]
\]

are mean vectors, and

\[
\Sigma_i^u = \left[\begin{array}{cccc}
\sigma_1^2 & \rho_{1,2}\sigma_1\sigma_2 & \cdots & \rho_{1,p}\sigma_1\sigma_p \\
\rho_{2,1}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2,p}\sigma_2\sigma_p \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{p,1}\sigma_p\sigma_1 & \rho_{p,2}\sigma_p\sigma_2 & \cdots & \sigma_p^2
\end{array}\right]
\]

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;

\[
o(H) = \sum_{i=1}^{g} c_i.
\]

(14)

where \( c_i \) represents “0” and “1” in elements of vector representations.

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

\[
\delta(H) = \arg\max_{ij}\|c_i - c_j\|, i, j = 1, \ldots, c_{max}
\]

(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.

**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 \( C_{\max} = \prod_{i=1}^{k} k_i \), 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 \( M_{total} = 2^{\prod_{i=1}^{k} k_i} - 1 \).
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 \( o(H) = \sum_{i=1}^{g} k_i \).

### E. INFORMATION CRITERIA FOR APPROPRIATE MIXTURE MODELS IN GRID STRUCTURE

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

\[
\log\text{L}(\pi, \mu, \Sigma) = \sum_{j=1}^{n} \log \left( \sum_{i=1}^{k} \pi_i f(x_j, \mu_i, \Sigma_i) \right)
\]

(16)

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

\[
\text{BIC} = 2\log\text{L}(\pi, \mu, \Sigma) - d\log n
\]

(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 $M_{\text{Total}}$ 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 $f(x; \Psi) = \sum_{i=1}^{g} \pi_i \mathcal{F}(x; \mu_i, \Sigma)$ 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.

**Figure 1**: The proposed approach for determining the number of clusters

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 used for this study. Each step of the study was done in the R software 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 $R^{dxn}$ 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 $(d xn \geq l xn)$. 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 $O(nd^2)$ [36] in multivariate models, for variable selection in univariate models information complexity is $O(nd)$. The information complexity of calculating model numbers to obtain vector representations of AGMMs is $O(md^2)$. Finally, the information complexity is $O(d^2 n + d mn)$ to obtain the information criteria for each AGMMs.

### 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, vareLCC, 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.

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

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

In this section, the principles of the proposed
MMSGM 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 components in the synthetic-1 data set and the number
of observations per component are given in Table 3.

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

$\mu_2 = [14.43 \ 44.69]$, $\Sigma_2 = [25.12 \ 1.104 \ 1.104 \ 26.43]$ and $\mu_3 = [9.64 \ 50.29]$, $\Sigma_3 = [23.294 \ 1.358 \ -1.266]$, $\Sigma_3 = [23.294 \ 1.358 \ 24.94 \ 0.335 \ 23.227]$ respectfully.

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.

The cluster number ranges, $C_{\text{min}} = \max(k_3) =
\max(1,2,3) = 3$ and $C_{\max} = \prod_{k=1}^3 k_3 = 1.23 = 6$, 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.

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

| 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 |

| 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 cluster number ranges, $C_{\text{min}} = \max(k_3) =
\max(1,2,3) = 3$ and $C_{\max} = \prod_{k=1}^3 k_3 = 1.23 = 6$, 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, $M_{\text{Total}}$ was computed as $M_{\text{Total}} = 2^6 - 1 = 63$ 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
\[
\mu_i^u = \begin{bmatrix} \mu_i^u \\ \mu_i^2 \\ \mu_i^3 \end{bmatrix} \quad \text{and} \quad \Sigma_i^u = \begin{bmatrix}
(\sigma_1^u)^2 & \rho_{1,2}^u \sigma_1^u \sigma_2^u & \rho_{1,3}^u \sigma_1^u \sigma_3^u \\
\rho_{2,1}^u \sigma_2^u \sigma_1^u & (\sigma_2^u)^2 & \rho_{2,3}^u \sigma_2^u \sigma_3^u \\
\rho_{3,1}^u \sigma_3^u \sigma_1^u & \rho_{3,2}^u \sigma_3^u \sigma_2^u & (\sigma_3^u)^2
\end{bmatrix}, \quad \text{where } i = 1, \ldots, k
\]
respectively. Here \(\sigma_1, \sigma_2 \) and \(\sigma_3\) 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:
\[
f_i^u = (x, \mu_i^u, \Sigma_i^u) = \sum_{i=1}^{n_i} \pi_i^u f_i(x, \mu_i^u, \Sigma_i^u), \quad \text{for } u = 1, 2, 3
\]
where \(\pi_i^u = \frac{\pi_i}{\sum_{s=1}^{n_s} \pi_s}\) 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 \(X_1, X_2\) and \(X_3\) and grid structures of appropriate models were obtained.

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

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 \(d(H) = \{1,3,4,6\}\).

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.
TABLE VI
RESULTS OF COMPARISON MMSCM AND MCLUST BASED METHODS ON THE SYNTHETIC-1 DATA SET

| Method     | 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| -4435 | -8942 | 4        | X2-X3     | EII        | 85   | 78   |
| VarselLCM  | -     | -12622| 4        | X1-X2-X3  | -          | 84   | 78   |
| selvarMix  | -     | -1664 | 4        | X2-X3     | EEI        | 85   | 78   |
| vscc       | -     | -     | 4        | X2-X3     | EEI        | 85   | 78   |

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

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

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

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 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 \(\mu_1 = [21.3, 57.1]\), \(\Sigma_1 = \begin{bmatrix} 84.38 & -1.67 \\ -1.67 & 119.88 \end{bmatrix}\) and \(\mu_2 = [21.4, 61.1]\), \(\Sigma_2 = \begin{bmatrix} 92.42 & 5.39 \\ 5.39 & 80.81 \end{bmatrix}\), respectively. The mean and standard deviation values for other homogeneous variables are \(\mu_{3,4,...,15} = \begin{bmatrix} 27.5, 32.3, 18.8, 23.5, 30.3, 56.8, 64.3, 20.7, 44.3, 39.1, 36.2, 28.1, 50.1 \end{bmatrix}\) and
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.

### Table VII

| 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      | -1471.1 | 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 |          |        |        |      |

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.

### Table VIII

| Variables | $X_1$ | $X_2$ | $X_3, X_4, ..., X_{15}$ |
|-----------|-------|-------|-------------------------|
| Components | $X_{1,1}$ | $X_{1,2}$ | $X_{2,1}$ | $X_{2,2}$ | - |
| # observations | 120 | 280 | 150 | 250 | 400 |

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. $\max(2,2) = 2$ 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.
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.

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.

According to the results in Table 11 for the Synthetic-2 data set, the success of the proposed method and the mclust based

| Variables          | Sepal length | Sepal width | Petal length | Petal width |
|--------------------|--------------|-------------|--------------|-------------|
| #Comp.             | 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     |

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

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.

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.

---

**TABLE XII**

| 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 | 70 | 77 | 50 | 64 | 61 | 83 | 61 |
| Total | 150 | 150 | 150 | 150 |

The range is calculated as $C_{min} = max\{k_j\} = max(2,1,2,2) = 2$, and $C_{max} = \prod_{s=1}^{p} k_s = 2 \times 1 \times 2 \times 2 = 8$ 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 $s = 1,2$, $s = 1,1$, $s = 1,2$, 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 $X_1, X_2$ and $X_4$, the clusters in the model, and the components forming each cluster are shown in Figure 9.

FIGURE 9. Components of variables $X_1$, $X_2$ and $X_4$ in reduced Iris data set.

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.

**TABLE XIII**

| #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 XIV**

| #Component | logL | BIC  | Vector Rep. |
|------------|------|------|-------------|
| 3          | -175.93 | -499.12 | 10010010   |

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 $j_1 = 2, j_2 = 2$ and $j_4 = 2$ correspond to components variables $X_1, X_3$ and $X_4$. $i_1, i_2, i_3$, 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.
FIGURE 10. The number of components (2) and covariance types for complete data with 4 variables

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

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

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

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 \((X_{11}, X_{31} \text{ and } X_{41})\), 4th centre \((X_{12}, X_{32} \text{ and } X_{42})\) and 7th centre \((X_{11}, X_{32} \text{ and } X_{42})\) are composed of components.

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.

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.

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.

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.

2) 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.

| Variable | $X_1$ | $X_2$ | $X_3$ |
|----------|-------|-------|-------|
| #Comp.   | #Param. | logL  | BIC   |
| 1        | 9      | $-13947.0$ | $-27897.0$ |
| 2        | 19     | $-13767.0$ | $-27538.0$ |
| 3        | 19     | $-13567.0$ | $-27242.0$ |
| 4        | 39     | $-13617.0$ | $-27246.0$ |
| 5        | 49     | $-13617.0$ | $-27249.0$ |

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 $C_{max} = k_1k_2k_3 = 3.3.3 = 27$, 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 the best model are...
shown in Figure 13 as \( o(H) = 5 \) and \( \delta(H) = \{1, 6, 13, 18, 20\} \).

| logL      | BIC     | Vector Representation |
|-----------|---------|-----------------------|
| -392715.4 | -785660.7 | 1000010000010000101000000 |

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.

| logL      | 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 | -       | -        | 9         | X1-X2-X3   | VVV  | 51   | 32   |
| vscc      | -       | -        | 9         | X1-X2-X3   | VVV  | 49   | 35   |

| Datasets  | mclust | clustvarsel | varselLCM | selvarMix | vscc | MMSCM |
|-----------|--------|-------------|-----------|-----------|------|-------|
| Synthetic-1 | 1.8381 | 6           | 80.3940   | 9.3431    | 5.2669 | 3.254 |
| Synthetic-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 |

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 XX also shows the acceleration and high-quality clustering results of MMSCM.

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

|        | 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          |

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.

Figure 16 shows the result of the clustering method proposed in the LSI data set as shown in Figure 16.

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] McLachlan, G. J., & Rathnayake, S. (2014). On the number of components in a Gaussian mixture model. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4(5), 341-355.

[2] Bozdogan, H. (1994). 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 (pp. 169-177). Springer Berlin Heidelberg.

[3] Ng, T. L. J., & Murphy, T. B. (2021). Model-based Clustering of Count Processes. Journal of Classification, 38(2), 188-211.

[4] McNicholas, P. D. (2016). Model-based clustering. Journal of Classification, 33(3), 331-373.
[5] Celeux, G., Frühwirth-Schnatter, S., & Robert, C. P. (2018). Model selection for mixture models perspectives and strategies. Handbook of mixture analysis, 121-160.

[6] Ye, X., Zhao, J., & Chen, Y. (2018). A nonparametric model for multi-manifold clustering with mixture of Gaussians and graph consistency. Entropy, 20(11), 830.

[7] Raftery, A. E., & Dean, N. (2006). Variable selection for model-based clustering. Journal of the American Statistical Association, 101(473), 168-178.

[8] Fraley, C. and Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association, 97, 611-631.

[9] Galimberti, G., & Soffritti, G. (2007). Model-based methods to identify multiple cluster structures in a data set. Computational statistics & data analysis, 52(1), 520-536.

[10] Galimberti, G., Manisi, A., & Soffritti, G. (2018). Modelling the role of variables in model-based cluster analysis. Statistics and Computing, 28(1), 145-169.

[11] Akogul, S., & Ersoyoglu, M. (2017). An approach for determining the number of clusters in a model-based cluster analysis. Entropy, 19(9), 452.

[12] Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). Model-based clustering and classification for data science: with applications in R (Vol. 50). Cambridge University Press.

[13] Shang, R., Meng, Y., Wang, W., Shang, F., & Jiao, L. (2019). Local discriminative based sparse subspace learning for feature selection. Pattern Recognition, 92, 219-230.

[14] Shang, R., Wang, W., Stolkin, R., & Jiao, L. (2017). Non-negative spectral learning and sparse regression-based dual-graph regularized feature selection. IEEE transactions on cybernetics, 48(2), 793-806.

[15] Maugis, C., Celeux, G., & Martin-Magniette, M. L. (2009). Variable selection for clustering with Gaussian mixture models. Biometrics, 65(3), 701-709.

[16] Fop, M., & Murphy, T. B. (2018). Variable selection methods for model-based clustering. Statistics Surveys, 12, 18-65.

[17] Sharma, K. K., Seal, A., Herrera-Viedma, E., & Krejcar, O. (2021). An Enhanced Spectral Clustering Algorithm with S-Distance. Symmetry, 13(4), 596.

[18] Sharma, K. K., & Seal, A. (2020). Clustering analysis using an adaptive fused distance. Engineering Applications of Artificial Intelligence, 96, 103928.

[19] Sharma, K. K., Seal, A., Yazidi, A., Selamat, A., & Krejcar, O. (2021). Clustering Uncertain Data Objects using Jeffreys-Divergence and Maximum Bipartite Matching based Similarity Measure. IEEE Access.

[20] Sharma, K. K., & Seal, A. (2021). Spectral embedded generalized mean based k-nearest neighbors clustering with s-distance. Expert Systems with Applications, 169, 114326.

[21] Wei, Y. and McNicholas, P. D. (2015). Mixture model averaging for clustering. Advanced Data Analysis and Classification. 9:197–217. DOI 10.1007/s11634-014-0182-6.

[22] Gogebakan, M., & Erol, H. (2018). A New Semi-supervised Classification Method Based on Mixture Model Clustering for Classification of Multispectral Data. Journal of the Indian Society of Remote Sensing, 46(8), 1323-1331.

[23] Melnykov, V. and Maitra, R. (2010). Finite mixture models and model based clustering. Statistics Surveys, 4 80–116.

[24] Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. Computational Statistics & Data Analysis 71 52 – 78.

[25] Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. Biometrics, 803-821.

[26] Schwarz, G. (1978). Estimating the dimension of a model, Ann. Statist. 6 pp. 461–464.

[27] Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. Pattern recognition, 28(5), 781-793.

[28] Fraley, C., & Raftery, A. E. (2003). Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST. Journal of Classification, 20(2), 263-286.

[29] Scrucca, L., & Raftery, A. E. (2018). clustervarsel: a package implementing variable selection for Gaussian model-based clustering in R. Journal of Statistical Software, 84.

[30] Marbac, M., & Sedki, M. (2017). 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.
[31] Sedki, M., Celeux, G., & Maugis-Rabusseau, C. (2017). SelvarMix: Regularization for variable selection in model-based clustering and discriminant analysis. R package version, 1(1).

[32] Andrews, J. L., & McNicholas, P. D. (2013). vscc: Variable selection for clustering and classification. R package version 0.2.

[33] Odell, P. L., & Duran, B. S. (1974). Cluster Analysis: A Survey, New York, Springer Berlin Heidelberg

[34] UCI Machine Learning Repository. Available online: http://archive.ics.uci.edu/ml (accessed on 6 March 2021).

[35] Erol, H., & Akdeniz, F. (1996). A multispectral classification algorithm for classifying parcels in an agricultural region. International Journal of Remote Sensing, 17(17), 3357-3371.

[36] Verbeek, J. J., Vlassis, N., & Kröse, B. (2003). Efficient greedy learning of Gaussian mixture models. Neural computation, 15(2), 469-485.

Dr. MARUF GÖGEBAKAN was born in 1981 in Gaziantep, Turkey. He completed his PhD in Mathematics/Statistics at Erciyes University in 2017.

Between 2011 and 2017, he worked as a research assistant in the Applied Mathematics department at Abdullah Gül University. Since 2019, he has been working as an Assistant Professor in the Department of Maritime Business Management at Bandirma Onyedi Eylül University. There are 6 articles and 3 book chapters on statistics and data analysis. His research interests include Applied statistics, classification and clustering analysis, big data, multivariate statistics applications.