Sinop Üniversitesi Fen Bilimleri Dergisi *Sinop Uni J Nat Sci 2(1): 121-132 (2017)*
ISSN: **2536-4383**

Research Article

Evaluation of Group Homogeneity in Gaussian Mixture Models Using Combined Cluster and Discriminant Analysis

Ezgi Nazman [*a], Semra Erbaş [a]

[a] *Gazi University, Faculty of Science, Department of Statistics, Ankara, Turkey*

Abstract

Cluster analysis is a widely used multivariate statistical method in many fields. Pairwise overlap is a measure of interaction between mixture components. Determining the number of homogeneous group is a difficult process due to the pairwise overlap. In this study, combined cluster and linear discriminant analysis is compared with combined cluster and quadratic discriminant analysis. Correctly classification rates of the Gaussian mixture components are obtained. Also, whether further division is necessary to obtain homogeneous groups is determined. The comparisons have been conducted by a simulation study for 81 different scenarios and an application is presented.

Keywords: Clustering, discriminant analysis, pairwise overlap, Gaussian mixture models

Birleştirilmiş Kümeleme ve Diskriminant Analizi Kullanarak Gauss Karma Modellerde Grup Homojenliğinin Değerlendirilmesi

Öz

Kümeleme analizi, pek çok alanda yaygın olarak kullanılan çok değişkenli bir istatistiksel analiz yöntemidir. İkili örtüşme, karma bileşenler arasındaki etkileşimin bir ölçüsüdür. İkili örtüşmeden ötürü homojen grup sayısını belirlemek zor bir süreçtir. Bu çalışmada, birleştirilmiş kümeleme ve lineer diskriminant analizi ile birleştirilmiş kümeleme ve karesel diskriminant analizi karşılaştırılmıştır. Gauss karma bileşenlerin doğru sınıflama oranları elde edilmiştir. Ayrıca, homojen grupları elde etmek için daha ileri bölünmenin gerekli olup olmadığı belirlenmiştir. Karşılaştırma 81 farklı senaryo ile yürütülmüş ve bir uygulama sunulmuştur.

Anahtar Kelimeler: Kümeleme, diskriminant analizi, ikili örtüşme, Gauss karma modeller

* Corresponding author :
e-mail: ezgicabuk@gazi.edu.tr

Received: 9.10.2016
Accepted: 21.2.2017

ISSN: **2536-4383**

Introduction

In recent years, mixture models have been often used to fit data where each of the mixture components are taken into the consideration as different groups or clusters in many fields. One of the widely used methods to extract information underlying components is cluster analysis [1,2,3]. Machine learning, artificial intelligence, pattern recognation, web mining and image segmentation in engineering; genetic, biology, microbiology, paleontology, psychiatry and clinic in medicine; geograpy, geology and remote sensing in earth science; sociology, psychology and archeology in social science; marketing and bussiness in economics can be given as example of the fields in which finite mixture models can be used to fit data [4,5,6]. On the other hand, how many groups should be homogeneously chosen has become a general question. It is known that the number of mixture components can not be always the same number with the number of groups [3,4,7,8]. Moreover, pairwise overlap of components might cause a complexity during determining group memberships of the observations, so there is need to simplify the clustering process.

One of the applicable methods for validation of cluster results is linear discriminant analysis [9,10]. However, assumption that homogeneity of variance-covariance matrix of the groups is not always provided in many situation. When the assumption that homogeneity of variance-covariance matrix of the groups is violated, quadratic discriminant analysis is generally preferred to discriminate observations to the groups and to classify new observations to the related groups [11,12].

There are several studies comparing the performance of linear discriminant analysis and quadratic discriminant analysis in literature [13,14]. Cherry [15] combined cluster and linear discriminant analysis to develop a social bond typology of runway youth. Hastie and Tibshirani [10] investigated the functionality of classification for Gaussian mixtures and applied a nonparametric regression method. Baudry et al. [16] determined the number of component in the mixture model by combining Bayesian Information Criterion (BIC) and Integrated Completed Likelihood (ICL) criterions. Tanos et al. [17] applied combined cluster and linear discriminant analysis to optimize monitoring network on the River Tisza. Morris and McNicholas [18] aimed to reveal underlying clusters in data set for generalized hyperbolic mixture models by applying clustering, classification and dimension reduction methods. Novak et al. [19] used combined cluster and linear discriminant analysis to verify chemically uniqueness of diesel fuel samples. Kovacs et al. [20] presented combined cluster and discriminant analysis and its applicability of the method with a case study on the water quality samples of Neusiedler See. In this regard, a comparison between linear discriminant analysis and quadratic discriminant analysis for Gaussian mixture models was aimed in order to reveal the change on correctly classification rates and detection of component homogeneity.

In this study, our purpose is to obtain homogeneous number of groups from Gaussian mixture components which do not require further division. The comparison between linear discriminant analysis and quadratic discriminant analysis is studied by using a combined cluster and discriminant analysis method with a simulation study for various number of components, number of variables, sample sizes and pairwise overlaps. Correctly classification rates are calculated by dividing the number of correctly classified observations of components by the sample size are computed. Meanwhile, the homogeneity of the components are

***ISSN:* 2536-4383**

investigated in order to reveal whether further division of the components are necessary. By this means, our study will able to offer an insight into the study plans of researchers for following studies before deciding sample structure and discrimination method.

This study is organized as follows: In Section 2, cluster analysis is presented. Linear discriminant analysis, quadratic discriminant analysis and combined cluster and discriminant analysis are presented in Section 3. In Section 4, Gaussian mixture model is introduced. The process of the simulation is described in Section 5 with the results. In Section 6, an application is presented. In the final section, conclusion is given with the highlighted inferences.

Cluster Analysis (CA)

Cluster analysis (CA) is a multivariate statistical method which helps grouping observations of which natural groups are certainly unknown [2]. There are fundamentally two cluster analysis: hierarchical cluster analysis (HCA) and non-hierarchical cluster analysis. Hierarchical cluster analysis proceeds by either a series of successive mergers or a series of successive divisions. Hierarchical cluster analysis seeks to build a nested hierarchy which can cause decrease or increase on the number of groups. Agglomerative hierarchical algorithm starts with the individual data points and merges the most similar groups [5].

Discriminant Analysis (DA)

Discriminant analysis (DA) is a multivariate statistical method concerned with separating distinct sets of objects or observations and with allocating new objects to previously defined groups. Discrimination terminology was introduced by Fisher in the first seperatory problems [21]. Both linear discriminant analysis and quadratic discriminant analysis are two most widely used statistical methods for classification problems [22].

G_1 and G_2 are the names of two groups and their number of observations are shown as n_1 and n_2, respectively. $\overline{X}_1$, $\overline{X}_2$ and S_1, S_2 indicate sample mean vectors, and estimated variance-covariance matrix based on sample sizes n_1 and n_2, respectively. The prior probability of G_i is given as prior probability p_i, $i=1,2$ and $p_1 + p_2 = 1$.

Linear Discriminant Analysis (LDA)

Linear discrimination method assumes that each group comes from a normal distribution with a common variance-covariance matrix.

LDA classifies an observation x_0 to group G_1 if

$$(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1} x_0 - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)' S_{pooled}^{-1}(\bar{x}_1 + \bar{x}_2) \geq \ln\left[\frac{c(1/2)}{c(2/1)}\left(\frac{p_2}{p_1}\right)\right] \quad (1)$$

where

$$S_{pooled} = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2} \quad (2)$$

and S_{pooled} is the pooled estimator of the common variance-covariance matrix.

Quadratic Discriminant Analysis (QDA)

Unlike LDA, QDA does not assume a common variance-covariance matrix. With this respect, Box-M test is one of the commonly used methods to test homogeneity of variance-covariance matrices of groups.

$$-\frac{1}{2}x_0'\left(S_1^{-1} - S_2^{-1}\right)x_0 + \left(\mu_1' S_1^{-1} - \mu_2' S_2^{-1}\right)x_0 - k \leq \ln\left[\left(\frac{c(1/2)}{c(2/1)}\left(\frac{p_2}{p_1}\right)\right)\right] \quad (3)$$

where

$$k = \frac{1}{2}\ln\left(\frac{|S_1|}{|S_2|}\right) + \frac{1}{2}\left(\bar{x}_1' S_1^{-1}\bar{x}_1 - \bar{x}_2' S_2^{-1}\bar{x}_2\right) \quad (4)$$

ISSN: **2536-4383**

Combined Cluster and Discriminant Analysis (CCDA)

Combined cluster and discriminant analysis is a method of which idea is to compare random grouping with preconceived grouping [23]. When the groups are not able to be discriminated by LDA, random groupings will be convenient to determine number of observations correctly classified. There exist the significance of random groupings behind of the idea. The package of this combined method can be found in *R software* under the *ccda* name and contains three main steps and these steps are given in the Section 3.3.1.

Combined Cluster and Linear Discriminant Analysis (CCLDA)

The steps of combined cluster and linear discriminant analysis as follows if the variance-covariance matrices of components are homogeneous after applying Box-M test [20].

Let k be the number of component $i = 1,...,K$.

I) Components are obtained by applying HCA using Ward's method.

IIa) Each observation is labeled according to the component membership.

IIb) The correctly classification rates of labeled observations are obtained by LDA ($ratio_i$).

IIc) The labels of components are obtained randomly by permuting the labels.

IId) The randomly obtained correctly classification rates of labeled observations are obtained by LDA. Then, 95% quantiles of these correctly classification rates are determined (q_{95}).

IIe) The difference between IIb and IId is calculated ($d_i = ratio_i - q_{i,95}$).

III) The number of components gives optimal group numbers in the situation where the difference value (d_i) is maximum. It is decided that investigated componets can not be divided further groups and it is homogeneous. Otherwise, the process of division groups should be proceeded.

Combined Cluster and Quadratic Discriminant Analysis (CCQDA)

In real life, assumption that a common variance-covariance matrix of groups tend to be violated in many situations especially for finite mixture models. QDA has a quadratic decision boundary, so application of this method might be more flexible.

The algorithm in Section 3.3.1 is conducted for QDA when the variance-covariance of the components are not homogeneous. So, if the Box-M test shows that the variance-covariance matrix of components are not homogeneous *step IIb* and *step IId* are conducted using QDA, respectively.

Gaussian Mixture Models

Gaussian mixture models containing more than one component have been studied in many fields. For example, wheats from different fields in agriculture, healthy and sick people in medicine, soil from different lands, water samples from different locations in the same river in geology and stock returns in crisis time and typical times in finance are some of the situations where researchers have to decide number of groups more carefully [23,24].

Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ consist of n independent and identically distributed p-dimensional observations from a finite mixture probability with multivariate Gaussian density $\phi(x; \mu_k, \Sigma_k)$ where μ_k is the mean vector and Σ_k is the covariance matrix of k th component.

In finite mixture modeling, each of the components are assumed that they have

***ISSN:* 2536-4383**

its own distribution and probability. The prior probability of the k th component is shown with π_k, and $\sum_{i=1}^{k}\pi_k = 1$. Let $\mathbf{X}$ be distributed according to the finite mixture model given in Eq.5

$$g(x;v) = \sum_{k=1}^{K}\pi_k \phi(x;\mu_k,\Sigma_k) \quad (5)$$

where

$$v = (\pi', v'_1, v'_2, ..., v'_K) \quad (6)$$

The pairwise overlap between i th and j th groups is shown in Eq. 7

$$w_{ij} = w_{i\backslash j} + w_{j\backslash i} \quad (7)$$

where

$w_{j\backslash i}$ presents the misclassification rate explained that X is mistakenly assigned to the j th group, although it is originated from the i th group. Similarly, $w_{i\backslash j}$ indicates misclassification rate that originated from the j th group, but it is mistakenly assigned to the i th group. $w_{j\backslash i}$ is given in the Eq. 8

$$w_{j\backslash i} = P\left[\pi_i \phi_i(X;\mu_i,\Sigma_i) < \pi_j \phi_j(X;\mu_j,\Sigma_j) \backslash X \sim N_p(\mu_i,\Sigma_i)\right] \quad (8)$$

Let's assume that we have already known that 1s, 2s, and 3s are originally from the components A, B and C, respectively. If we obtain nine observations 2, 1, 2, 3, 1, 2, 2, 3, 1 belong to the components C, A, C, A, B, B, B, A, A, respectively. $w_{A\backslash B} = 1/9$, $w_{A\backslash C} = 2/9$ and $w_{B\backslash C} = 2/9$.

Simulation Study

Performances of both CCLDA and CCQDA were evaluated with a simulation study. *MixSim* [23] was used to generate Gaussian distributed data set. *ccda* [20] were used to obtain the correctly classification rates of LDA and QDA investigating homogeneity of the components. Besides, for components contained different pairwise overlaps were investigated to reveal whether further divisions are needed.

In the simulation design, 81 scenarios were conducted and shown in Table1 where the number of components (k) are 5, 10 and 20; number of variables (p) are 2, 3 and 5; sample sizes (n) are 200, 500 and 1000; pairwise overlaps ($\overline{w}$) are 0.01, 0.05 and 0.1, respectively. These notations are validated for all tables. The iteration was repeated 10000 times.

According to Table 2, the highest correctly classification rates for CCLDA and CCQDA were computed for the Scenario8 and Scenario5, whereas the lowest correctly classification rates were computed for the Scenario73 and Scenario75, respectively. For Scenario71, correctly classification rates of CCLDA and CCQDA were computed as equal. Correctly classification rates of both LDA and QDA tend to decrease, when the number of component increase. It was seen that the correctly classification rates of CCQDA is higher than the correctly classification rates of CCLDA for seven scenarios when the number of components are 5 and 20.

Table 3 indicates the number of obtained homogeneous groups after applying both CCLDA and CCQDA. If obtained number of group is the same as the number of component, the components have already been homogeneous and there is no need division of components into less groups. The least number of homogeneous groups are obtained 4, 6 and 13 for number of components 5, 10 and 20, respectively.

***ISSN:* 2536-4383**

Table 1: Scenarios for number of components, sample sizes, number of variables and pairwise overlaps

$k=5$	$\bar{w}=0.01$			$\bar{w}=0.05$			$\bar{w}=0.1$		
	$p=2$	$p=3$	$p=5$	$p=2$	$p=3$	$p=5$	$p=2$	$p=3$	$p=5$
$n=200$	Scenario1	Scenario4	Scenario7	Scenario10	Scenario13	Scenario16	Scenario19	Scenario22	Scenario25
$n=500$	Scenario2	Scenario5	Scenario8	Scenario11	Scenario14	Scenario17	Scenario20	Scenario23	Scenario26
$n=1000$	Scenario3	Scenario6	Scenario9	Scenario12	Scenario15	Scenario18	Scenario21	Scenario24	Scenario27
$k=10$	$\bar{w}=0.01$			$\bar{w}=0.05$			$\bar{w}=0.1$		
	$p=2$	$p=3$	$p=5$	$p=2$	$p=3$	$p=5$	$p=2$	$p=3$	$p=5$
$n=200$	Scenario28	Scenario31	Scenario34	Scenario37	Scenario40	Scenario43	Scenario46	Scenario49	Scenario52
$n=500$	Scenario29	Scenario32	Scenario35	Scenario38	Scenario41	Scenario44	Scenario47	Scenario50	Scenario53
$n=1000$	Scenario30	Scenario33	Scenario36	Scenario39	Scenario42	Scenario45	Scenario48	Scenario51	Scenario54
$k=20$	$\bar{w}=0.01$			$\bar{w}=0.05$			$\bar{w}=0.1$		
	$p=2$	$p=3$	$p=5$	$p=2$	$p=3$	$p=5$	$p=2$	$p=3$	$p=5$
$n=200$	Scenario55	Scenario58	Scenario61	Scenario64	Scenario67	Scenario70	Scenario73	Scenario76	Scenario79
$n=500$	Scenario56	Scenario59	Scenario62	Scenario65	Scenario68	Scenario71	Scenario74	Scenario77	Scenario80
$n=1000$	Scenario57	Scenario60	Scenario63	Scenario66	Scenario69	Scenario72	Scenario75	Scenario78	Scenario81

***ISSN:* 2536-4383**

Table 2: Correctly classification rates of CCLDA and CCQDA for number of components, sample sizes, number of variables and pairwise overlaps

$k=5$	$\bar{w}=0.01$						$\bar{w}=0.05$						$\bar{w}=0.1$					
	$p=2$		$p=3$		$p=5$		$p=2$		$p=3$		$p=5$		$p=2$		$p=3$		$p=5$	
	CCLDA	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**
$n=200$	0.975	0.950	0.960	0.975	0.975	0.970	0.930	0.895	0.895	0.905	0.945	0.885	0.805	0.815	0.810	0.805	0.870	0.795
$n=500$	0.980	0.958	0.972	0.976	0.988	0.972	0.914	0.962	0.884	0.882	0.926	0.894	0.860	0.900	0.810	0.792	0.846	0.816
$n=1000$	0.981	0.952	0.975	0.969	0.984	0.976	0.914	0.911	0.895	0.886	0.918	0.897	0.850	0.917	0.820	0.781	0.844	0.805
$k=10$	$\bar{w}=0.01$						$\bar{w}=0.05$						$\bar{w}=0.1$					
	$p=2$		$p=3$		$p=5$		$p=2$		$p=3$		$p=5$		$p=2$		$p=3$		$p=5$	
	CCLDA	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**
$n=200$	0.965	0.925	0.950	0.955	0.960	0.950	0.865	0.825	0.795	0.800	0.860	0.770	0.635	0.610	0.660	0.640	0.715	0.745
$n=500$	0.964	0.946	0.968	0.926	0.970	0.950	0.810	0.760	0.818	0.792	0.868	0.790	0.656	0.592	0.718	0.606	0.768	0.664
$n=1000$	0.956	0.937	0.966	0.932	0.969	0.942	0.797	0.799	0.810	0.774	0.841	0.789	0.665	0.606	0.692	0.646	0.734	0.665
$k=20$	$\bar{w}=0.01$						$\bar{w}=0.05$						$\bar{w}=0.1$					
	$p=2$		$p=3$		$p=5$		$p=2$		$p=3$		$p=5$		$p=2$		$p=3$		$p=5$	
	CCLDA	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**
$n=200$	0.950	0.895	0.940	0.942	0.935	0.735	0.725	0.675	0.740	0.745	0.770	0.750	0.560	0.565	0.655	0.580	0.640	0.595
$n=500$	0.908	0.912	0.950	0.908	0.928	0.868	0.750	0.736	0.768	0.666	0.644	0.644	0.662	0.602	0.644	0.528	0.624	0.536
$n=1000$	0.914	0.917	0.947	0.903	0.949	0.893	0.704	0.752	0.779	0.883	0.777	0.667	0.616	0.501	0.630	0.511	0.645	0.532

***ISSN:* 2536-4383**

Table 3: Obtained group numbers of CCLDA and CCQDA for number of components, sample sizes, number of variables and pairwise overlaps

$k=5$	$\bar{w}=0.01$						$\bar{w}=0.05$						$\bar{w}=0.1$					
	$p=2$		$p=3$		$p=5$		$p=2$		$p=3$		$p=5$		$p=2$		$p=3$		$p=5$	
	CCLDA	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**
$n=200$	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
$n=500$	5	5	5	5	5	5	5	4	5	5	5	5	5	4	5	5	5	5
$n=1000$	5	5	5	5	5	5	5	5	5	5	5	5	5	4	5	5	5	5
$k=10$	$\bar{w}=0.01$						$\bar{w}=0.05$						$\bar{w}=0.1$					
	$p=2$		$p=3$		$p=5$		$p=2$		$p=3$		$p=5$		$p=2$		$p=3$		$p=5$	
	CCLDA	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**
$n=200$	10	10	10	10	10	9	9	9	10	10	10	9	9	10	10	10	10	6
$n=500$	10	10	10	10	9	10	10	10	10	10	9	10	10	10	10	10	9	10
$n=1000$	10	10	10	10	10	10	10	9	10	10	10	10	10	9	9	10	10	10
$k=20$	$\bar{w}=0.01$						$\bar{w}=0.05$						$\bar{w}=0.1$					
	$p=2$		$p=3$		$p=5$		$p=2$		$p=3$		$p=5$		$p=2$		$p=3$		$p=5$	
	CCLDA	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**	**CCLDA**	**CCQDA**
$n=200$	18	20	19	19	19	18	18	20	20	20	17	18	18	17	14	19	20	18
$n=500$	20	19	19	18	20	20	17	14	17	19	13	19	13	20	17	19	20	18
$n=1000$	20	20	18	18	18	17	17	13	18	18	17	20	20	17	18	18	17	19

ISSN: **2536-4383**

Table 4: The superiority results of CCLDA and CCQDA for number of components, sample sizes, number of variables and pairwise overlaps

$k=5$	$\bar{w}=0.01$			$\bar{w}=0.05$			$\bar{w}=0.1$		
	$p=2$	$p=3$	$p=5$	$p=2$	$p=3$	$p=5$	$p=2$	$p=3$	$p=5$
$n=200$	CCLDA	CCQDA	CCLDA	CCLDA	CCQDA	CCLDA	CCQDA	CCLDA	CCLDA
$n=500$	CCLDA	CCQDA	CCLDA	CCQDA	CCLDA	CCLDA	CCQDA	CCLDA	CCLDA
$n=1000$	CCLDA	CCLDA	CCLDA	CCLDA	CCLDA	CCLDA	CCQDA	CCLDA	CCLDA
$k=10$	$\bar{w}=0.01$			$\bar{w}=0.05$			$\bar{w}=0.1$		
	$p=2$	$p=3$	$p=5$	$p=2$	$p=3$	$p=5$	$p=2$	$p=3$	$p=5$
$n=200$	CCLDA	CCQDA	CCLDA	CCLDA	CCQDA	CCLDA	CCLDA	CCLDA	CCQDA
$n=500$	CCLDA	CCLDA	CCLDA	CCLDA	CCLDA	CCLDA	CCLDA	CCLDA	CCLDA
$n=1000$	CCLDA	CCLDA	CCLDA	CCQDA	CCLDA	CCLDA	CCLDA	CCLDA	CCLDA
$k=20$	$\bar{w}=0.01$			$\bar{w}=0.05$			$\bar{w}=0.1$		
	$p=2$	$p=3$	$p=5$	$p=2$	$p=3$	$p=5$	$p=2$	$p=3$	$p=5$
$n=200$	CCLDA	CCQDA	CCLDA	CCLDA	CCQDA	CCLDA	CCQDA	CCLDA	CCLDA
$n=500$	CCQDA	CCLDA	CCLDA	CCLDA	CCLDA	CCLDA	CCLDA	CCLDA	CCLDA
$n=1000$	CCQDA	CCLDA	CCLDA	CCQDA	CCQDA	CCLDA	CCLDA	CCLDA	CCLDA

***ISSN:* 2536-4383**

Table 4 is prepared to understand the comparison results of CCLDA and CCQDA. When the number of component is 5, CCQDA shows superiority for seven scenarios. For twenty scenarios, CCLDA has more correctly classification rates than CCQDA. When the number of component is 10, CCQDA shows superiority for only four scenarios. It is seen that LDA also tend to divide components into less homogeneous groups for Scenario71.

CCLDA is more succesfull in terms of correctly classification rates for most of the scenarios for Gaussian mixture data. However, the correctly classification rates both CCLDA and CCQDA generally decrease when the number of component increases. It is also clear that large average pairwise overlap can cause low correctly classification rates for both CCLDA and CCQDA.

Application Study

Glass identification data set from UCI Machine Learning Repository was used for application study. The sample size of data set is 214 and number of variables is 9. The description of the variables are refractive index (RI), and mass percentages of the elements Na, Mg, Al, Si, K, Ca, Ba and Fe. The data set was created by B. German (Central Research Establishment, England) and it is available in the *MASS package* of *R*. The purpose of classification considering glass sample is to investigate forensic studies. The high correctly classification rates and correctly clustered components are helpful to recognize crime behind the variables. Building windows float processed (1), building windows non-float processed (2), vehicle windows float processed (3), containers (4), tableware (5) and headlamps (6) are the components of data. According to the Box-M results, p-value was found 0.0001. Therefore, at individual significance level ($\alpha = 0.05$), there is sufficient evidence to reject the null hypothesis H_0 that the variance-covariance matrices of the Glass data set are equal.
Thus, we used CCQDA in order to obtain correctly classification rate and the number of homogeneous groups.
After applying CCQDA, the correctly classification rate was calculated as 94.3%. Besides, number of component was reduced from 6 to 4. Together the components number 1, 2 and 3 constitute one group. The remained components number 4, 5 and 6 still remain only one group after applying CCQDA.

Conclusion

The basic idea of our study is to overcome uncertainity of the number of homogeneous groups for Gaussian mixture components. With this purpose, we evaluated correctly classification rates of CCLDA and CCQDA and determined the number of homogeneous groups in terms of various number of components, number of variables, sample sizes and pairwise overlaps using 81 scenarios. In general, despite the correctly classification rates of CCLDA is obtained higher than correctly classification rates of CCQDA for totally 62 scenarios, there is no large differences between CCLDA and CCQDA. It can be inferred that high overlap rate and high number of component may cause decrease in the correctly classification rates for both CCLDA and CCQDA. It has been known that LDA has superiority to QDA because of the estimated number of parameters during classification process [22]. This study reveals that the number of component and the overlap rate are the other important considerations in addition to the number variable and sample size. While making a decision about number of homogeneous groups, CCLDA can be preferred instead of CCQDA in most cases.

Since CCDA does not provide satisfactory results when there exist outliers

in data set, outliers and noise variables are still open to be a part of discussion. This study will able to offer an insight into the study plans of researchers for following studies before deciding sample structure and discrimination method.

References

[1] Reynolds D, 2009. Gaussian mixture models. In Encyclopedia of Biometrics, 1st ed. New York: Springer Science and Business Media, 659–663.

[2] Everitt B, Landau S, Leese M, Stahl D,2011. Cluster Analysis. 5th ed. Wiley Series.

[3] Melnykov V, Maitra R, 2010. Finite mixture models and model based clustering Statistics Surveys, 4:80-116.

[4] Melnykov V, 2016. Merging mixture components for clustering through pairwise overlap Journal of Computational and Graphical Statistics, 25(1):66-90.

[5] Duda RO, Hart PE, Stork DG, 2000.Pattern Classification, 2nd ed. New York: John Wiley & Sons, Inc.

[6] Johnson RA, Wichern DW, 1998. Applied multivariate analysis, 4th ed.New Jersey: Prentice Hall, Englewood Cliffs.

[7] Baudry JP, Raftery AE, Celeux G, Lo K, Gottardo R, 2008. Combining Mixture Components for Clustering Technical Report 540, University of Washington, Seattle.

[8] Goldberger J, Roweis S, 2005. Hierarchical clustering of a mixture model. In Advances in Neural Information Processing Systems MIT Press, 17:505–512.

[9] El-Hanjouri MMR, Hamad BS, 2015. Using cluster analysis and discriminant analysis methods in classification with application on standart of living family in Palestinian Areas International Journal of Statistics, 5(5):213-222.

[10] Hastie T, Tibshirani R, 1996. Discriminant analysis by Gaussian mixtures Journal of the Royal Statistical Society, 5(1):155-176.

[11] Flury B W, Schmid M J, 1992. Quadratic discriminant functions with constraints on the covariance matrices some asymptotic results Journal of Multivariate Analysis 4:244-261.

[12] Ganesalingam S, Nanthakumar A, Ganesh S, 2011. An analytical expression for the error rate associated with the quadratic discriminant function Journal of Statistics and Management Systems, 14(6):1027-1040.

[13] Goswamiand S, Wegman EJ, 2016. Comparison of Different Classification Methods on Glass Identification for Forensic Research Journal of Statistical Science and Application, 4(3-4):65-84.

[14] Engelhardt A, Kanawade R, Knipfer C, Schmid M, Stelzle F, Adler W, 2014. Comparing classification methods for diffuse reflectance spectra to improve tissue specific laser surgery BMC Medical Research Methodology, 14(91):1-15.

[15] Cherry A, 1993. Combining cluster and discriminant analysis to develop a social bond typology of runaway youth Research on Social Work Practice, 3(2):175-190.

[16] Baudry JP, Raftery AE, Celeux G, Lo K, Gottardo R, 2010. Combining mixture components for clustering Journal of Computational and Graphical Statistics, 19(2):332-353.

[17] Tanos P, Kovacs J, Kovacs S, 2015. Optimization of the monitoring network on the River Tisza (Central Europe, Hungary) using combined cluster and discriminant analysis, taking seasonality into account Environmental Monitoring and Assessment DOI: 10.1007/s10661-015-4777-y.

[18] Morris K, McNicholas PD, 2016. Clustering, classification, discriminant analysis and dimension reduction via generalized hyperbolic mixtures Computational Statistics and Data Analysis, 97:133-150.

[19] Novak M, Palya D, Bodai Z, Nyiri Z, Magyar N, Kovacs J, Eke Z, 2017. Combined cluster and discriminant analysis: An efficient chemometric approach in diesel fuel characterization Forensic Science International 270:61-69.

[20] Kovacs J, Kovacs S, Magyar N, Tanos P, Hatvani IG, 2014. Classification into homogeneous groups using combined cluster and discriminant analysis Environmental Modelling & Software, 57: 52-59.

[21] Fisher RA, 1936. The use of multiple measurements in taxonomic problems, Annals of Eungenics, 7:179-188.

[22] McLachlan G, 2004. Discriminant analysis and statistical pattern recognition. 2th ed. John Wiley&Sons.

[23] Melnykov V, Chen WC, Maitra R, 2012. MixSim: An R package for simulating data to study performance of clustering algorithms Journal of Statistical Software, 51(12):1-25.

[24] Baudry JP, Raftery AE, Celeux G, Lo K, Gottardo R, 2010. Combining mixture components for clustering, Journal of Computational and Graphical Statistics, 19(2):332-353.