Covariance-based Clustering for Classification

Theophilus Anim Bediako

Abstract

Classification involves assigning observations to known classes based on some common features. Fisher's discriminant analysis commonly known as Linear and Quadratic Discriminant Analysis (LDA and QDA, respectively) are famous methods used to solve classification problems. Both LDA and QDA assume the given classes follow a multivariate normal distribution. In the case of LDA, the classes share a common covariance matrix which could be too restrictive. In the case of QDA, each class has its own covariance matrix, which could lead to a problem while estimation. Clustering methods search for natural groups of similar observations in data. Model-based clustering, particularly mixtures of Gaussian, remains one of the popular clustering methods where the search for classes relies on the assumption that samples within a cluster come from a specific normal distribution. This work proposes a covariance-based clustering method to build a classifier to overcome the simplicity of LDA and the complexity of QDA. For this, a finite mixture of Wishart is developed for clustering the cross-product matrices with similar structures. The parameter estimates of the mixtures are obtained through the Expectation Maximization (EM) algorithm. Initialization of parameters for the EM algorithm is proposed. Having identified similar covariance structures in the data, we proposed a method, cluster-based LDA, which uses the result of the clustering to build a classifier. The proposed method is studied through simulated data and applied to glass fragment classification.

Keywords:

1. Introduction

Commonly known methods for solving classification problems include linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). LDA and QDA assume a known number of classes and that new observations can be assigned to one of the known classes. The observations in each class are assumed to follow a multivariate normal, with each class having a class-specific mean vector. The assumption that covariance matrices are equal across classes makes LDA different from QDA [1]. LDA results in a linear classification boundary between classes [1]. QDA assumes Σ_k is unique to each class and that within-class observations follow a multivariate normal distribution [2]. Such a flexible assumption results in a non-linear (curve) classification boundary between classes. The equal covariance assumption in LDA is often described as almost unrealistic and simplistic in real-life applications, resulting in less flexibility and high bias. The problem with QDA is that it requires the estimation of a lot of parameters, making it computationally expensive. Estimating class-specific covariance matrices becomes a problem when there are more classes with few observations. This work proposes a method in the middle of LDA and QDA using mixtures of Wishart to be used in solving classification problems. In subsequent sections, we provide a general overview of Wishart distribution, Mixtures of Wishart, and derivations of the E and M steps of the EM-algorithm for fitting mixtures of Wishart.

2. Methods

2.1. Wishart Distribution

Let $\mathbf{X_j} = \left[\mathbf{x_1}, \cdots, \mathbf{x_{n_j}}\right]$ be p x n_j such that $\mathbf{x_1} \cdots, \mathbf{x_{n_j}}$ be independent $N_p(0, \Sigma_j)$. Then $\mathcal{S}_{\mathrm{j}} = \sum\limits_{\mathrm{i=1}}^{\mathrm{n_j}} \mathbf{x_i}\mathbf{x_i}^{\mathrm{T}} = \mathbf{X_j}\mathbf{X_j}^{\mathrm{T}} \in \Re^{p \text{ x } p}$ is said to follow a Wishart distribution with n_j degrees of freedom and a scale matrix, Σ_j. The probability density function of the Wishart is given by:

$$f(\mathcal{S}|\mathbf{\Sigma}, \mathbf{n}) = \frac{|\mathcal{S}|^{\frac{\mathbf{n-p-1}}{\mathbf{2}}} \mathbf{exp}\left\{-\frac{\mathbf{1}}{\mathbf{2}}\mathbf{tr}\left(\mathbf{\Sigma^{-1}}\mathcal{S}\right)\right\}}{\mathbf{2^{\frac{pn}{2}}}.|\mathbf{\Sigma}|^{\mathbf{n/2}}.\mathbf{\Gamma_p}\left(\frac{\mathbf{n}}{\mathbf{2}}\right)}$$

where $\Gamma_p(.)$ is the multivariate gamma function expressed as:

$$\Gamma_p\left(n/2\right) = \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \Gamma\left(\frac{n-j+1}{2}\right)$$

and Γ is the ordinary gamma function, $\Gamma(z) = \int\limits_0^{+\infty} t^{z-1}e^{-t}dt$, $z > 0$ [3]. S is a p x p positive definite. We assume the number of degrees of freedom, n is greater than p and that the elements of the random vector, X are independent [4]. The expectation of a Wishart random matrix, $\mathcal{S}$ is nΣ. The Wishart distribution is a multivariate extension of the univariate chi-square distribution applicable to a set of random matrices generated from a random Gaussian vector. In a univariate setting where p = 1 and Σ = 1, the pdf of the Wishart reduces to the pdf of a chi-square distribution with n degrees of freedom [3].

2.2. Wishart Mixture Model

Given a set of N positive definite random matrices of dimension p, $\mathfrak{S} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_N\}$, the Wishart mixture model is a convex combination of Wishart densities f_k's expressed as

$$f\left(\mathcal{S}_i; \Theta\right) = \sum_{k=1}^{K} \pi_k f_k\left(\mathcal{S}_i; \Sigma_k, n_k\right) \tag{1}$$

where π_k's are the mixing proportions which sum up to one and $\Theta = (\Sigma_1, \Sigma_2, \cdots, \Sigma_K; n_1, n_2, \cdots, n_K)$ are the unknown parameters.

The log-likelihood of (1) is seen as:

$$log(\ell(\Theta)) = \sum_{i=1}^{N} log\Big[\sum_{k=1}^{K} f_k(\mathcal{S}_i|\theta_k)\Big]$$

which is difficult to maximize for closed-form solutions. Here, the $\mathcal{S}_i$'s are regarded as incomplete. Alternatively, argument each $\mathcal{S}_i$ with a latent variable Z_i and regard $\{\mathcal{S},\, Z\}$ as complete data. The complete-data log-likelihood for the mixture model is therefore given by:

$$log\ell_c(\Theta) = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} log\big\{\pi_k f_k(\mathcal{S}_i; \Sigma_k, n_k)\big\} \tag{2}$$

where

$$z_{ik} = \begin{cases} 1 & \text{if } \mathcal{S}_i \in \text{class k} \\ 0 & \text{otherwise} \end{cases}, \qquad \text{and each } \mathbf{z_i} \sim Mult_g\left(1, \pi_1, \cdots, \pi_g\right).$$

2.3. Expectation-Maximization

The EM algorithm is an iterative two-step process for fitting mixture models. Maximizing (2) is not straightforward. We take the expectation of equation (2) instead, which is mostly defined as a Q function:

$$\begin{aligned} Q\left(\Theta, \hat{\Theta}\right) &= E_{\hat{\Theta}}\left(\ell_c\left(\Theta|\mathfrak{S}\right)\right) \\ &= E_{\hat{\Theta}}\Big(\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik} log\big\{\pi_k f(\mathcal{S}_i; \Sigma_k, n_k)\big\}\Big) \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} E\big(z_{ik}\big) log\big\{\pi_k f(\mathcal{S}_i; \Sigma_k, n_k)\big\} \end{aligned} \tag{3}$$

$$\text{Set } \tau_{ik} = E(z_{ik}|\mathfrak{S}, \Theta^{t-1})$$

E step : Involves computing the conditional expectation of complete-data log-likelihood given data and parameter estimates, $\hat{\Theta}$ [5].

$$E(z_{ik}|\mathfrak{S}, \Theta^{t-1}) = \tau_{ik}^{(t)} = \frac{\pi_k^{(t)} f_k\left(\mathcal{S}_i|\Sigma_k^{(t)}, n_k^{(t)}\right)}{\sum_{l=1}^{K} \pi_l^{(t)} f_k\left(\mathcal{S}_i|\Sigma_l^{(t)}, n_l^{(t)}\right)}$$

E step gives an estimate of the posterior probability that $\mathcal{S}_i$ is in the kth cluster.

M step : Involves maximizing the conditional expectation of (2) w.r.t π_k and Θ_k given the data to find new estimates for $\hat{\Theta}^{(t+1)}$ [5]. This is equivalent to taking the partial derivatives of 3 w.r.t to parameters

$$\begin{aligned}
\frac{\partial Q}{\partial \Sigma_k} &= \sum_{i=1}^{N} E\big(z_{ik}\big)\Big\{\frac{\partial}{\partial \Sigma_k} log\big(f_k(\mathfrak{S}|\Theta_k)\big)\Big\} \\
&= \sum_{i=1}^{N} \tau_{ik} \frac{\partial}{\partial \Sigma_k}\Big\{\frac{n_k - p - 1}{2} log|\mathcal{S}| - \frac{1}{2} tr\big(\Sigma_k^{-1}\mathcal{S}\big) \\
&\qquad - \frac{pn_k}{2} log(2) - \frac{n_k}{2} log|\Sigma_k| - log\big(\Gamma_p(n_k/2)\big)\Big\} \\
&= \sum_{i=1}^{N} \tau_{ik} \frac{\partial}{\partial \Sigma_k}\Big\{-\frac{1}{2} tr\big(\Sigma_k^{-1}\mathcal{S}\big) - \frac{n_k}{2} log|\Sigma_k|\Big\} \\
&\implies \sum_{i=1}^{N} \tau_{ik}\Big\{-\frac{1}{2}\Sigma_k^{-1}\mathcal{S}_i\Sigma_k^{-1} - \frac{n_k}{2}\Sigma_k^{-1}\Big\} = 0
\end{aligned}$$

Pre and post multiply by Σ_k

$$\sum_{i=1}^{N} \tau_{ik}\mathcal{S}_i - \Sigma_k \sum_{i=1}^{N} \tau_{ik} n_k = 0$$

$$\hat{\Sigma}_k = \frac{\sum\limits_{i=1}^{N} \tau_{ik}\mathcal{S}_i}{\sum\limits_{l=1}^{N} \tau_{lk} n_k}$$

New estimates for $\hat{\pi}_k^{(t+1)}$ and $\Sigma_k^{(t+1)}$ are obtained as:

$$\hat{\pi}_k^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} \hat{\tau}_{ik}^{(t)} \qquad\qquad \hat{\Sigma}_k^{(t+1)} = \frac{\sum\limits_{i=1}^{N} \hat{\tau}_{ik}^{(t)}\mathcal{S}_i}{\sum\limits_{l=1}^{N} \hat{\tau}_{lk}^{(t)} n_k}$$

Repeat steps until some pre-specified convergence criterion is met. Here, we adopt the relative difference convergence criterion given by $\frac{\ell(\Theta^{(t+1)};\mathcal{S})-\ell(\Theta^{(t)};\mathcal{S})}{|\ell(\Theta^{(t)};\mathcal{S})|} < \Delta$. Δ is the pre-specified threshold value and it is usually greater than 0.
We often solve for n_k through numerical optimization when n_k is unknown [5]. Here, we assume n_k is known and equal for all clusters.

2.4. Proposed method

This paper proposes a method in the middle of LDA and QDA. The method relies on the covariance-based clustering method to identify clusters with similar covariance structures. The proposed method, cluster-based LDA, is fitted to observations identified as having a similar covariance structure. The method is compared to the ordinary LDA and QDA via their prediction accuracies. To apply the proposed method to real-life problems, we also develop an initialization algorithm for fitting mixtures of Wishart.

Initializing the EM-algorithm in the case of simulated data is usually straightforward. However, this is not the case in practice. For instance, it is unrealistic to predefine the number of components when dealing with real-life cases. Finding the initial parameters and the number of components is equivalent to a model selection problem [6]. Model selection methods frequently used for mixture models include Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) [7]. To the best of our knowledge, there are several initialization methods for EM algorithm. However, these methods are predominantly used in the case of Gaussian mixture modeling. We propose an initialization method for Wishart mixture modeling. The proposed method is similar to the Rnd-EM algorithm used in Gaussian mixture modeling.

The initialization method is combined with BIC to help choose the number of components. For mixtures of Wishart,

$$BIC = -2log(L(\hat{\Theta}) + Mlog(N), \text{ M is the number of parameters in the model}$$
$$M = K - 1 + K * \Big(\frac{p(p+1)}{2} + 1\Big) \quad \text{set 1 to 0 when df is known}$$

```
Algorithm 1: Initialization Algorithm for Mixtures of Wishart
  Input: $K$ & $\mathfrak{S} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_N\}$
  Output: $\Theta^{(0)} = \{\Sigma_1, \Sigma_2, \cdots, \Sigma_K, \pi_1, \pi_2, \cdots, \pi_K\}$
  Draw $K$ samples from $\mathfrak{S}$. Assume the samples represent the cluster centers
  Compute Frobenius norm between the cluster centers and each $\mathcal{S}_i$
    $||\mathcal{S}_i - \mathcal{S}_k|| = \sqrt{\text{trace}\big[(\mathcal{S}_i - \mathcal{S}_k)(\mathcal{S}_i - \mathcal{S}_k)^\top\big]} \quad for\ i = 1 \cdots, N \quad \& \quad k = 1, 2, \cdots, \quad K$
  Form a hard partitioning, $\tau_{ik}$, of observations close to a cluster center. $\tau_{ik}$ is an n by K
    matrix of ones and zeros
  Based on the hard partitioning, compute parameter estimates
    $\pi_k = \frac{1}{N}\sum_{i=1}^{N} \tau_{ik}$
    $\Sigma_k = \frac{\sum_{i=1}^{N} \tau_{ik}\mathcal{S}_i}{\sum_{j=1}^{N} \tau_{ik} n_k}$   $n_k$ is the degree of freedom associated with the Wishart
  Compute loglikelihood using estimates in step 4
  Repeat steps $1 - 5$ a large number of times
  Return a proposed set of parameters, $\Theta^{(0)}$, that results in the highest loglikelihood
```

Algorithm 2: Cluster-Based LDA

```
1 Clustering: K component mixture
  Input: $\mathfrak{S} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_N\}$
  Output: $\{\hat{Z}_1, \cdots, \hat{Z}_N, \hat{\Sigma}_1, \cdots, \hat{\Sigma}_K, \hat{\pi}_1, \cdots, \hat{\pi_K}\}$
    (i) Initialize: $\Theta^{(0)} = \{\Sigma_1^0, \cdots, \Sigma_K^0, \pi_1^0, \cdots, \pi_K^0\}$
    (ii) Estep: $\tau_{ik}^{(t)}$ for $i = 1, \cdots, N$ & $k = 1, \cdots, K$
    (iii) Mstep: $\Theta^{(t)} = \{\Sigma_1^t, \cdots, \Sigma_K^t, \pi_1^t, \cdots, \pi_K^t\}$
    (iv) Repeat (ii) and (iii) until $\frac{\ell(\Theta^{(t+1)};\mathcal{S}) - \ell(\Theta^{(t)};\mathcal{S})}{|\ell(\Theta^{(t)};\mathcal{S})|} < \Delta$, for some small $\Delta > 0$
    (v) Obtain $\hat{\Theta} = \{\hat{\Sigma}_1, \cdots, \hat{\Sigma}_K, \hat{\pi}_1, \cdots, \hat{\pi_K}\}$ & $\hat{Z}_i = argmax(\tau_{ik})$ for $i = 1, \cdots, N$
2 Classification $G$ classes Input:
    Class labels $\mathbf{y_{ji}}$,   $\mathbf{j = 1}, \cdots, \mathbf{G}$ & $i = 1, \cdots, n_j$
    Predictors   $\mathbf{x_{ji}}$
    Cluster labels   $\{\hat{Z}_1, \cdots, \hat{Z}_N\}$
    Covariances   $\{\hat{\Sigma}_1, \cdots, \hat{\Sigma}_K\}$
  Output: $j = argmax\big(P(g = j|\mathbf{x}^*, \delta_{\mathbf{j}}\big)$
  for $k = 1, \cdots, K$
    (i) use $\hat{\Sigma}_k$ and compute discriminate functions $\delta_1, \cdots, \delta_G$ , where
        $\delta_g(\mathbf{x}^*) = \mathbf{x}^T \hat{\Sigma}_{k_g} \mu_g - 1/2 \mu_g^T \hat{\Sigma}_{k_g} \mu_g + log \pi_g$
    (ii) Given $X^*$, compute $\tau_j = P\big(g = j|X^*, \delta_j\big)$
    (iii) Return $j = argmax\big(P(g = j|X^*, \delta_j\big)$
```

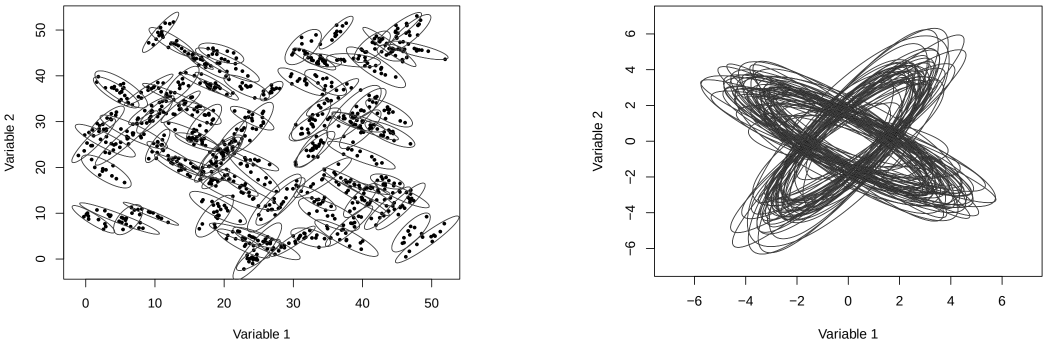
3. Simulation Study

3.1. Illustrative Example 1

We simulate 100 two-dimensional Gaussian random variables with 100 unique means. Each $\mu_{2 \text{ x } 1}$ has its entries simulated from $unif(0, 50)$. The mixing proportions are 0.45 and 0.55 with corresponding covariance matrices as:

$$\Sigma_1 = \begin{bmatrix} 1.42 & 1.57 \\ 1.57 & 2.53 \end{bmatrix} \qquad \Sigma_2 = \begin{bmatrix} 2.39 & -1.61 \\ -1.61 & 1.57 \end{bmatrix}$$

The generated 100 class two-dimensional data with 10 observations per class and 95% confidence ellipsoids are shown below:



(a) Scatterplot of points with 95% confidence ellipsoid for each class before mean centering

(b) 95% confidence ellipses based on the sample covariance matrix of each class centered at zero

Figure 1: Class Confidence ellipsoid before and after mean centering

Mean center the class observations and transform to cross-product matrices via $\mathcal{S} = \mathrm{XX}^{\mathrm{T}}$. Each $\mathcal{S}$ follows a Wishart distribution with $n = 10$ degrees of freedom and a scale matrix Σ_k. Before the transformation, Figure 1a shows the observations for each class and their respective 95% confidence ellipsoid. Each $\mathcal{S}$ follows a Wishart distribution with 10 degrees of freedom and a scale matrix Σ_k. For all classes, we obtain $\mathfrak{S} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_{100}\}$. Figure 1b shows the structure of the sample covariances after meaning centering.

We apply the EM algorithm for fitting mixtures of Wishart to the 100 cross-product matrices. EM algorithm provides the following parameter estimates.

$$\hat{\Sigma}_1 = \begin{bmatrix} 1.42 & 1.57 \\ 1.57 & 2.53 \end{bmatrix} \qquad \hat{\Sigma}_2 = \begin{bmatrix} 2.39 & -1.61 \\ -1.61 & 1.57 \end{bmatrix} \qquad \hat{\pi}_1 = 0.45 \qquad \hat{\pi}_2 = 0.55$$

We also obtain a hard classification of the covariance structures using the maximum a-posterior criterion on the posterior probabilities in the Estep. No element of $\mathfrak{S}$ is misclassified. Here, 45 classes were clustered as having a covariance structure with the major axis pointed to the right. The rest of the 55 classes were identified as having a similar covariance structure with the major axis pointed to the right [see Figure 1].

To test our proposed method, we fit a cluster-based LDA to each set of classes the EM algorithm identified as having a similar covariance structure. Table 1 provides the number of parameters to be estimated, training accuracy, test accuracy, and leave-one-out cross-validation(LOOCV) for the proposed cluster-based LDA compared with ordinary LDA and QDA.

Table 1: Accuracy & Estimated Parameters for the classification methods

Classification	**Training**	**Test**	**LOOCV**	**Parameters**
LDA	68.9%	64.0%	65.0%	203
Cluster-based LDA	86.8%	85.0%	83.3%	207
QDA	82.0%	68.8%	71.9%	500

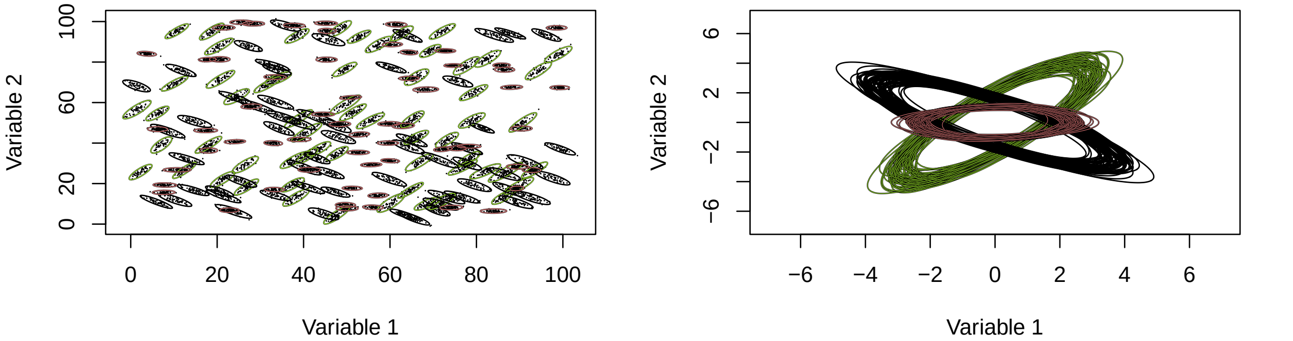
3.2. Illustrative Example 2

Next, we consider a 180-class problem with the following scale matrices and mixing proportions.

$$\Sigma_1 = \begin{bmatrix} 1.42 & 1.57 \\ 1.57 & 2.53 \end{bmatrix} \quad \Sigma_2 = \begin{bmatrix} 2.39 & -1.61 \\ -1.61 & 1.57 \end{bmatrix} \quad \Sigma_3 = \begin{bmatrix} 1.00 & 0 \\ 0 & 0.20 \end{bmatrix} \quad \pi_1 = \frac{1}{3} \quad \pi_2 = \frac{1}{3} \quad \pi_3 = \frac{1}{3}$$

μ_{2x1} of each of the 180 classes was obtained from $Unif(0, 100)$

Figure 2a shows the classes of the data with each class having 50 observations. By similar transformation of the within-class observations, each $\mathcal{S}$ follows a Wishart distribution with 50 degrees of freedom.



(a) Scatterplot of points with 95% confidence ellipsoid for each class before mean centering

(b) 95% confidence ellipses based on the sample covariance matrix of each class centered at zero

Figure 2: Class Confidence ellipsoid before and after mean centering

EM-algorithm(initialized with true parameters) on $\mathfrak{S} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_{180}\}$ provides the following estimates:

$$\hat{\Sigma}_1 = \begin{bmatrix} 1.46 & 1.59 \\ 1.59 & 2.55 \end{bmatrix} \qquad \hat{\Sigma}_2 = \begin{bmatrix} 2.28 & -1.56 \\ -1.56 & 1.55 \end{bmatrix} \qquad \hat{\Sigma}_3 = \begin{bmatrix} 0.97 & 0.01 \\ 0.01 & 0.20 \end{bmatrix}$$

$$\hat{\pi}_1 = 0.33333 \qquad \hat{\pi}_2 = 0.33333 \qquad \hat{\pi}_3 = 0.33333$$

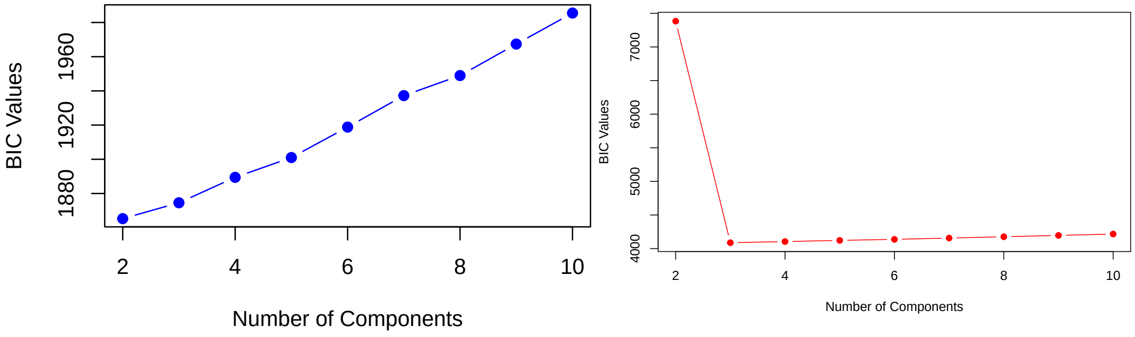
No element of $\mathfrak{S}$ is misclassified. Subset classes based on similar covariance structures from the EM algorithm output and perform cluster-based LDA. 60 classes form one cluster. We have three clusters based on the results from the EM algorithm. Table 2 shows the training accuracy, the test accuracy(35 observations training and 15 observations testing per class), leave-one-out cross-validation, and the number of parameters to be estimated.

Table 2: Accuracy & Estimated Parameters for the classification methods

Classification	**Training**	**Test**	**LOOCV**	**Parameters**
LDA	87.9%	88.0%	87.6%	363
Cluster-based LDA	98.3%	98.5%	98.3 %	367
QDA	92.8%	92.7%	92.2%	900

3.3. Test Proposed Initialization

We test the proposed initialization method[see Algorithm 1] on the simulated data. The proposed method correctly identifies the number of components for the two simulated data



(a) Finding K for $\mathfrak{S} = \{\mathcal{S}_1, \cdots, \mathcal{S}_{100}\}$

(b) Finding K for $\mathfrak{S} = \{\mathcal{S}_1, \cdots, \mathcal{S}_{180}\}$

Figure 3: Model selection using BIC

Initial parameters for the EM algorithm using the proposed initialization method

For K = 2,

$$\hat{\Sigma}_1 = \begin{bmatrix} 1.44 & 0.88 \\ 0.88 & 2.13 \end{bmatrix} \qquad \hat{\Sigma}_2 = \begin{bmatrix} 2.78 & -1.90 \\ -1.90 & 1.76 \end{bmatrix} \qquad \hat{\pi}_1 = 0.65 \qquad \hat{\pi}_2 = 0.35$$

For K = 3,

$$\hat{\Sigma}_1 = \begin{bmatrix} 1.31 & 0.92 \\ 0.92 & 1.60 \end{bmatrix} \qquad \hat{\Sigma}_2 = \begin{bmatrix} 2.31 & -1.58 \\ -1.58 & 1.57 \end{bmatrix} \qquad \hat{\Sigma}_3 = \begin{bmatrix} 0.76 & 0.01 \\ 0.01 & 0.20 \end{bmatrix}$$

$$\hat{\pi}_1 = 0.57 \qquad \hat{\pi}_2 = 0.32 \qquad \hat{\pi}_3 = 0.11$$

4. Application

The glass data is from the Institute of Forensic Research in Krakow, Poland [8]. The data has nine variables and 2400 observations. The item variable is the main classification variable of interest. The data has 200 unique items. Each item has four fragments measured three times to form the fragment variable. The rest of the variables represent the glass fragments' elemental compositions. The elemental compositions are log(NaO), log(MgO), log(AlO), log(SiO), log(KO), log(CaO), log(CaO) and log(FeO). For instance, log(NaO) = log_{10}(Na/O).

Considering the 200 unique items(windows), we obtain 200 7x7 cross-product matrices. Given $\mathfrak{S} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_{200}\}$, four observations are found not to be positive definite, leaving 196 cross-product matrices such that $\mathcal{S}_i | Z_i = k \sim W(n, \Sigma_k)$, where $n = 12$. We adopt the proposed initialization method and the BIC model selection approach to choose the number of components for the mixtures of Wishart. Next, we subset observations based on unique ID from the results of the cluster analysis and fit cluster-based LDA. The results of the cluster-based LDA are compared to the ordinary LDA and QDA in Table 3 using different re-sampling methods. 70%(8 observations) training and 30%(4 observations) testing per class.

5. Discussion

Based on the results of this study, we proposed a classification method using cluster analysis. We compared our proposed method with traditional LDA and QDA methods and evaluated the performance using different re-sampling methods. The study also proposed an initialization method for mixtures of Wishart. While our proposed method seems to be performing better than traditional methods in the context of simulated data, the rank deficiency problem encountered in the case of the glass data made it difficult to see how the method compares with QDA in application settings

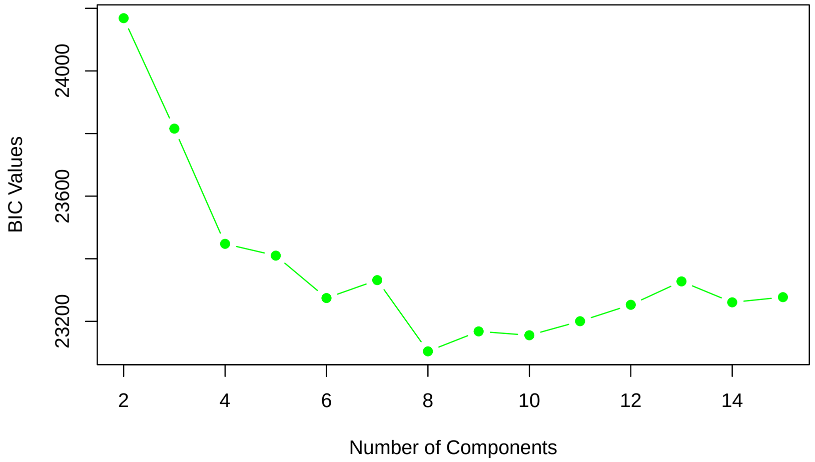


Figure 4: Model selection for glass data

Table 3: Accuracy & Estimated Parameters for the classification methods

Classification	**Training**	**Test**	**LOOCV**	**Parameters**
LDA	42.0%	36.4%	37.5%	1400
Cluster-based LDA	79.3%	73.9%	74.4%	1603
QDA	91.9%	–	59.4%	6860

References

[1] J. Gareth, W. Daniela, H. Trevor, T. Robert, An introduction to statistical learning: with applications in R, Spinger, 2013.

[2] R. Johnson, D. Wichern, Applied multivariate statistical analysis, Prentice Hall, 2002.

[3] A. G. Wilson, Z. Ghahramani, Generalised wishart processes, arXiv preprint arXiv:1101.0240 (2010).

[4] Y.-K. Yu, Y.-C. Zhang, On the anti-wishart distribution, Physica A: Statistical Mechanics and its Applications 312 (1-2) (2002) 1–22.

[5] S. Hidot, C. Saint-Jean, An expectation–maximization algorithm for the wishart mixture model: Application to movement clustering, Pattern Recognition Letters 31 (14) (2010) 2318–2324.

[6] Z. Hu, Initializing the EM algorithm for data clustering and sub-population detection, The Ohio State University, 2015.

[7] G. Celeux, S. Frühwirth-Schnatter, C. P. Robert, Model selection for mixture models–perspectives and strategies, in: Handbook of mixture analysis, Chapman and Hall/CRC, 2019, pp. 117–154.

[8] C. G. Aitken, G. Zadora, D. Lucy, A two-level model for evidence evaluation, Journal of forensic sciences 52 (2) (2007) 412–419.

[9] C. Fraley, A. E. Raftery, Model-based clustering, discriminant analysis, and density estimation, Journal of the American statistical Association 97 (458) (2002) 611–631.

[10] K. P. Murphy, Machine learning: a probabilistic perspective, MIT press, 2012.

[11] V. Melnykov, R. Maitra, Finite mixture models and model-based clustering, Statistics Surveys 4 (none) (2010) 80 – 116.

[12] G. J. McLachlan, S. X. Lee, S. I. Rathnayake, Finite mixture models, Annual review of statistics and its application 6 (2019) 355–378.

[13] S. Goswami, E. J. Wegman, Comparison of different classification methods on glass identification for forensic research, J. Stat. Sci. App 4 (2016) 65–84.

[14] S. Michael, V. Melnykov, An effective strategy for initializing the em algorithm in finite mixture models, Advances in Data Analysis and Classification 10 (2016) 563–583.

[15] R. A. Horn, C. R. Johnson, Matrix analysis, Cambridge university press, 2012.

[16] C. M. Bishop, N. M. Nasrabadi, Pattern recognition and machine learning, Vol. 4, Springer, 2006.