

# Covariance-Based Clustering for Classification

Theophilus Anim Bediako  
Andrew Simpson  
Semhar Michael, PhD

South Dakota State University, Department of Mathematics & Statistics  
Brookings, SD

April 28, 2023

# Table of Contents

- 
- 1 Introduction
- 2 Methodology
- 3 Simulation Study
- 4 Application
- 5 Conclusion
- 6 Reference
- 

## Classification & Clustering

**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 classification methods

- 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.
- 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.

# Objective

We propose a covariance-based clustering method that can identify similar covariance structures in the data.

- Our method involves using a finite mixture of Wishart distributions to cluster the cross-product matrices
- The parameter estimates of the mixtures are obtained through the Expectation Maximization (EM) algorithm.
- Identify observations with similar covariance structure and build a classifier using cluster-based LDA
- We also propose an initialization method for the parameters of the EM algorithm

# Table of Contents

- 
- 1 Introduction
- 2 Methodology
- 3 Simulation Study
- 4 Application
- 5 Conclusion
- 6 Reference
- SOUTH DAKOTA**  
**STATE UNIVERSITY**

## Wishart Distribution

Let  $\mathbf{X}_j = [\mathbf{x}_1, \dots, \mathbf{x}_{n_j}]$  such that  $\mathbf{x}_1, \dots, \mathbf{x}_{n_j}$  be independent  $N_p(0, \Sigma_j)$ .

Then  $\mathcal{S}_j = \sum_{i=1}^{n_j} \mathbf{x}_i \mathbf{x}_i^T = \mathbf{X}_j \mathbf{X}_j^T \in \Re^{p \times p}$  is said to follow a Wishart distribution with  $n_j$  degrees of freedom and a scale matrix,  $\Sigma_j$

# Wishart Distribution

## Probability Density Function

The probability density function (pdf) of the Wishart is given by:

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

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

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

$\Gamma$  is the ordinary gamma function,  $\Gamma(z) = \int_0^{+\infty} t^{z-1} e^{-t} dt, z > 0$



# Wishart Distribution

## Points to Note

- $S$  is a  $p \times p$  positive definite
- $n$  is greater than  $p$  and that the elements of the data matrix,  $X$  are independent
- The expectation of a Wishart random matrix,  $S$  is  $n\Sigma$
- 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

## Definition

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

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

$\Theta = (\Sigma_1, \Sigma_2, \dots, \Sigma_K; n_1, n_2, \dots, n_K)$  are the unknown parameters

$\pi_k$ 's are the mixing proportions,  $\sum_{k=1}^K \pi_k = 1$

# Loglikelihood function

The log-likelihood of (1) is

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

Difficult to maximize for closed-form solution

Alternatively, argument each  $\mathcal{S}_i$  with a latent variable  $Z_i$  and regard  $\{\mathcal{S}, Z\}$  as complete data.

# Complete Data Log-likelihood

The complete-data log-likelihood for the mixture model is given by:

$$\log \ell_c(\Theta) = \sum_{i=1}^N \sum_{k=1}^K z_{ik} \log \{ \pi_k f(S_i; \Sigma_k, n_k) \} \quad (2)$$

$$z_{ik} = \begin{cases} 1 & \text{if } S_i \in \text{class } k \\ 0 & \text{otherwise} \end{cases}, \quad \text{and each } \mathbf{z}_i \sim \text{Mult}_g(1, \pi_1, \dots, \pi_K).$$

# Expectation-Maximization Algorithm

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:

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

# Expectation-Maximization Algorithm

## E Step

Involves computing the conditional expectation of complete-data log-likelihood given data and parameter estimates,  $\hat{\Theta}$

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

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

## M step

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

$$\pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^N \hat{\tau}_{ik}^{(t)} \quad \Sigma_k^{(t+1)} = \frac{\sum_{i=1}^N \hat{\tau}_{ik}^{(t)} \mathcal{S}_i}{\sum_{l=1}^N \hat{\tau}_{lk}^{(t)} n_k}$$

Here, we assume  $n_k$  is known and equal for all clusters

Repeat steps until some pre-specified convergence criterion is met. Here, we adopt the relative difference convergence criterion given by

$$\frac{\ell(\Theta^{(t+1)}; \mathcal{S}) - \ell(\Theta^{(t)}; \mathcal{S})}{|\ell(\Theta^{(t)}; \mathcal{S})|} < \Delta, \text{ for some small } \Delta > 0$$

---

## Algorithm 1: Cluster-Based LDA

---

### 1 Clustering: K component mixture

**Input:**  $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_N\}$

**Output:**  $\{\hat{Z}_1, \dots, \hat{Z}_N, \hat{\Sigma}_1, \dots, \hat{\Sigma}_K, \hat{\pi}_1, \dots, \hat{\pi}_K\}$

- (i) Initialize:  $\Theta^{(0)} = \{\Sigma_1^0, \dots, \Sigma_K^0, \pi_1^0, \dots, \pi_K^0\}$
- (ii) Estep:  $\tau_{ik}^{(t)}$  for  $i = 1, \dots, N$  &  $k = 1, \dots, K$
- (iii) Mstep:  $\Theta^{(t)} = \{\Sigma_1^t, \dots, \Sigma_K^t, \pi_1^t, \dots, \pi_K^t\}$
- (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  $\hat{\Theta} = \{\hat{\Sigma}_1, \dots, \hat{\Sigma}_K, \hat{\pi}_1, \dots, \hat{\pi}_K\}$  &  $\hat{Z}_i = \operatorname{argmax}(\tau_{ik})$  for  $i = 1, \dots, N$



## 2 Classification $G$ classes **Input:**

Class labels  $\mathbf{y}_{ji}$ ,  $\mathbf{j} = 1, \dots, \mathbf{G}$  &  $i = 1, \dots, n_j$

Predictors  $\mathbf{x}_{ji}$

Cluster labels  $\{\hat{Z}_1, \dots, \hat{Z}_N\}$

Covariances  $\{\hat{\Sigma}_1, \dots, \hat{\Sigma}_K\}$

**Output:**  $j = \operatorname{argmax}(P(g = j | \mathbf{x}^*, \delta_j))$

for  $k = 1, \dots, K$

- (i) use  $\hat{\Sigma}_k$  and compute discriminate functions  $\delta_1, \dots, \delta_G$ , 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  $\tau_j = P(g = j | X^*, \delta_j)$
- (iii) Return  $j = \operatorname{argmax}(P(g = j | X^*, \delta_j))$

# Parameters

Let's look at the number of parameters to estimate in each method

Table 1: Parameters to Estimate

Classification Method	Parameters
LDA	$Gp + \frac{p(p+1)}{2}$
Cluster-based LDA	$Gp + K \frac{p(p+1)}{2} + (K - 1)$
QDA	$Gp + G \frac{p(p+1)}{2}$

$G$  denotes the number of classes

$K$  denotes the number of clusters(components) for the mixtures of Wishart

# Initialization

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

We propose an initialization method for mixtures of Wishart

---

## Algorithm 2: Initialization Algorithm for Mixtures of Wishart

---

**Input:**  $K$  &  $\mathfrak{S} = \{S_1, S_2, \dots, S_N\}$

**Output:**  $\Theta^{(0)} = \{\Sigma_1^0, \dots, \Sigma_K^0, \pi_1^0, \dots, \pi_K^0\}$

- 1 Draw  $K$  samples from  $\mathfrak{S}$  to represent the cluster centers
- 2 Compute Frobenius norm between the cluster centers and each  $S_i$
- 3  $\|S_i - S_k\| = \sqrt{\text{trace}[(S_i - S_k)(S_i - S_k)^\top]}$  for  $i = 1 \dots, N$  &  
 $k = 1, 2, \dots, K$
- 4 Form a hard partitioning of observations close to a cluster center,  $\tau_{ik}$ ,  $\tau_{ik}$  is an  $n$  by  $K$  matrix of ones and zeros

# Initialization

- 5 Based on the hard partitioning, compute parameter estimates

$$\hat{\pi}_k = \frac{1}{N} \sum_{i=1}^N \tau_{ik} \quad \hat{\Sigma}_k = \frac{\sum_{i=1}^N \tau_{ik} \mathcal{S}_i}{\sum_{i=1}^N \tau_{ik} n_k}$$

- 6 Compute loglikelihood using estimates in step 4  
7 Repeat steps 1 – 5 a large number of times  
8 Return a proposed set of parameters,  $\Theta^{(0)}$ , that results in the highest loglikelihood

# Initialization

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

## Model Selection


$$\text{BIC} = -2\log(L(\hat{\Theta})) + M\log(N)$$

M is the number of parameters in the model

$$M = K - 1 + K * \left( \frac{p(p+1)}{2} + 1 \right)$$

Change 1 to 0 when  $n_k$  is known

# Table of Contents

- 
- 1 Introduction
- 2 Methodology
- 3 **Simulation Study**
- 4 Application
- 5 Conclusion
- 6 Reference
- SOUTH DAKOTA**  
**STATE UNIVERSITY**

# Illustrative Example 1

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

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

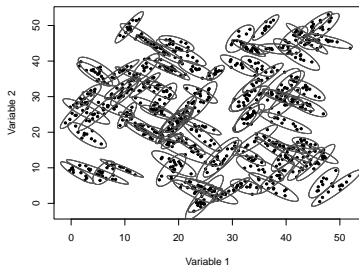
Mean center the class observations and transform to cross-product matrices via  $\mathcal{S} = \mathbf{X}\mathbf{X}^T$ . Each  $\mathcal{S}$  follows a Wishart distribution with  $n = 10$  degrees of freedom and a scale matrix  $\Sigma_k$ .

For all classes, we obtain  $\mathfrak{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_{100}\}$

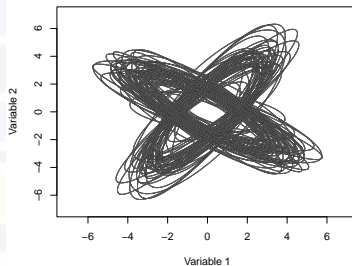


# Illustrative Example 1

The generated 100 classes with  $n = 10$  per class and 95% confidence ellipsoids



(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

# Illustrative Example 1

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

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

We also obtain a soft classification of the covariance structures using the maximum a-posteri criterion on the posterior probabilities in the Estep. No element of  $\mathcal{G}$  is misclassified.

# Illustrative Example 1

45 classes were clustered as having a similar covariance structure with the major axis pointed to the right

55 classes were identified as having a similar covariance structure with the major axis pointed to the left

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

## Illustrative Example 1 Results

Table 2: 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

We split the data into a 70%(7 observations) training set and a 30%(3 observations) testing set per class

## Illustrative Example 2

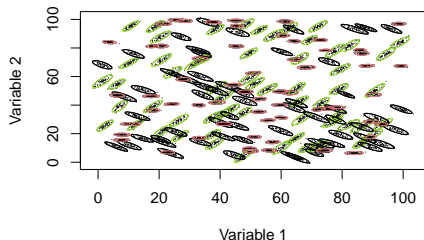
Using the following scale matrices and mixing proportions, simulate 180-class problem with unique  $\mu_{2 \times 1}$  per class. The mean of each variable follows an  $Unif(0, 100)$

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

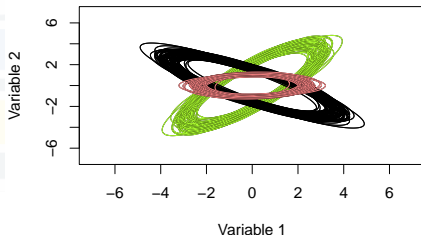
# Illustrative Example 2

The generated 180 classes with  $n = 50$  per class and 95% confidence ellipsoids

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

## Illustrative Example 2

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

$$\begin{aligned}\hat{\Sigma}_1 &= \begin{bmatrix} 1.46 & 1.59 \\ 1.59 & 2.55 \end{bmatrix} & \hat{\Sigma}_2 &= \begin{bmatrix} 2.28 & -1.56 \\ -1.56 & 1.55 \end{bmatrix} & \hat{\Sigma}_3 &= \begin{bmatrix} 0.97 & 0.01 \\ 0.01 & 0.20 \end{bmatrix} \\ \hat{\pi}_1 &= 0.33333 & \hat{\pi}_2 &= 0.33333 & \hat{\pi}_3 &= 0.33333\end{aligned}$$

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.

## Illustrative Example 2 Results

**Table 3:** Accuracy & Estimated Parameters for the classification methods

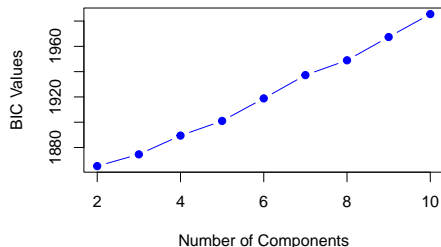
Classification	Training	Test	LOOCV	Parameters
LDA	87.9%	88.0%	87.6%	363
Cluster-based LDA	98.3%	98.5%	98.3 %	367
QDA	92.8%	92.7%	92.2%	900

70%(35 observations) training and 30%(15 observations) testing per class

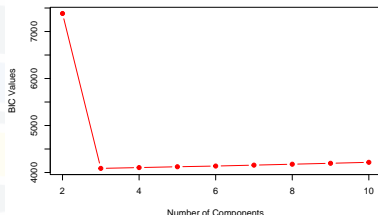


# Test Proposed Initialization

We test the method on the simulated data. The proposed method correctly identifies the number of components for the two simulated data



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



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

Figure 3: Model selection using BIC

# Test Proposed Initialization

Initial parameters for the EM algorithm using the proposed initialization method

$K = 2$

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

$K = 3$

$$\hat{\Sigma}_1 = \begin{bmatrix} 1.31 & 0.92 \\ 0.92 & 1.60 \end{bmatrix} \quad \hat{\Sigma}_2 = \begin{bmatrix} 2.31 & -1.58 \\ -1.58 & 1.57 \end{bmatrix} \quad \hat{\Sigma}_3 = \begin{bmatrix} 0.76 & 0.01 \\ 0.01 & 0.20 \end{bmatrix}$$
$$\hat{\pi}_1 = 0.57 \quad \hat{\pi}_2 = 0.32 \quad \hat{\pi}_3 = 0.11$$

# Table of Contents

- 
- 1 Introduction
  - 2 Methodology
  - 3 Simulation Study
  - 4 Application**
  - 5 Conclusion
  - 6 Reference

In forensic investigations, a common problem forensic scientists encounter is whether glass fragments obtained from a suspect or crime scene originated from a specific window

Examining chemical properties in glass fragments is an alternative for forensic analysts. We test the proposed method on the glass data and compare the performance to existing classification methods.

# Data Description

The glass data is from the Institute of Forensic Research in Krakow, Poland

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

- $\log(\text{NaO})$
- $\log(\text{MgO})$
- $\log(\text{AlO})$
- $\log(\text{SiO})$
- $\log(\text{KO})$
- $\log(\text{CaO})$
- $\log(\text{FeO})$

For instance,  $\log(\text{NaO}) = \log_{10}(\frac{\text{Na}}{\text{O}})$

# Proposed Method on the Glass Data

Considering the 200 unique items(windows), we obtain 200  $7 \times 7$  cross-product matrices. Given  $\mathfrak{S} = \{\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_{200}\}$ , 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

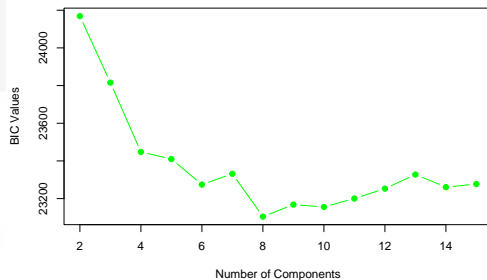


Figure 4: Model selection for glass data

# Results

Next, we subset observations based on unique ID from the results of the cluster analysis and fit cluster-based LDA.

**Table 4:** 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

70%(8 observations) training and 30%(4 observations) testing per class

# Table of Contents

- 
- 1 Introduction
  - 2 Methodology
  - 3 Simulation Study
  - 4 Application
  - 5 Conclusion**
  - 6 Reference



# Conclusion

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

# Future Work

This provides grounds for further investigation into the proposed cluster-based LDA method

The proposed initialization method presents the opportunity to examine the theoretical justification and appropriateness of such a method

# Acknowledgement

I would like to express my sincerest gratitude to Professor Christopher Saunders of the Department of Mathematics & Statistics, SDSU for his invaluable contributions to this research project. His constructive feedback is deeply appreciated

# Table of Contents

- 
- A large, stylized logo of South Dakota State University (SDSU) is centered in the background. It features a large 'S' and 'D' in light blue and yellow, with 'SOUTH DAKOTA' and 'STATE UNIVERSITY' written in a smaller font below it.
- 1 Introduction
  - 2 Methodology
  - 3 Simulation Study
  - 4 Application
  - 5 Conclusion
  - 6 Reference

# References


- Aitken, C. G., Zadora, G., & Lucy, D. (2007). A two-level model for evidence evaluation. *Journal of forensic sciences*, 52(2), 412–419.
- Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4) (No. 4). Springer.
- Celeux, G., Frühwirth-Schnatter, S., & Robert, C. P. (2019). Model selection for mixture models—perspectives and strategies. In *Handbook of mixture analysis* (pp. 117–154). Chapman and Hall/CRC.
- Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458), 611–631.
- Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: with applications in r*. Springer.

# References

- Goswami, S., & Wegman, E. J. (2016). Comparison of different classification methods on glass identification for forensic research. *J. Stat. Sci. App*, 4, 65–84.
- Hidot, S., & Saint-Jean, C. (2010). An expectation–maximization algorithm for the wishart mixture model: Application to movement clustering. *Pattern Recognition Letters*, 31(14), 2318–2324.
- Horn, R. A., & Johnson, C. R. (2012). *Matrix analysis*. Cambridge university press.
- Hu, Z. (2015). *Initializing the em algorithm for data clustering and sub-population detection*. The Ohio State University.
- McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, 6, 355–378.
- Melnykov, V., & Maitra, R. (2010). Finite mixture models and model-based clustering.


# References

- Michael, S., & Melnykov, V. (2016). An effective strategy for initializing the em algorithm in finite mixture models. *Advances in Data Analysis and Classification*, 10, 563–583.
- Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.
- Wilson, A. G., & Ghahramani, Z. (2010). Generalised wishart processes. *arXiv preprint arXiv:1101.0240*.
- Yu, Y.-K., & Zhang, Y.-C. (2002). On the anti-wishart distribution. *Physica A: Statistical Mechanics and its Applications*, 312(1-2), 1–22.



Questions???





Thank you!!!