Covariance-Based Clustering for Classification

Theophilus Anim Bediako Andrew Simpson Semhar Michael, Ph.D.

Department of Mathematics and Statistics, South Dakota State University

South Dakota State University

Introduction

- Classification involves assigning a given observation to known classes
- This project focuses on problems with a high number of classes and features with few numbers of observations within each class e.g. Forensic source identification problems
- Linear discriminant analysis (LDA) and Quadratic discriminant analysis (QDA) are common classification methods [1]
- Both assume a multivariate normal distribution for each class [3]
- LDA assumption of equal class covariance matrices is simplistic, leading to less flexibility and high bias [3]
- In QDA, estimating class-specific covariance matrices becomes a problem when there is a large number of classes

Objectives

This work proposes a covariance-based clustering to build a classifier

- Use a finite mixture of Wishart to cluster cross-product matrices
- Obtain parameter estimates of the mixtures using EM-algorithm
- Identify observations with similar covariances and build a cluster-based LDA

Wishart Distribution

Let $\mathbf{X_j} = [\mathbf{x_1}, \cdots, \mathbf{x_{n_j}}]$, $\mathbf{x_i} \sim N_p(0, \Sigma_j)$. Then $\mathcal{S}_j = \sum_{i=1}^{n_j} \mathbf{x_i}\mathbf{x_i}^T = \mathbf{X_j}\mathbf{X_j}^T \in \Re^{p \times p}$ follows a Wishart distribution with n_j degrees of freedom and a scale matrix, Σ_j

The pdf of the Wishart is given by:

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

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

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

Γ is the ordinary gamma function

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

S is a positive definite scale matrix; $n > p$; E$(S) = n\Sigma$

Wishart Mixture Model [2]

Given $\mathfrak{S} = \{\mathcal{S}_1, \mathcal{S}_2, \cdots, \mathcal{S}_N\}$, WMM is expressed as:

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

where $\pi_k \in (0,1), \quad \sum_{k=1}^{K} \pi_k = 1$

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

$$\ell(\Theta) = \sum_{i=1}^{N} log\left[\sum_{k=1}^{K} \pi_k f(\mathcal{S}_i; \Sigma_k, n_k)\right]$$

For parameter estimation using the Expectation Maximization (EM) algorithm, argument $\mathcal{S}_i$ with a latent variable z_i to get complete data log-likelihood as

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

$$z_{ik} = \begin{cases} 1 & \text{if } \mathcal{S}_i \in \text{component } k \\ 0 & \text{otherwise} \end{cases}$$

EM Algorithm

An iterative two-step process for fitting mixture models

1. E-step: compute the conditional expectation of complete-data log-likelihood function given data and parameter estimates, $\Theta^{(t)}$ [2].

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

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

2. M-step: maximize the conditional expectation of (2) w.r.t π and Θ_k given the data to find new estimates for $\Theta^{(t+1)}$ [2]

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

We solve for n_k through numerical optimization when n_k is unknown [2]

Repeat steps 1 & 2 until some pre-specified convergence criterion is met

Here, we assume n_k is known and equal for all classes

Cluster-Based LDA Algorithm

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	$\mathbf{y_{ji}}$, $\mathbf{j = 1, \cdots, G}$ & $i = 1, \cdots, n_j$
Predictors	$\mathbf{x_{ji}}$
Cluster labels	$\{\hat{Z}_1, \cdots, \hat{Z}_N\}$
Covariances	$\{\hat{\Sigma}_1, \cdots, \hat{\Sigma}_K\}$

Output: $j = argmax\left(P(g = j|\mathbf{x}^*, \delta_\mathbf{j})\right)$

For $k = 1, \cdots, K$

(i) Use $\hat{\Sigma}_k$ and compute discriminant 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 $\tau_j = P(g = j|X^*, \delta_j)$

(iii) Return $j = argmax\left(P(g = j|X^*, \delta_j)\right)$

Simulation Study

Here, we simulate 100 two-dimensional Gaussian random vectors uniquely centered such that the entries of each class mean, μ_{2x1} come from $\mathbf{runif}(0, 50)$. The mixing proportions and the scale covariance matrices are

$$\pi_1 = \mathbf{0.45} \qquad \pi_2 = \mathbf{0.55}$$

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

The generated 100-class two-dimensional data with 10 observations per class, where there are two underlying covariance matrices. The 95% confidence ellipsoids for 100 classes are shown in Figure 1.

Covariance based clustering

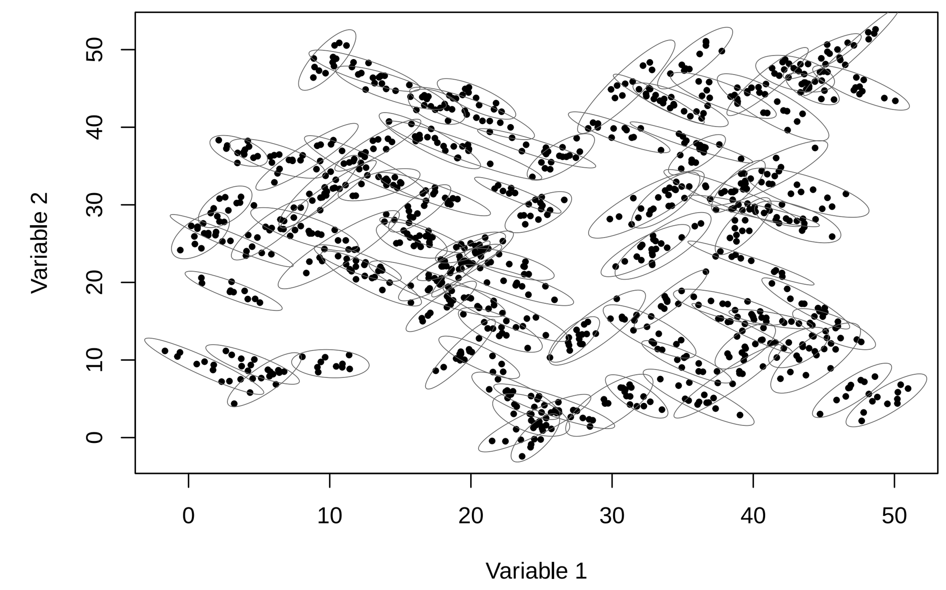

Figure 1. Scatterplot of points with 95% confidence ellipsoid for each class before mean centering

After centering the observations within each class, we obtain the following cross-product matrices via $\mathcal{S}_j = X_jX_j^\top$ to get $\mathfrak{S} = \{\mathcal{S}_1, \cdots, \mathcal{S}_{100}\}$ seen in Figure 2

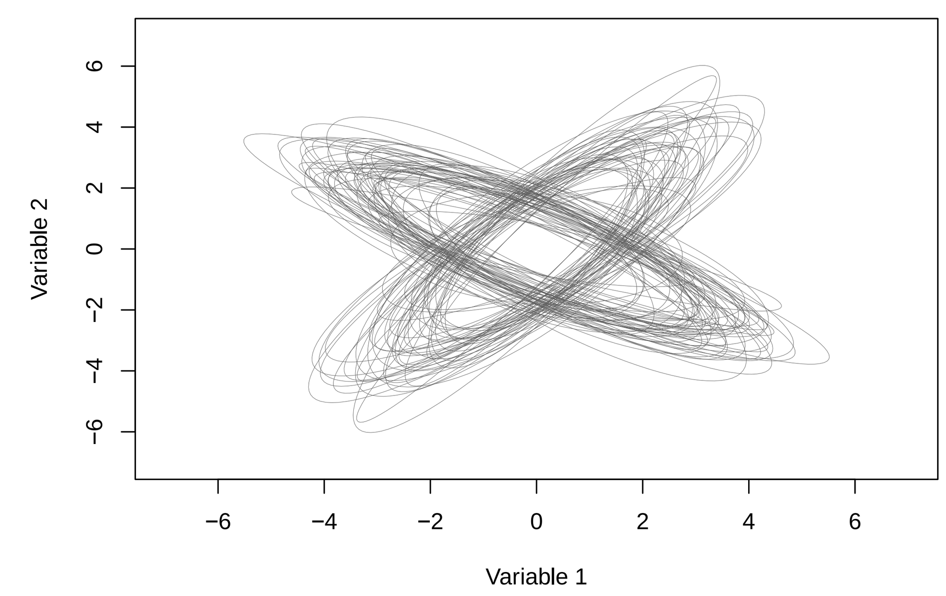

Figure 2. 95% confidence ellipses based on the sample covariance matrix of each class centered at zero

EM algorithm on $\mathfrak{S}$ initialized with the true parameters gives the following parameter estimates

$$\hat{\mathbf{\Sigma}}_\mathbf{1} = \begin{bmatrix} 1.42 & 1.63 \\ 1.63 & 2.69 \end{bmatrix}, \quad \hat{\mathbf{\Sigma}}_\mathbf{2} = \begin{bmatrix} 2.30 & -1.50 \\ -1.50 & 1.44 \end{bmatrix}$$

$$\hat{\pi}_\mathbf{1} = \mathbf{0.45} \qquad \hat{\pi}_\mathbf{2} = \mathbf{0.55}$$

- 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 $\mathfrak{S}$ was misclassified.

Classification

- With 70% train data and 30% test data, we compare the prediction accuracy of LDA, QDA, and Cluster-based LDA
- Cluster-based LDA has 85.0% prediction accuracy compared to 64.0% for LDA and 68.8% for QDA

Current Work

The current work is focused on applying the proposed cluster-based LDA to a glass identification problem. We are also working on developing an initialization algorithm for initializing the EM algorithm when dealing with mixtures of Wishart.

References

[1] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

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

[3] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[4] Volodymyr Melnykov and Ranjan Maitra. Finite mixture models and model-based clustering. 2010.

[5] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

theophilus.animbediako@jacks.sdstate.edu