Outlier Detection in Gaussian Mixture Models

OUTLIER DETECTION IN GAUSSIAN MIXTURE MODELS

BY

KATHARINE M. CLARK, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2020)	McMaster University
(Mathematics & Statistics)	Hamilton, Ontario, Canada

TITLE:	Outlier Detection in Gaussian Mixture Models
AUTHOR:	Katharine M. Clark B.Sc., (Mathematics and Statistics) McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Paul D. McNicholas
NUMBER OF PAGES:	ix, 43

To my parents, friends, and family

Abstract

Unsupervised classification is a problem often plagued by outliers, yet there is a paucity of work on handling outliers in unsupervised classification. Mixtures of Gaussian distributions are a popular choice in model-based clustering. A single outlier can affect parameters estimation and, as such, must be accounted for. This issue is further complicated by the presence of multiple outliers. Predicting the proportion of outliers correctly is paramount as it minimizes misclassification error. It is proved that, for a finite Gaussian mixture model, the log-likelihoods of the subset models are distributed according to a mixture of beta-type distributions. This relationship is leveraged in two ways. First, an algorithm is proposed that predicts the proportion of outliers by measuring the adherence of a set of subset log-likelihoods to a beta-type mixture reference distribution. This algorithm removes the least likely points, which are deemed outliers, until model assumptions are met. Second, a hypothesis test is developed, which, at a chosen significance level, can test whether a dataset contains a single outlier.

Acknowledgements

I would like to thank my supervisor, Dr. Paul McNicholas for his academic, professional, and personal support, guidance, and encouragement throughout my undergraduate and master's degrees.

I would like to express my gratitude to Dr. Roman Viveros, Dr. Sharon McNicholas, and Dr. Paul McNicholas for their roles on my examination committee.

I would like to thank the Natural Sciences and Engineering Research Council (NSERC), the Milos Novotny Fellowship, the Department of Mathematics and Statistics, and the MacDATA Institute for their financial support.

Finally, I would like to thank my friends and family for their love and support. I would not be here without them.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

An outlier is an observation "that appears to deviate markedly from other members of the sample in which it occurs" (Grubbs, 1969). Outliers may occur due to unlikely random chance, or they may arise due to experimental, measurement, or recording error. The treatment of outliers is a long-studied topic in the field of applied statistics.

In the mixture model-based clustering framework, outlying observations can have detrimental effects on classification accuracy. If an outlier is far enough away from the other data points, it may be classified into its own group. This may misrepresent the number of groups present or force the model to merge unrelated groups. At the other extreme, an outlier can affect parameter estimates by moving the mean or inflating the variance, ultimately leading to unsatisfactory clustering results. Thus, outlier detection is an important task when classifying data.

Outliers are often treated in one of three ways. We can treat them as regular data points and keep them in the model, we can down-weight their effects on the model, or we can remove them entirely. This thesis will focus on the third approach—trimming outliers. Two methods presented will identify outliers and decide when to remove them from the dataset.

Chapter 2

Background

2.1 Mixture Model-Based Clustering

Classification aims to partition data into a set number of groups, whereby observations in the same group are in some sense similar to one another. Clustering is unsupervised classification, in that none of the group memberships are known *a priori*. Most clustering algorithms originate from one of three major methods: hierarchical clustering, k-means clustering, and mixture model-based clustering. Although hierarchical and k-means clustering are still used, the mixture modelling approach has become increasingly popular due to its robustness and mathematical interpretability. In the mixture modelling framework for clustering, each component is usually taken to be a cluster. Although the model can employ almost any component distribution, Gaussian components remain popular due to the distribution's versatility and ubiquity. Most mixture model-based clustering methods assume, either explicitly or implicitly, that the data are free of outliers.

Mixture model-based clustering involves maximizing the likelihood of the mixture model. The density of a Gaussian mixture model is a convex linear combination of each component density, and is given by

$$f(\mathbf{x} \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{2.1}$$

where

$$\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}_g|}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x} - \boldsymbol{\mu}_g)\right\}$$

is the density of a p-dimensional random variable $\mathbf{X}$ from a Gaussian distribution with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$, $\pi_g > 0$ is the mixing proportion such that $\sum_{g=1}^{G} \pi_g = 1$, and $\boldsymbol{\vartheta} = \{\pi_1, \ldots, \pi_G, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_G, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_G\}$.

2.2 Outlier Detection Methods

Outliers, particularly those with high leverage, can significantly affect the parameter estimates. It is thus beneficial to remove, or reduce, the effect of outliers by accounting for them in the model. In model-based clustering, we can incorporate outliers in several ways. The first method, proposed by Banfield and Raftery (1993), includes outliers in an additional uniform component over the convex hull. If outliers are cluster-specific, we can incorporate them into the tails if we cluster using mixtures of t-distributions (Peel and McLachlan, 2000). Punzo and McNicholas (2016) introduce mixtures of contaminated Gaussian distributions, where each cluster has a proportion $\alpha_g \in (0, 1)$ of ‘good’ points with density $\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, and a proportion $1 - \alpha_g$ of ‘bad’ points, with density $\phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \eta_g \boldsymbol{\Sigma}_g)$. Each distribution has the same centre, but the ‘bad’ points have an inflated variance, where $\eta_g > 1$.

Instead of fitting outliers in the model, it may be of interest to trim them from the dataset. Cuesta-Albertos *et al.* (1997) developed an impartial trimming approach for k-means clustering; however, this method maintains the drawback of k-means clustering, where the clusters are spherical with equal — or, in practice, similar — radii. García-Escudero *et al.* (2008) improved upon trimmed k-means with the TCLUST algorithm. TCLUST places a restriction on the eigenvalue ratio of the covariance matrix, as well as implementing a weight on the clusters, allowing for clusters of various elliptical shapes and sizes. An obvious challenge with these methods is that the eigenvalue ratio must also be known *a priori.* There exists an estimation scheme for the proportion of outliers, denoted α, but it is heavily influenced by the choices for number of clusters and eigenvalue ratio. It is of great interest to bring trimming into the model-based clustering domain, especially when α is unknown, as is the case for most real datasets — in fact, for all but very low dimensional data.

2.3 Outlier Hypothesis Tests

It may be useful to use a hypothesis test to reject suspected outliers at a particular significance level. In the univariate case, Grubbs's test (Grubbs *et al.*, 1950) tests whether the maximum (or minimum) value is an outlier based on its distance from the sample mean relative to the sample's standard deviation. The test uses the statistic

$$T_n = \frac{x_{(n)} - \bar{x}}{s}$$

in the maximum case, and we conclude that $x_{(n)}$ is an outlier when T_n is larger than some reference value, given in Grubbs *et al.* (1950). This method requires the data to be normally distributed.

The quartile method is a non-parametric outlier rejection method where we find the lower (Q1) and upper (Q3) quartiles, the values under which 25% and 75% of the sample data reside, respectively. A point is considered an outlier if $x < Q1 - 1.5(Q3 - Q1)$ or $x > Q3 + 1.5(Q3 - Q1)$ (Tukey, 1977). This method requires the underlying distribution to be symmetric (Hubert and Vandervieren, 2008).

In the multivariate Gaussian case, one can leverage the fact that the Mahalanobis distance (MD) is chi-squared distributed (Mardia *et al.*, 1979). Points which are far from the mean relative to the standard deviation will have large MDs. We can reject, at a specific significance level, points where the MD is too large.

There is a dearth of hypothesis tests for outliers in the unsupervised Gaussian mixture-model framework.

Chapter 3

Methodology

3.1 Distribution of Log-Likelihoods

In this section, the distribution of subset log-likelihoods is derived. Note that, in Section 3.1.1 we use population parameters whereas, in Section 3.1.2, we use parameter estimates. The ideas developed in this chapter build on those introduced in the author's undergraduate thesis.

3.1.1 Distribution of Subset Log-Likelihoods using Population Parameters

Consider a dataset $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in p-dimensional Euclidian space $\mathbb{R}^p$. Define the jth subset as $\mathcal{X} \setminus \mathbf{x}_j = \{\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_n\}$. Suppose each $\mathbf{x}_i \in \mathcal{X}$ has Gaussian mixture model density $f(\mathbf{x}_i \mid \boldsymbol{\vartheta})$ as in (2.1). The log-likelihood of dataset

$\mathcal{X}$ under the Gaussian mixture model is

$$\ell_{\mathcal{X}} = \sum_{i=1}^{n} \log \left[\sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right]. \tag{3.1}$$

Assumption 1. *The clusters are non-overlapping and well separated.*

Assumption 1 is required to simplify the model density to the component density, as shown in Lemma 1. In practice, however, these assumptions may be relaxed. For more information on the effect of cluster separation on the density, see Appendix A.

Write $\mathbf{x}_i \in \mathcal{C}_g$ to indicate that $\mathbf{x}_i$ belongs to the gth cluster. Let $\mathbf{z}_i = (z_{i1}, \dots, z_{iG})'$, where $z_{ig} = 1$ if $\mathbf{x}_i \in \mathcal{C}_g$ and $z_{ig} = 0$ if $\mathbf{x}_i \notin \mathcal{C}_g$.

Lemma 1. *As the separation between the clusters increases, $\ell_{\mathcal{X}} \simeq Q_{\mathcal{X}}$. In other words, the log-likelihood in* (3.1) *converges asymptotically to $Q_{\mathcal{X}}$, where*

$$Q_{\mathcal{X}} = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \log \left[\pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right] = \sum_{\mathbf{x}_i \in \mathcal{C}_g} \log \left[\pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right].$$

A proof of Lemma 1 may be found in Appendix B. We will maintain Assumption 1 throughout this paper. Using Lemma 1, an approximate log-likelihood for the mixture model is

$$Q_{\mathcal{X}} = \sum_{\mathbf{x}_i \in \mathcal{C}_g} \left[\log \pi_g + \log \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right],$$

which can be regarded as the approximate log-likelihood for the entire dataset $\mathcal{X}$. Define $Q_{\mathcal{X} \setminus \mathbf{x}_j}$ as the approximate log-likelihood for the jth subset $\mathcal{X} \setminus \mathbf{x}_j$.

Proposition 1. *If $Y_j = Q_{\mathcal{X}\setminus\mathbf{x}_j} - Q_{\mathcal{X}}$ and $\mathbf{x}_j \in \mathcal{C}_h$, then $Y_j \sim f_{\text{gamma}}\left(y_j - c \mid p/2, 1\right)$, where $c = -\log \pi_h + \frac{p}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{\Sigma}_h|$, and*

$$f_{\text{gamma}}\left(w \mid k, \theta\right) = \frac{1}{\Gamma(k)\theta^k} w^{k-1} \exp\{-w/\theta\},$$

for $w > 0, k > 0$, and $\theta > 0$.

The requisite mathematical results are given in the following lemmata.

Lemma 2. *For $\mathbf{x}_j \in \mathcal{C}_h$,*

$$Q_{\mathcal{X}\setminus\mathbf{x}_j} - Q_{\mathcal{X}} = -\log \pi_h + \frac{p}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{\Sigma}_h| + \frac{1}{2}\tau_j,$$

where

$$\tau_j = (\mathbf{x}_j - \boldsymbol{\mu}_h)'\mathbf{\Sigma}_h^{-1}(\mathbf{x}_j - \boldsymbol{\mu}_h).$$

Proof. Population parameters $\boldsymbol{\mu}_g$ and $\mathbf{\Sigma}_g$, $g \in [1, G]$, are impervious to the sample drawn from the dataset and remain constant for each subset $\mathcal{X} \setminus \mathbf{x}_j$, $j \in [1, n]$. Thus, the approximate log-likelihood for the jth subset, $\mathcal{X} \setminus \mathbf{x}_j$, when $\mathbf{x}_j \in \mathcal{C}_h$ is

$$Q_{\mathcal{X}\setminus\mathbf{x}_j} = Q_{\mathcal{X}} - \log \pi_h - \log \phi(\mathbf{x}_j \mid \boldsymbol{\mu}_h, \mathbf{\Sigma}_h). \tag{3.2}$$

Rearranging (3.2) yields

$$Q_{\mathcal{X}\setminus\mathbf{x}_j} - Q_{\mathcal{X}} = -\log \pi_h + \frac{p}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{\Sigma}_h| + \frac{1}{2}\tau_j. \tag{3.3}$$

□

Lemma 3. $\tau_j \sim f_{\text{chi-squared}}(p)$

This result is stated as Corollary 3.2.1.1 in Mardia *et al.* (1979).

Lemma 4. $\frac{1}{2}\tau_j \sim f_{\text{gamma}}(p/2, 1)$.

Proof. If $\tau_j \sim f_{\text{chi-squared}}(p)$, then $\tau_j \sim f_{\text{gamma}}(p/2, 2)$. Thus, $\frac{1}{2}\tau_j \sim f_{\text{gamma}}(p/2, 1)$ by the scaling property of the gamma distribution. □

Let $Y_j = Q_{\mathcal{X}\setminus\mathbf{x}_j} - Q_{\mathcal{X}}$, $\mathbf{x}_j \in \mathcal{C}_h$, and $c = -\log \pi_h + \frac{p}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{\Sigma}_h|$. Then,

$$Y_j \sim f_{\text{gamma}}\left(y_j - c \mid p/2, 1\right),$$

for $y_j - c > 0$.

3.1.2 Distribution of Subset Log-Likelihoods using Sample Parameter Estimates

Generally, population parameters $\boldsymbol{\mu}_g$ and $\mathbf{\Sigma}_g$ are unknown *a priori.* We can replace the population parameters with parameter estimates

$$\begin{aligned}\hat{\boldsymbol{\mu}}_g &= \bar{\mathbf{x}}_g = \frac{1}{n_g}\sum_{\mathbf{x}_i \in \mathcal{C}_g} \mathbf{x}_i,\\ \hat{\mathbf{\Sigma}}_g &= \frac{1}{n_g - 1}\sum_{\mathbf{x}_i \in \mathcal{C}_g} (\mathbf{x}_i - \bar{\mathbf{x}}_g)(\mathbf{x}_i - \bar{\mathbf{x}}_g)' =: \mathbf{S}_g,\end{aligned}$$

where $n_g = \sum_{i=1}^{n} z_{ig}$ is the number of observations in $\mathcal{C}_g$.

Assumption 2. *The number of observations in each cluster, n_g, is large.*

This is assumption required for the following lemmata.

Lemma 5. *Sample parameter estimates are asymptotically equal for all subsets:*

$$\bar{\mathbf{x}}_{g\setminus j} \simeq \bar{\mathbf{x}}_g,$$

$$\mathbf{S}_{g\setminus j} \simeq \mathbf{S}_g,$$

where $\bar{\mathbf{x}}_g$ *and* $\mathbf{S}_g$ *are the sample mean and sample covariance, respectively, for the* g*th cluster considering all observations in the entire dataset* $\mathcal{X}$*, and* $\bar{\mathbf{x}}_{g\setminus j}$ *and* $\mathbf{S}_{g\setminus j}$ *are the sample mean and sample covariance, respectively, for the* g*th cluster considering only observations in the* j*th subset* $\mathcal{X} \setminus \mathbf{x}_j$.

Proof. If $\mathbf{x}_j \in \mathcal{C}_h$, then the equality trivially holds for all $g \neq h$. For $g = h$,

$$\bar{\mathbf{x}}_{h\setminus j} = \frac{n_h \bar{\mathbf{x}}_h - \mathbf{x}_j}{n_h - 1}.$$

Thus $\bar{\mathbf{x}}_{h\setminus j} \to \bar{\mathbf{x}}_h$ as $n_h \to \infty$. Therefore, $\bar{\mathbf{x}}_{h\setminus j} \simeq \bar{\mathbf{x}}_h$ and so

$$\mathbf{S}_{h\setminus j} \simeq \frac{(n_h - 1)\mathbf{S}_k - (\mathbf{x}_j - \bar{\mathbf{x}}_h)(\mathbf{x}_j - \bar{\mathbf{x}}_h)'}{n_h - 2}.$$

Thus $\mathbf{S}_{h\setminus j} \to \mathbf{S}_h$ as $n_h \to \infty$, so $\mathbf{S}_{h\setminus j} \simeq \mathbf{S}_h$. □

Using the sample parameter estimates, (3.3) becomes

$$Q_{\mathcal{X}\setminus \mathbf{x}_j} - Q_{\mathcal{X}} = -\log \pi_h + \frac{p}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{S}_h| + \frac{1}{2}t_j,$$

where $t_j = (\mathbf{x}_j - \bar{\mathbf{x}}_h)'\mathbf{S}_h^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}_h)$.

Lemma 6. *(From Gnanadesikan and Kettenring, 1972) When* $\mathbf{X} \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*,*

$$\frac{n}{(n-1)^2}T_j \sim f_{\text{beta}}\left(\frac{n}{(n-1)^2}t_j \;\middle|\; \frac{p}{2}, \frac{n-p-1}{2}\right),$$

for $t_j \geq 0, \alpha > 0, \beta > 0.$

Ververidis and Kotropoulos Ververidis and Kotropoulos (2008) prove this result for all n, p satisfying $p < n < \infty$.

Proposition 2. *For* $\mathbf{x}_j \in \mathcal{C}_h$*, with* $Y_j = Q_{\mathcal{X}\setminus\mathbf{x}_j} - Q_{\mathcal{X}}$ *and* $c = -\log\pi_h + \frac{p}{2}\log(2\pi) + \frac{1}{2}\log|\mathbf{S}_h|$*,*

$$Y_j \sim f_{\text{beta}}\left(\frac{2n_h}{(n_h-1)^2}(y_j - c) \;\middle|\; \frac{p}{2}, \frac{n_h-p-1}{2}\right),$$

for $y_j - c \geq 0, \alpha > 0, \beta > 0.$

Proof. We will perform a change of variables. Let $X_j = \frac{n_h}{(n_h-1)^2}T_j$ and $Y_j = \frac{1}{2}T_j + c$. Then

$$Y_j = \frac{(n_h-1)^2}{2n_h}X_j + c.$$

The inverse function is

$$x_j = v(y_j) = \frac{2n_h}{(n_h-1)^2}(y_j - c).$$

The absolute value of the derivative of x_j with respect to y_j is

$$\left|\frac{dx_j}{dy_j}\right| = \left|\frac{2n_h}{(n_h-1)^2}\right| = \frac{2n_h}{(n_h-1)^2}.$$

Because X_j is beta-distributed, its density is

$$f_{\text{beta}}(x_j \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_j^{\alpha-1}(1-x_j)^{\beta-1},$$

for $x_j \geq 0$, $\alpha > 0$, and $\beta > 0$. The transformation of variables allows the density of Y_j to be written

$$f_Y(y_j) = f_X\left(v(y_j)\right)\left|\frac{dx_j}{dy_j}\right|.$$

The density of Y_j becomes

$$f_Y(y_j) = \frac{2n_h}{(n_h-1)^2}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\left[\frac{2n_h}{(n_h-1)^2}(y_j-c)\right]^{\alpha-1}\left[1-\frac{2n_h}{(n_h-1)^2}(y_j-c)\right]^{\beta-1}, \tag{3.4}$$

for $y_j - c \geq 0$, $\alpha > 0$, and $\beta > 0$. Thus, Y_j has a beta-type density with

$$Y_j \sim f_{\text{beta}}\left(\frac{2n_h}{(n_h-1)^2}(y_j-c) \;\middle|\; \frac{p}{2}, \frac{n_h-p-1}{2}\right).$$

□

Because $f(y_j)$ applies to any $Y_j = Q_{\mathcal{X}\backslash\mathbf{x}_j} - Q_{\mathcal{X}}$, with $\mathbf{x}_j \in \mathcal{C}_h$, let $f(y_j) = f_h(y)$. Proposition 2 can be applied to generate the density of the mixture model with variable $Y = Q_{\mathcal{X}\backslash\mathbf{x}_j} - Q_{\mathcal{X}}$ for any $\mathbf{x}_j \in \mathcal{X}$. The density is given by

$$f(y \mid \boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(y \mid \boldsymbol{\theta}_g), \tag{3.5}$$

where $f_g(y \mid \boldsymbol{\theta}_g)$ is the beta-type density given in (3.4), and $\boldsymbol{\theta}_g = \{n_g, p, \pi_g, \mathbf{S}_g\}$.

Remark 1. *Y has density $f(y \mid \boldsymbol{\vartheta})$ from* (3.5) *when typical model assumptions hold.*

If the density in (3.5) *does not describe the distribution of subset log-likelihoods, that is, they are not distributed according to a mixture of beta-type densities, then we can conclude that at least one model assumption fails. In this case, we will assume that only the outlier assumption has been violated and that there are, in fact, outliers in the model.*

3.2 OCLUST Algorithm

Let $\mathcal{Y}$ be the set of subset log-likelihoods generated from the data. Thus, $\mathcal{Y}$ is the realization of random variable Y. We propose testing the adherence of $\mathcal{Y}$ to the reference distribution in (3.5) as a way to test for the presence of outliers. In other words, if $\mathcal{Y}$ does not have a beta-type mixture distribution, then outliers are present in the model. Because $Q_{\mathcal{X}}$ is asymptotically equal to $\ell_{\mathcal{X}}$, we will use $\ell_{\mathcal{X}}$. This is important because we will need $\ell_{\mathcal{X}\setminus\mathbf{x}_j}$ for outlier identification and, additionally, it is outputted by many existing clustering algorithms. The algorithm described below uses the log-likelihood and parameter estimates calculated using the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977) for Gaussian model-based clustering; however, other methods may be used to estimate parameters and the overall log-likelihood.

OCLUST both identifies likely outliers and determines the proportion of outliers within the dataset. The OCLUST algorithm assumes all model assumptions hold, except that outliers are present. The algorithm involves removing points one-by-one until the density in (3.5) describes the distribution of $\mathcal{Y}$, which is determined using Kullback-Leibler (KL) divergence, estimated via relative frequencies. Notably, KL divergence generally decreases as outliers are removed and the model improves. Once all outliers are removed, KL divergence increases again as points are removed from the tails. We select the number of outliers as the location of the global minimum. With each iteration, we remove the most likely outlier.

Definition 1 (Most likely outlier)**.** *With each iteration, we define the most likely outlier as* $t = \mathbf{x}_k$*, where*

$$k = \arg\max_{j\in[1,n]} \ell_{\mathcal{X}\setminus\mathbf{x}_j}.$$

In other words, we assign the kth point as outlying if the log-likelihood is greatest when point $t = \mathbf{x}_k$ is removed. The OCLUST algorithm is outlined in Algorithm 1.

Algorithm 1 OCLUST algorithm.

```
Initialize parameters:
1: Cluster the data into $G$ clusters using the EM algorithm, and calculate the log-likelihood of the clustering solution, $\ell_{\mathcal{X}}$.
2: Calculate the sample covariance $\mathbf{S}_g$, the number of points $n_g$, and the proportion of points $\pi_g = n_g/n$ for each cluster.
Calculate KL divergence:
3: Create $n$ new datasets $\mathcal{X} \setminus \mathbf{x}_j$, each with one $\mathbf{x}_j$ removed.
4: Cluster each of the $n$ datasets into $G$ clusters, calculating the log-likelihood $\ell_{\mathcal{X}\setminus\mathbf{x}_j}$ for each solution.
5: Create a new set $\mathcal{Y} = \{\ell_{\mathcal{X}\setminus\mathbf{x}_j} - \ell_{\mathcal{X}}\}_{j=1:n}$ of realized values for variable $Y$.
6: Generate the density of $Y$ using (3.5) and the parameters from Step 1.
7: Calculate the approximate KL divergence of $\mathcal{Y}$ to the generated density, using relative frequencies.
Determine the most likely outlier $t$ as per Definition 1.
Update:
8: $n \hookleftarrow n - 1$.
9: $\mathcal{X} \hookleftarrow \mathcal{X} \setminus t$.
Perform: $(F+1)$ iterations of Steps 1–9 until an upper bound, $F$, of desired outliers is obtained and the resulting KL divergence is calculated.
Choose: the number of outliers as the value for which the KL divergence is minimized.
```

3.3 Hypothesis Test

Consider a dataset with a single suspected outlier, $\mathbf{x}_k$. Let $\mathbf{x}_k$ be the point corresponding to the maximum subset log-likelihood. We may wish to determine at a specified significance level if $\mathbf{x}_k$ is an outlier. We generate a test of hypothesis, where

$$H_0 : \text{no outliers are present}$$
$$H_A : \mathbf{x}_k \text{ is an outlying point.}$$

Under the null hypothesis, $Y = Q_{\mathcal{X}\setminus\mathbf{x}_j} - Q_{\mathcal{X}}$ is distributed according to a mixture of beta-type densities, with the probability density function given in (3.5). The cumulative distribution function is given by

$$\begin{aligned} F(y \mid \boldsymbol{\vartheta}) &= \int_{-\infty}^{y} \sum_{g=1}^{G} \pi_g f_g(z \mid \boldsymbol{\theta}_g) dz && (3.6) \\ &= \sum_{g=1}^{G} \pi_g \int_{-\infty}^{y} f_g(z \mid \boldsymbol{\theta}_g) dz && (3.7) \\ &= \sum_{g=1}^{G} \pi_g F_g(y \mid \boldsymbol{\theta}_g), && (3.8) \end{aligned}$$

where $F_g(y \mid \boldsymbol{\theta}_g)$ is the cumulative distribution function of the scaled and shifted beta function given in (3.4). Thus, by a property of the maximum, the cumulative density for the maximum subset log-likelihood is given by

$$\begin{aligned} F(y_{(n)} \mid \boldsymbol{\vartheta}) &= [F(y \mid \boldsymbol{\vartheta})]^n && (3.9) \\ &= \left(\sum_{g=1}^{G} \pi_g F_g(y \mid \boldsymbol{\theta}_g) \right)^n . && (3.10) \end{aligned}$$

Finally, we can calculate the p-value of our test,

$$\text{P-value} = P(Y_k > y) = 1 - P(Y_k \leq y) = 1 - \left(\sum_{g=1}^{G} \pi_g F_g(y \mid \boldsymbol{\theta}_g) \right)^n .$$

We reject H_0 when the p-value is less than our specified significance level α.

Chapter 4

Analyses

4.1 OCLUST

4.1.1 Simulation Study

The first simulation study tests the performance of OCLUST against the following three popular outlier detection algorithms:

a. TCLUST (García-Escudero *et al.*, 2008);

b. Contaminated normal mixtures (CNMix; Punzo and McNicholas, 2016); and

c. Noise component mixtures (NCM), mixtures of Gaussian clusters and a uniform component (Banfield and Raftery, 1993).

The datasets were generated to closely mimic those used by García-Escudero *et al.* (2008) and, as such, the simulation scheme and notation used here are borrowed therefrom. Datasets containing three clusters with means $\boldsymbol{\mu}_1 = (0, 8, 0, \ldots, 0)'$, $\boldsymbol{\mu}_2 = (8, 0, 0, \ldots, 0)'$, and $\boldsymbol{\mu}_3 = (-8, -8, 0, \ldots, 0)'$, respectively, were generated with $n =$

1000, and $p = 2$ or $p = 6$. Covariance matrices were generated of the forms:

$$\mathbf{\Sigma}_1 = \text{diag}(1, a, 1, \ldots, 1), \quad \mathbf{\Sigma}_2 = \text{diag}(b, c, 1, \ldots, 1), \quad \mathbf{\Sigma}_3 = \left(\begin{array}{cc|c} d & e & \multirow{2}{*}{\mathbf{0}} \\ e & f & \\ \hline \multicolumn{2}{c|}{\mathbf{0}} & \mathbf{I} \end{array}\right).$$

With different combinations for (a, b, c, d, e, f), we generate five different models:

I. $(a, b, c, d, e, f) = (1, 1, 1, 1, 0, 1)$, spherical clusters with equal volumes;

II. $(a, b, c, d, e, f) = (5, 1, 5, 1, 0, 5)$, diagonal clusters with equal covariance matrices;

III. $(a, b, c, d, e, f) = (5, 5, 1, 3, -2, 3)$, clusters with equal volumes, but varying shapes and orientations;

IV. $(a, b, c, d, e, f) = (1, 20, 5, 15, -10, 15)$, clusters with varying volumes, shapes, and orientations; and

V. $(a, b, c, d, e, f) = (1, 45, 30, 15, -10, 15)$, clusters with varying volumes, shapes, and orientations but two with severe overlap.

To fix the proportion of outliers to $\alpha = 0.1$, each dataset had 900 'regular' observations and 100 outliers. Outliers were generated uniformly in the p-parallelotope defined by the coordinate-wise maxima and minima of the 'regular' observations, accepting only those points with Mahalanobis squared distances greater than $\chi^2_{p,0.995}$. Datasets either had equal cluster proportions ($\pi_1 = \pi_2 = \pi_3 = 1/3$) or unequal proportions ($\pi_1 = 1/5, \pi_2 = \pi_3 = 2/5$). Ten datasets were generated with each combination of parameters (dimension, cluster proportions, model).

We would like to test each method for classification accuracy and accuracy in predicting α. Each method was run with $G = 3$. OCLUST was run using the `oclust` (Clark and McNicholas, 2019) package for `R` (R Core Team, 2018), with an upper bound $F = 125$ ($\alpha = 0.125$). TCLUST was run using the `tclust` (Fritz *et al.*, 2012) package, with eigenvalue restriction $c = 50$. The proportion of outliers for each dataset was estimated as the location of the 'elbow' of the plot generated by the `ctlcurves` function. CNMix was run using the `CNmixt` function (Punzo *et al.*, 2018) with default initialization. NCM was run using the `Mclust` function (Scrucca *et al.*, 2016), initializing the noise component as a random sample of points with probability 1/4. Both CNMix and NCM inherently estimate the proportion of outliers.

Table 4.1 shows the average estimated proportion of outliers predicted by each method, over the ten datasets. It is paramount that we correctly predict the proportion of outliers, lest we introduce errors in outlier detection. CNMix and NCM generally over-specify α, while TCLUST generally under-specifies α. As a result, the former methods tend to have larger errors in labelling 'regular' points as outliers, and the latter tends to have larger error in labelling outliers as 'regular'. Crucially, OCLUST predicts α very well overall, with the predicted value for α always falling within one standard deviation of the mean. On average, OCLUST predicts closest to the true value of α, and as such any point mislabelled as an outlier usually has a corresponding point mislabelled as 'regular'. A breakdown of each type of error is available in Table C.1 in Appendix C.

We evaluate each method using outlier misclassification error. Table 4.2 lists the outlier misclassification errors for each method, and Figure C.1 in Appendix C displays the results graphically. All four methods perform with similar misclassification

Table 4.1: Means and standard deviations for the proportion of outliers predicted by each method for the simulated datasets, where "E" denotes equal and "U" denotes unequal mixing proportions π_g.

			OCLUST		TCLUST		CNMix		NCM	
π_g	p	Mod.	Mean	SD	Mean	SD	Mean	SD	Mean	SD
E	2	I	.1008	.0055	.0868	.0027	.1066	.0045	.1066	.0024
		II	.0987	.0035	.0876	.0027	.1143	.0083	.1132	.0035
		III	.0999	.005	.0901	.0046	.1068	.0082	.1111	.0035
		IV	.1001	.0038	.1029	.0033	.2083	.0719	.1442	.0096
		V	.1022	.0078	.1089	.0041	.147	.197	.1949	.0266
	6	I	.1018	.0051	.087	.0047	.1022	.0033	.1073	.002
		II	.1016	.0025	.0857	.0033	.1082	.0061	.1092	.0028
		III	.0997	.0027	.0907	.0021	.1068	.0051	.1112	.0049
		IV	.1036	.0062	.0928	.0029	.1488	.0351	.124	.0063
		V	.101	.0052	.0988	.0034	.0859	.0844	.1248	.0077
U	2	I	.1	.0045	.0954	.0024	.1072	.0081	.108	.0032
		II	.0991	.003	.0947	.0039	.1203	.0097	.1135	.0035
		III	.101	.0045	.0986	.0027	.1076	.0064	.1123	.0044
		IV	.1028	.0051	.0962	.0038	.1489	.0343	.1417	.0059
		V	.102	.0051	.1107	.0034	.2629	.1659	.1814	.0236
	6	I	.1003	.0037	.0956	.0025	.1036	.0049	.1067	.0027
		II	.1012	.0058	.093	.0017	.1079	.0028	.1119	.004
		III	.1023	.0037	.0937	.0052	.1109	.0225	.1105	.0028
		IV	.0985	.005	.0906	.0034	.1132	.0364	.119	.0044
		V	.0968	.0061	.0973	.0047	.1383	.0808	.1253	.0064

rates in Models I–III, but OCLUST and TCLUST significantly outperform CNMix and NCM in Models IV and V. This may be due to the fact that clusters one and two in Models IV and V are close together or overlapping. In this case, the contamination for each cluster is non-symmetrical, so CNMix classifies the outliers into one cluster with large contamination parameter. NCM consistently underestimates the variance of each cluster, which over-specifies α and results in outlier miclassification error. OCLUST and TCLUST perform consistently with similar misclassification rates, and

OCLUST has the lowest misclassification error in 13 of the 20 models.

Table 4.2: Outlier misclassification rate from running each method on the simulated datasets. Classifications for OCLUST and TCLUST were taken to be those produced when α was estimated, the average of which is detailed in Table 4.1, where "E" denotes equal and "U" denotes unequal mixing proportions π_g.

			OCLUST		TCLUST		CNMix		NCM	
π_g	p	Model	Mean	SD	Mean	SD	Mean	SD	Mean	SD
E	2	I	.0074	.0021	.0134	.0028	.0072	.0038	.0066	.0024
		II	.0071	.0018	.0142	.0031	.0143	.0083	.0132	.0035
		III	.0065	.0031	.0123	.0037	.0102	.0044	.0111	.0035
		IV	.0079	.0022	.0075	.0018	.1085	.0717	.0442	.0096
		V	.0198	.0075	.0117	.0026	.167	.1014	.0949	.0266
	6	I	.0074	.0022	.013	.0047	.0066	.0018	.0075	.0022
		II	.0076	.0022	.0157	.0031	.0084	.006	.0092	.0028
		III	.0053	.0028	.0103	.0019	.0108	.0036	.0122	.0045
		IV	.0088	.0036	.0116	.0032	.056	.0262	.0264	.0055
		V	.0074	.0044	.0074	.0037	.0691	.0465	.0256	.0068
U	2	I	.007	.0027	.0068	.0028	.01	.0047	.008	.0032
		II	.0071	.0025	.0093	.0038	.0205	.0095	.0135	.0035
		III	.0068	.0034	.0066	.0022	.01	.0034	.0123	.0044
		IV	.0098	.004	.0106	.0031	.0535	.0258	.0419	.0061
		V	.014	.0046	.0143	.0029	.2029	.1061	.0814	.0236
	6	I	.0067	.0031	.0078	.003	.0082	.0032	.0069	.0025
		II	.0086	.0042	.0106	.0027	.0091	.0014	.0119	.004
		III	.0075	.002	.0103	.004	.0239	.0129	.0123	.0026
		IV	.0113	.0035	.0132	.0039	.0354	.0154	.0222	.004
		V	.0108	.0027	.0101	.0021	.0689	.0546	.0259	.0061

4.1.2 Crabs Study

Next we compare the performance of OCLUST on a real dataset. For this, we use the crabs dataset from Campbell and Mahon (1974). This study closely mimics the study done by Peel and McLachlan (2000) and again by Punzo and McNicholas (2016).

The dataset contains observations for 100 blue crabs, 50 of which are male, and 50 of which are female. The aim for each classification is to recover the sex of the crab. For this study, we will focus on measurements of rear width (RW) and carapace length (CL). We substitute the CL value of 25th point to one of eight values in $[-15, 20]$. The leftmost plot in Figure 4.1 plots the crabs dataset by sex, with the permuted value in blue taking value CL $= -5$. OCLUST, as well as the three other comparative methods in Section 4.1.1 were run for each dataset. With the exception of TCLUST, each method was run, restricting the model to one where the clusters had equal shapes and volumes, but varying orientations. Solutions for OCLUST, CNMix, and NCM for the dataset with CL $= -5$ are also plotted in Figure 4.1. Table 4.3 summarizes the results for each method, listing the number of misclassifications (M), the predicted number of outliers (n_O), and whether the model identified the permuted point as an outlier (bad).

Table 4.3: Results for running each method on the crabs dataset, where "M" and "n_O" designate the number of misclassified points and number of predicted outliers, respectively, and 'bad' indicates whether the substituted point was labelled as an outlier.

	OCLUST			TCLUST			CNMix			NCM		
CL	M	n_O	bad	M	n_O	bad	M	n_O	bad	M	n_O	bad
−15	10	4	✓	20	1	✓	13	1	✓	13	1	✓
−10	10	4	✓	20	1	✓	13	1	✓	13	1	✓
−5	10	4	✓	20	1	✓	13	1	✓	13	1	✓
0	10	4	✓	20	1	✓	13	1	✓	13	1	✓
5	10	4	✓	20	1	✓	13	1	✓	13	2	✓
10	10	4	✓	20	1	✓	13	1	✓	11	3	✓
15	10	4	✓	20	1	✓	13	1	✓	10	4	✓
20	10	4	✓	20	1	✓	13	1	✓	9	5	✓

Every method identifies the permuted value correctly as an outlier. OCLUST, TCLUST, and CNMix are robust as they retain the same classifications for each

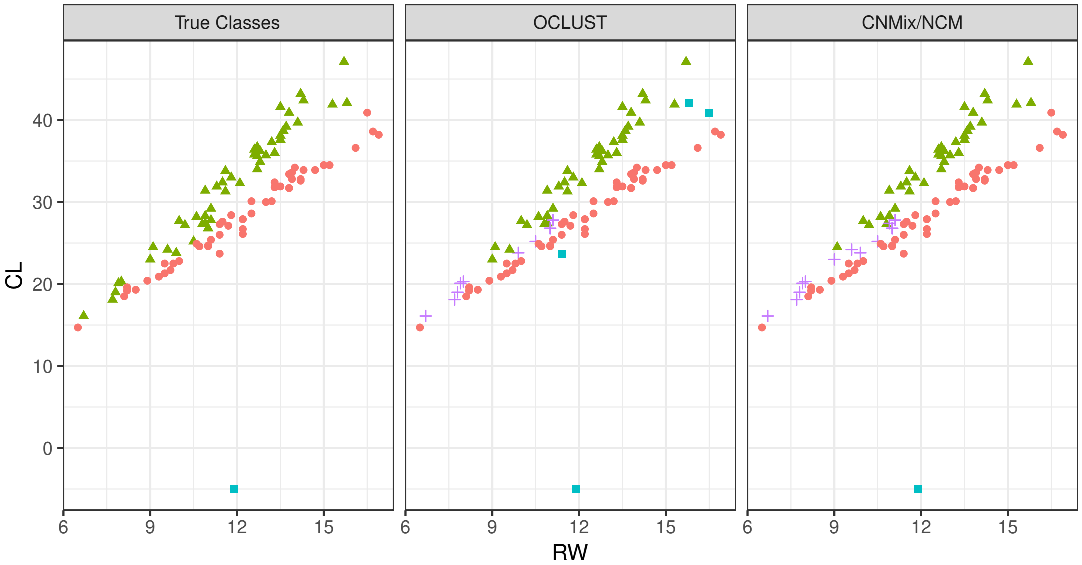

Figure 4.1: Predicted classifications for OCLUST, and the shared solution for CNMix and NCM, when CL $= -5$. Green triangles, red circles, purple crosses, and blue squares indicate male, female, misclassified, and outlying points, respectively.

dataset, regardless of the value for CL. NCM classifies more points as outliers as the permuted CL value becomes less extreme. TCLUST has the highest misclassification rate, which is expected as it employs k-means clustering, which tends to fail when the clusters are elliptical. NCM outputs the same classifications as CNMix when CL $\in [-15, 0]$, but differs when CL > 0, at which point NCM begins to classify points like OCLUST (see Figure 4.2 for comparison of OCLUST, CNMix, and NCM on the dataset with permuted point having CL $= 20$). Although the sum of 'n_O' and 'M' are always the same for OCLUST and CNMix, it is important to note that the points misclassified by CNMix are not the points labelled as outliers by OCLUST . Instead, as seen in Figure 4.2, OCLUST identifies two points between the clusters as technical outliers. This removes the points with high leverage, allowing the clusters to rotate and improve the classification among low values of RW.

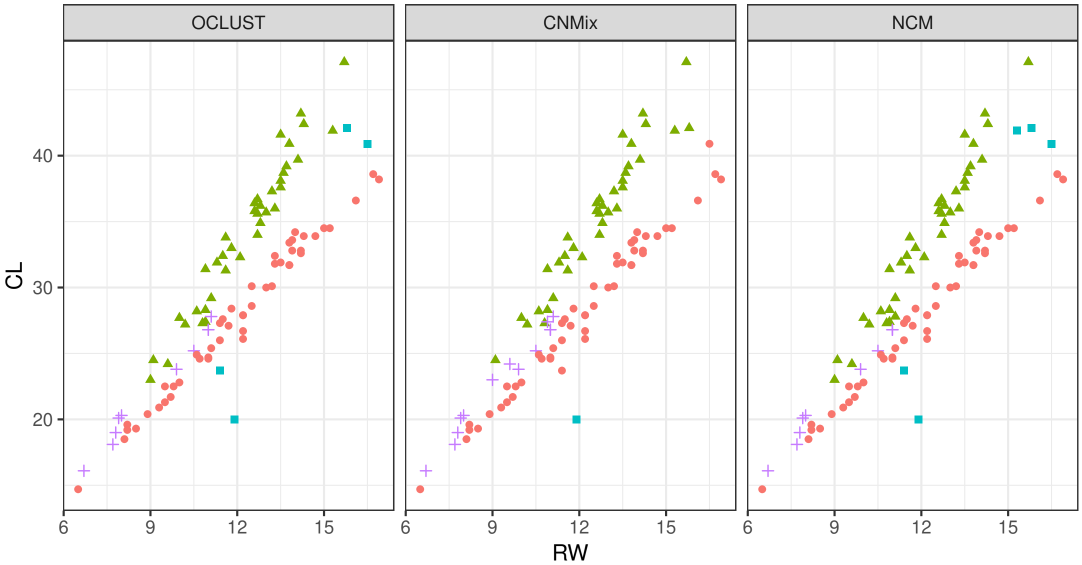

Figure 4.2: Predicted classifications for OCLUST, CNMix, and NCM, when CL = 20. Green triangles, red circles, purple crosses, and blue squares indicate male, female, misclassified, and outlying points, respectively.

4.2 Hypothesis Test

4.2.1 Simulation Study

The second simulation study tests the effectiveness of the hypothesis test for different values of n. Data are generated in two and six dimensions using Model IV in Section 4.1.1 with unequal proportions ($\pi_1 = 1/5, \pi_2 = \pi_3 = 2/5$). Different values of n are tested with $n \in \{50, 100, 200, 400, 800\}$. For each n, 100 datasets are generated without outliers, and 100 are created by adding a single outlier to the 100 aforementioned datasets. The outlier is generated randomly between two and seven units away from the global maximum or minimum value in each dimension.

In each case, the first 100 datasets evaluate the hypothesis test's specificity, while the latter 100 evaluate the test's sensitivity. Each outlier is tested at the $\alpha = 0.05$

Table 4.4: Error rate for each combination of p and n.

p	outlier?	n	# rejected	error rate
2	yes	50	100	0%
		100	100	0%
		200	100	0%
		400	100	0%
		800	100	0%
	no	50	42	42%
		100	15	15%
		200	17	17%
		400	9	9%
		800	3	3%
6	yes	50	100	0%
		100	100	0%
		200	100	0%
		400	97	3%
		800	100	0%
	no	50	97	97%
		100	85	85%
		200	55	55%
		400	25	25%
		800	10	10%

significance level. A table of the results are shown in Table 4.4. The hypothesis test's sensitivity remains nearly constant at 100% for every n, but specificity remains low for $n = 50$ and increases as n increases. Thus, the power of the test increases with n, which is a common feature of most tests. A graphical representation of the error rates is shown in Figure 4.3.

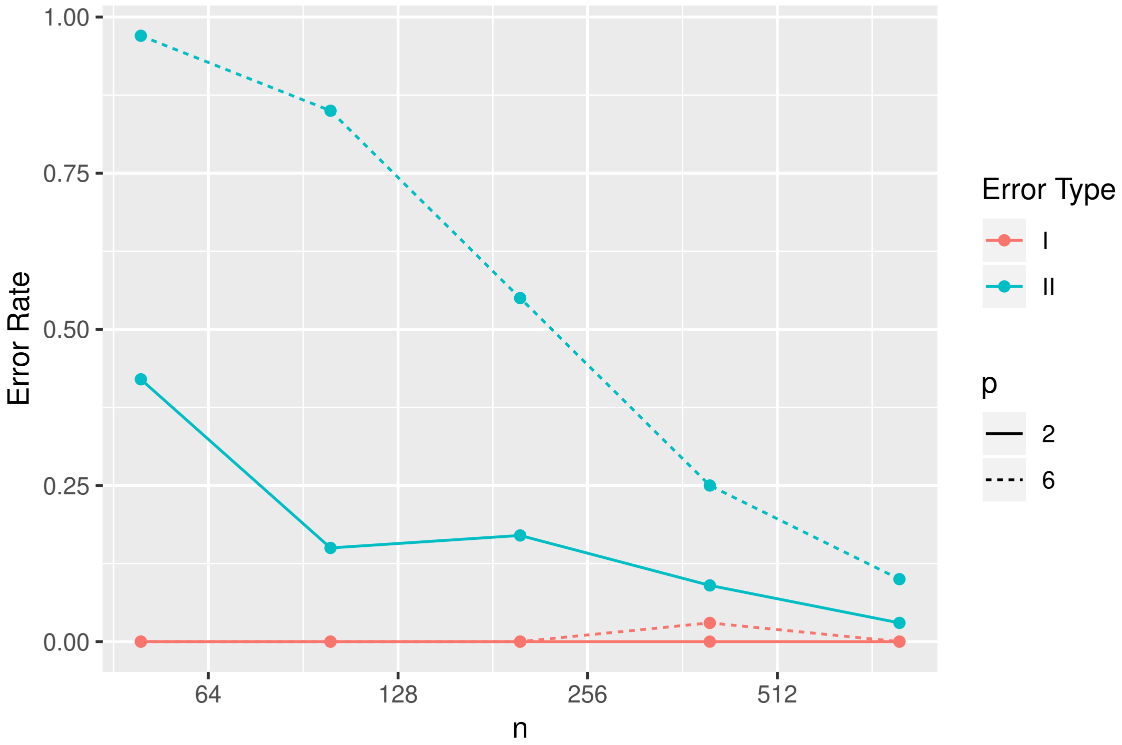

Figure 4.3: Type I and type II error rates for each value of n. Note that n is plotted with a base-2 logarithmic scale.

4.2.2 Real Data

Iris Dataset

The iris dataset originates from the `datasets` package. It contains measurements of sepal length, sepal width, petal length, and petal width for 50 flowers each of three species of iris. A pairs plot of the data is shown in Figure 4.4. By visual inspection, there do not appear to be outliers.

The hypothesis test was applied to this dataset, which resulted in a p-value of 0.0753. We do not reject H_0 at the $\alpha = 0.05$ significance level and we conclude that the iris dataset does not contain an outlier. This result is expected and reflects the

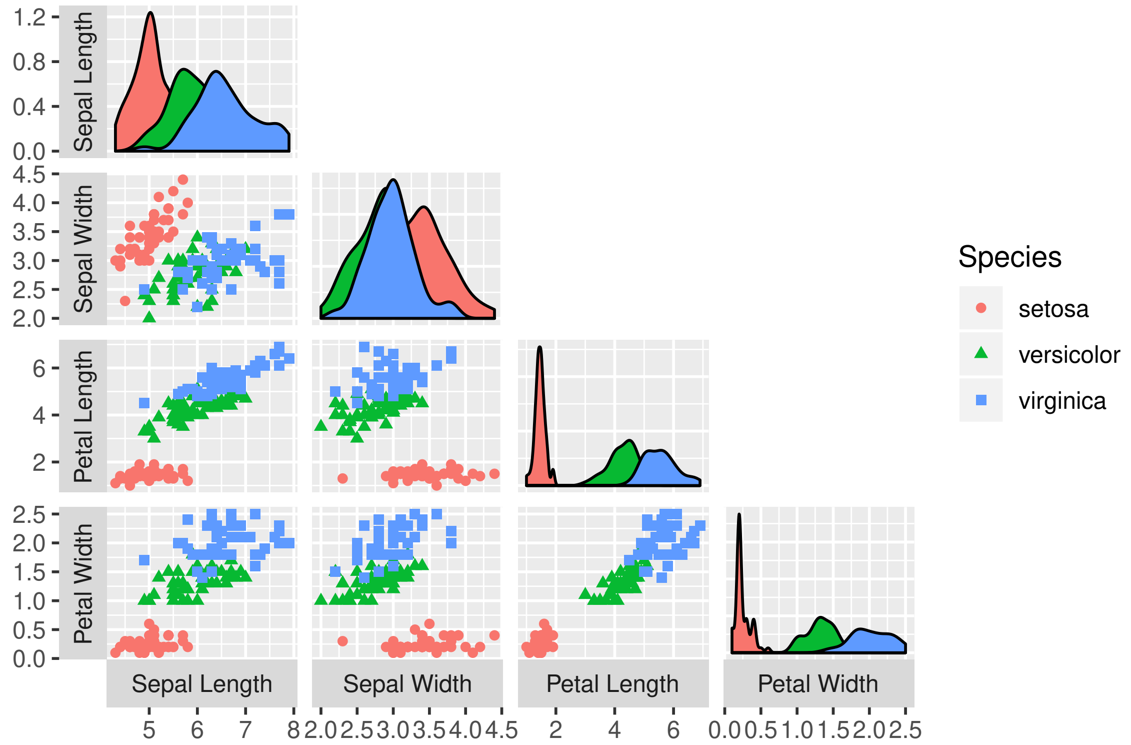

Figure 4.4: A pairs plot of the iris dataset. Points are coloured based on true classes.

information from the pairs plot.

Crabs Dataset

We extend the crabs study performed in Section 4.1.2, but we substitute the CL value of 25th point to one of twenty values in $[-15, 80]$. This serves to evaluate the responsiveness of the test. The results are shown in Table 4.5. We reject the null hypothesis at the $\alpha = 0.05$ significance level and conclude that there is an outlier when CL is in $\{\text{CL} \leq 20\} \cup \{\text{CL} \geq 40\}$.

The three datasets are plotted in Figure 4.5 when the hypothesis is not rejected (i.e. CL $\in \{25, 30, 35\}$). By visual inspection, these points are not outliers as they seem to be reasonable for the dataset. The hypothesis test performs as expected.

Table 4.5: P-values for the crabs dataset with substituted CL values.

CL	p-value	CL	p-value
-15	0	35	0.0940
-10	0	40	0.0004
-5	0	45	0
0	0	50	0
5	0	55	0
10	0	60	0
15	0	65	0
20	0	70	0
25	0.1417	75	0
30	0.1187	80	0

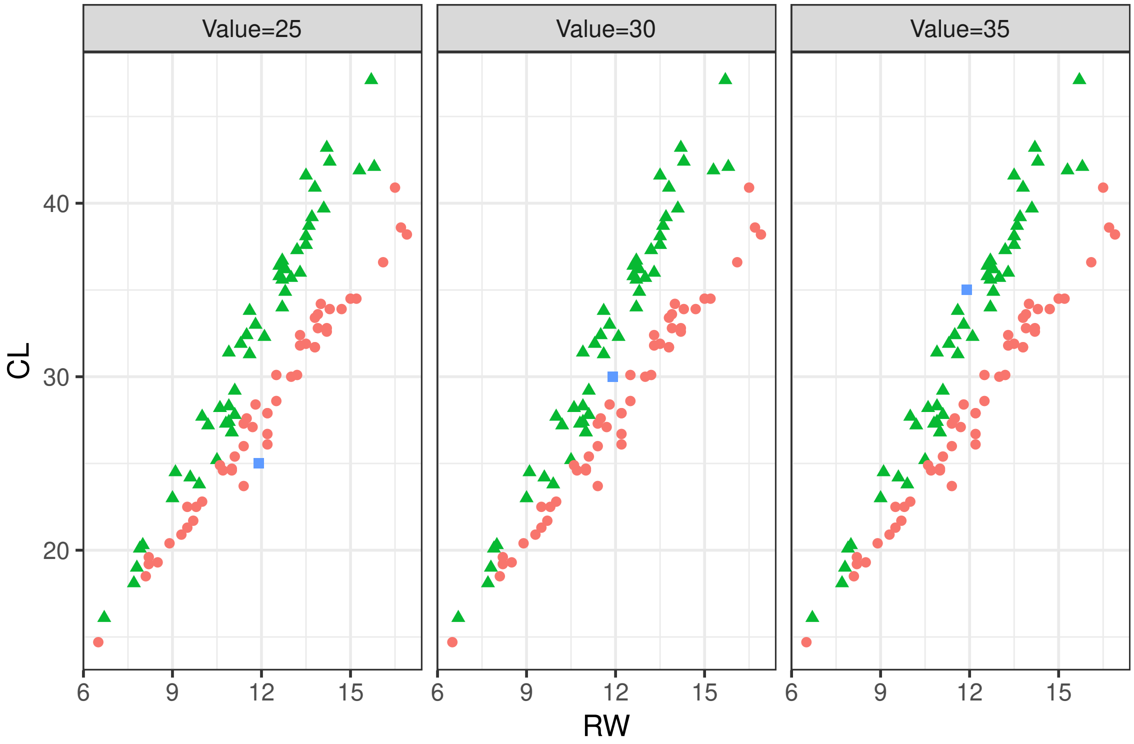

Figure 4.5: The three datasets when we fail to reject the null hypothesis. The colours correspond to the true classes, with blue squares indicating the outliers.

Chapter 5

Conclusions and Future Work

It was proved that, for data from a Gaussian mixture, the log-likelihoods of the subset models are distributed according to a mixture of beta-type distributions. This result was used in two ways, first to determine the number of outliers by removing outlying points until the subset log-likelihoods followed this derived distribution. The result is the OCLUST algorithm, which trims outliers from a dataset and predicts the proportion of outliers. Second, the distribution was used to develop a hypothesis test to determine if a single outlier is present.

In simulations, the trimming methods OCLUST and TCLUST outperform the additional-outlier-component methods CNMix and NCM, and OCLUST outperfoms all methods 65% of the time. Crucially, however, OCLUST produces the best estimation for the proportion of outliers, and as such does not consistently misclassify outliers as 'regular', as is the case with TCLUST, or consistently misclassify 'regular' points as outliers, as is the case with CNMix and NCM. In the crabs study, OCLUST trims technical outliers with high leverage, which improves the classification among small values of carapace length.

The hypothesis test performs with nearly 100% sensitivity and increasing specificity as n increases. It predicts that the iris dataset is free of outliers and it demonstrates excellent responsiveness on the crabs dataset when carapace length is gradually changed.

Although this work used the distribution of the log-likelihoods of the subset models to test for the presence of outliers, the derived distribution may be used to verify other underlying model assumptions, such as whether the clusters are Gaussian. Note that the OCLUST algorithm could be used with other clustering methods and should be effective so long as is it reasonable to assume that the underlying distribution of clusters is Gaussian. Of course, one could extend this work by deriving the distribution of subset log-likelihoods for mixture models with non-Gaussian components. This would allow direct consideration of asymmetric clusters with outliers. Finally, one could extend this approach to high-dimensional data by using an analogue of the mixture of factor analyzers model or its extensions (see Ghahramani and Hinton (1997), McNicholas and Murphy (2008), McNicholas and Murphy (2010)). Based on the comparisons conducted herein, one might expect the resulting method to perform favourably, or at least comparably, when compared to the approaches used by Wei and Yang (2012) and Punzo *et al.* (2020).

Appendix A

Relaxing Assumptions

Lemma 1 assumes that the clusters are well separated and non-overlapping to simplify the model density to the component density. This section, however, serves to show that this assumption may be relaxed in practice. Following Qiu and Joe (2006), we can quantify the separation between clusters using the separation index J^*; in the univariate case,

$$J^* = \frac{L_2(\alpha/2) - U_1(\alpha/2)}{U_2(\alpha/2) - L_1(\alpha/2)},$$

where $L_i(\alpha/2)$ is the sample lower $\alpha/2$ quantile and $U_i(\alpha/2)$ is the sample upper $\alpha/2$ quantile of cluster i, and cluster 1 has lower mean than cluster 2. In the multivariate case, the separation index is calculated along the projected direction of maximum separation. Clusters with $J^* > 0$ are separated, clusters with $J^* < 0$ overlap, and clusters with $J^* = 0$ are touching.

To measure the effect of separation index on the approximate log-likelihood $Q_{\mathcal{X}}$, 100 random datasets with $n = 1800$ for each combination were generated using the `clusterGeneration` (Qiu and Joe, 2015) package in R. Data were created with three clusters with equal cluster proportions with dimensions $p \in \{2, 4, 6\}$ and separation indices in $[-0.9, 0.9]$. Covariance matrices were generated using random eigenvalues $\lambda \in (1, 10)$. The parameters were estimated using the `mclust` package. The log-likelihoods using the full and approximate densities were calculated using the parameter estimates. $[Q_{\mathcal{X}} - \ell_{\mathcal{X}}]/\ell_{\mathcal{X}}$, the average proportional change in log-likelihood over the 100 datasets between the full log-likelihood $\ell_{\mathcal{X}}$ and the approximate log-likelihood $Q_{\mathcal{X}}$, is reported in Table A.1. A graphical representation of the results is shown in Figure A.1. As one would expect, the approximation of $\ell_{\mathcal{X}}$ by $Q_{\mathcal{X}}$ improves

Table A.1: The proportional change in log-likelihood between the log-likelihood $\ell_{\mathcal{X}}$ and the approximate log-likelihood $Q_{\mathcal{X}}$, for different values of p and varying values for separation, where 0^* indicates no computationally-detectable difference.

Separation	Difference in Log-Likelihoods (%)		
Value	$p = 2$	$p = 4$	$p = 6$
−0.9	1.13E+01	5.14E+00	3.03E+00
−0.8	1.03E+01	5.00E+00	2.99E+00
−0.7	9.70E+00	4.82E+00	2.89E+00
−0.6	9.15E+00	4.38E+00	2.67E+00
−0.5	7.74E+00	3.90E+00	2.35E+00
−0.4	6.46E+00	3.21E+00	2.00E+00
−0.3	4.86E+00	2.43E+00	1.44E+00
−0.2	3.25E+00	1.57E+00	9.76E-01
−0.1	1.83E+00	8.86E-01	5.54E-01
0	8.18E-01	4.06E-01	2.59E-01
0.1	2.58E-01	1.32E-01	8.51E-02
0.2	5.05E-02	2.51E-02	1.60E-02
0.3	4.24E-03	2.29E-03	1.24E-03
0.4	1.63E-05	8.48E-06	2.52E-05
0.5	3.94E-10	6.49E-11	2.85E-10
0.6	0*	0*	0*
0.7	0*	0*	0*
0.8	0*	0*	0*
0.9	0*	0*	0*

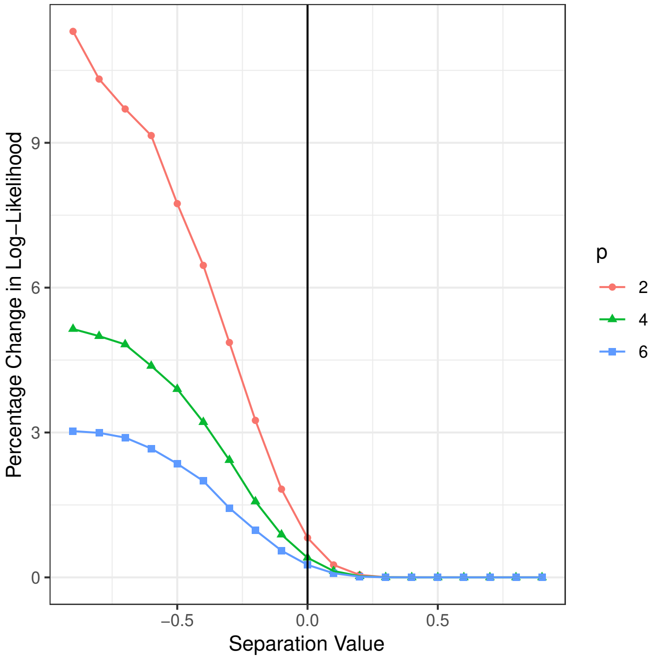

Figure A.1: Graphical representation of the results in Table A.1, showing the effect of cluster separation on the approximate log-likelihood of the model, where the vertical line represents the threshold between separated and overlapping clusters.

as the separation index increases (see Lemma 1). However, the difference is negligible for touching and separated clusters ($J^* \geq 0$). In the simulations in Section 4.1.1, the most overlapping clusters have $J^* = -0.09732371$, which produces an error of less than 2% in two dimensions, and less than 0.6% in six dimensions. In this case, the approximation is appropriate.

Appendix B

Mathematical Results

B.1 Proof of Lemma 1

Proof. Suppose $\mathbf{\Sigma}$ is positive definite. Then, $\mathbf{\Sigma}^{-1}$ is also positive definite and there exists $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$ such that $\mathbf{\Sigma}^{-1} = \mathbf{Q}'\mathbf{\Lambda}\mathbf{Q}$ and $\mathbf{\Lambda}$ is diagonal with $\mathbf{\Lambda}_{ii} = \lambda_i > 0, i \in [1, p]$. Let $\mathbf{x} - \boldsymbol{\mu} = \mathbf{Q}'\mathbf{w}$, where $\mathbf{w} \neq \mathbf{0}$. Now,

$$
\begin{aligned}
(\mathbf{x} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{w}'\mathbf{Q}\mathbf{\Sigma}^{-1}\mathbf{Q}'\mathbf{w} = \mathbf{w}'\mathbf{\Lambda}\mathbf{w} &= \sum_{i=1}^{p} \lambda_i w_i^2 \\
&\geq \inf_i(\lambda_i) \sum_{i=1}^{p} w_i^2 = \inf_i(\lambda_i)\|\mathbf{w}\|^2 = \inf_i(\lambda_i)\|\mathbf{x} - \boldsymbol{\mu}\|^2
\end{aligned}
$$

because $\|\mathbf{x} - \boldsymbol{\mu}\|^2 = \|\mathbf{Q}'\mathbf{w}\|^2 = \mathbf{w}'\mathbf{Q}\mathbf{Q}'\mathbf{w} = \|\mathbf{w}\|^2$. Thus, as $\|\mathbf{x} - \boldsymbol{\mu}\| \to \infty, (\mathbf{x} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) \to \infty$ and

$$
\phi(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p|\mathbf{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\} \to 0.
$$

Suppose $\mathbf{x}_i \in \mathcal{C}_h$. Then, as the clusters separate, $\|\mathbf{x}_i - \boldsymbol{\mu}_g\| \to \infty$ and $\phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \to 0$ for $g \neq h$. Thus, for $\mathbf{x}_i \in \mathcal{C}_h$,

$$\sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) = \sum_{g \neq h} \pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) + \pi_h \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h) \simeq \pi_h \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h).$$

Thus,

$$\ell_{\mathcal{X}} = \sum_{i=1}^{n} \log \left[\sum_{g=1}^{G} \pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right] \simeq \sum_{\mathbf{x}_i \in \mathcal{C}_g} \log \left[\pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \right] = Q_{\mathcal{X}}.$$

□

Remark 2. *Although covariance matrices need only be positive semi-definite, we restrict* $\boldsymbol{\Sigma}$ *to be positive definite so that* $\mathbf{X}$ *is not degenerate.*

Appendix C

Additional Tables and Figures

Table C.1: Average outlier detection error for each method. Classifications for OCLUST and TCLUST used the estimated α parameter, the average of which is detailed in Table 4.1. Numbers without parentheses indicate the proportion of 'good' points classified as outliers, and those in parentheses indicate the proportion of outliers classified as 'good'.

			OCLUST		TCLUST		CNMix		NCM	
π_g	p	Model	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Equal	2	I	.0046(.033)	.0038(.0245)	.0001(.133)	.0004(.0271)	.0077(.003)	.0045(.0067)	.0073(0)	.0027(0)
		II	.0032(.042)	.0017(.0235)	.001(.133)	.0008(.0279)	.0159(0)	.0092(0)	.0147(0)	.0039(0)
		III	.0036(.033)	.0021(.0371)	.0013(.111)	.0009(.0409)	.0094(.017)	.0063(.0343)	.0123(0)	.0039(0)
		IV	.0044(.039)	.002(.0247)	.0058(.023)	.0027(.0116)	.1204(.001)	.0798(.0032)	.0491(0)	.0106(0)
		V	.0122(.088)	.007(.0432)	.0114(.014)	.0028(.0237)	.1189(.6)	.1644(.5164)	.1054(0)	.0296(0)
	6	I	.0051(.028)	.0034(.0253)	0(.13)	0(.0474)	.0049(.022)	.0025(.0148)	.0082(.001)	.0023(.0032)
		II	.0051(.03)	.002(.0149)	.0008(.15)	.0009(.0309)	.0092(.001)	.0067(.0032)	.0102(0)	.0031(0)
		III	.0028(.028)	.0022(.0187)	.0006(.098)	.0008(.0187)	.0098(.02)	.0045(.017)	.013(.005)	.0052(.0071)
		IV	.0069(.026)	.005(.0237)	.0024(.094)	.0016(.0267)	.0582(.036)	.0336(.0695)	.028(.012)	.0064(.014)
		V	.0047(.032)	.0035(.0368)	.0034(.043)	.0026(.0267)	.0306(.416)	.0513(.5014)	.028(.004)	.008(.0097)
Unequal	2	I	.0039(.035)	.0036(.0184)	.0012(.057)	.0013(.0231)	.0096(.014)	.0066(.0284)	.0089(0)	.0036(0)
		II	.0034(.04)	.002(.0211)	.0022(.073)	.0017(.035)	.0227(.001)	.0107(.0032)	.015(0)	.0039(0)
		III	.0043(.029)	.0025(.0325)	.0029(.04)	.0012(.0221)	.0098(.012)	.005(.0244)	.0137(0)	.0049(0)
		IV	.007(.035)	.0046(.019)	.0038(.072)	.0014(.0326)	.0569(.023)	.0327(.0727)	.0464(.001)	.0066(.0032)
		V	.0089(.06)	.0047(.0236)	.0139(.018)	.0031(.0155)	.2032(.2)	.1474(.4216)	.0904(0)	.0263(0)
	6	I	.0039(.032)	.0024(.0262)	.0019(.061)	.0015(.0242)	.0066(.023)	.0039(.0216)	.0076(.001)	.0029(.0032)
		II	.0054(.037)	.0049(.0245)	.002(.088)	.001(.0204)	.0094(.006)	.002(.0126)	.0132(0)	.0044(0)
		III	.0054(.026)	.0027(.0165)	.0022(.083)	.0016(.0445)	.0193(.065)	.0158(.1158)	.0127(.009)	.0029(.0074)
		IV	.0054(.064)	.0032(.032)	.0021(.113)	.0016(.034)	.027(.111)	.0184(.2253)	.0229(.016)	.0044(.0135)
		V	.0042(.07)	.0027(.0408)	.0041(.064)	.0026(.028)	.0596(.153)	.0672(.3313)	.0284(.003)	.0069(.0067)

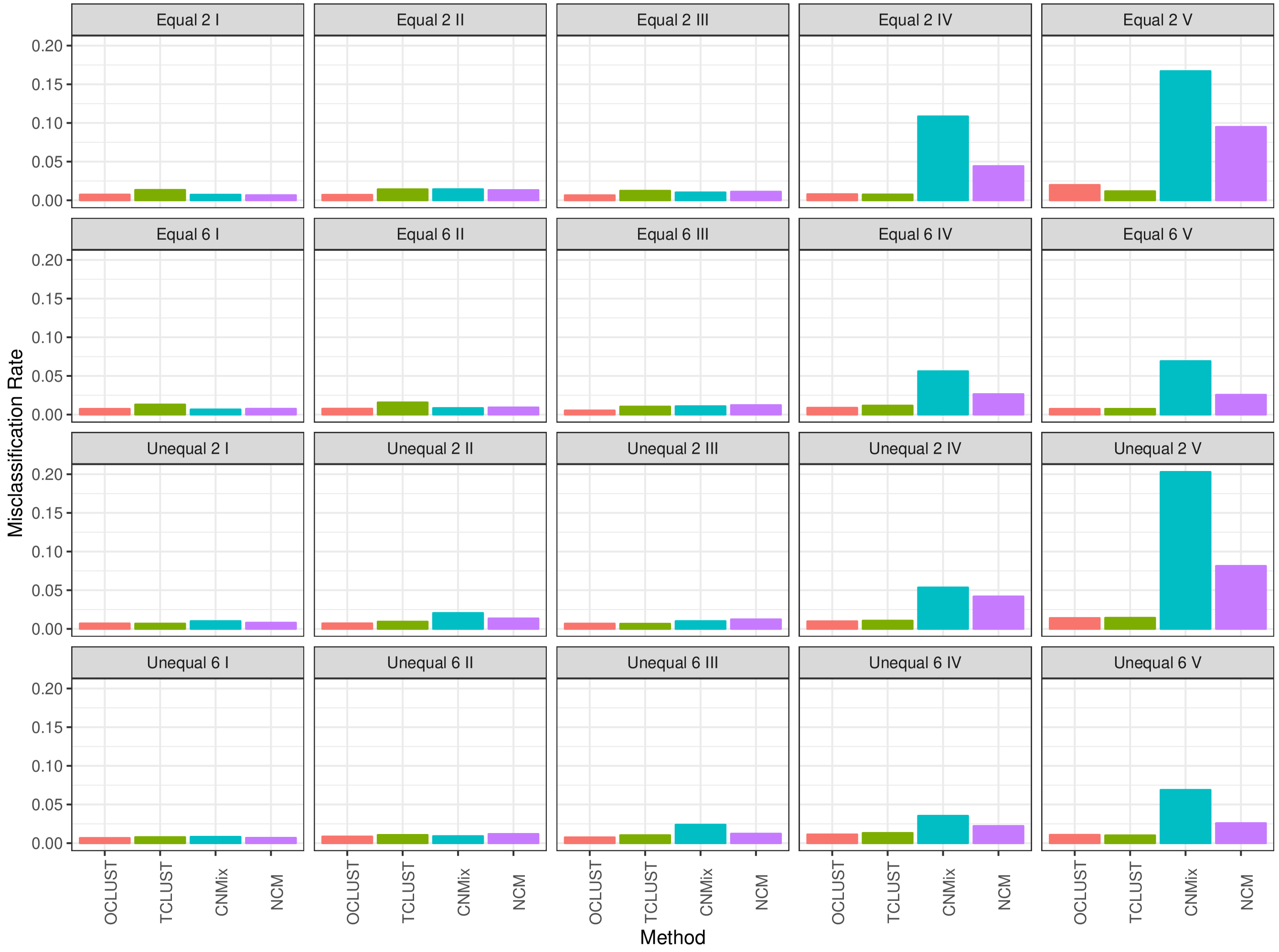

Figure C.1: Graphical representation of Table 4.2, outlier misclassification error by model.

Bibliography

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**(3), 803–821.

Campbell, N. A. and Mahon, R. J. (1974). A multivariate study of variation in two species of rock crab of genus leptograpsus. *Australian Journal of Zoology*, **22**, 417–425.

Clark, K. M. and McNicholas, P. D. (2019). *oclust: Gaussian Model-Based Clustering with Outliers*. R package version 0.1.0.

Cuesta-Albertos, J. A., Gordaliza, A., and Matrán, C. (1997). Trimmed k-means: an attempt to robustify quantizers. *The Annals of Statistics*, **25**(2), 553–576.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, **39**(1), 1–38.

Fritz, H., García-Escudero, L. A., and Mayo-Iscar, A. (2012). tclust: An R package for a trimming approach to cluster analysis. *Journal of Statistical Software*, **47**(12), 1–26.

García-Escudero, L. A., Gordaliza, A., Matrán, C., and Mayo-Iscar, A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, **36**(3), 1324–1345.

Ghahramani, Z. and Hinton, G. E. (1997). The EM algorithm for factor analyzers. Technical Report CRG-TR-96-1, University of Toronto, Toronto, Canada.

Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, **28**(1), 81–124.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, **11**(1), 1–21.

Grubbs, F. E. *et al.* (1950). Sample criteria for testing outlying observations. *The Annals of Mathematical Statistics*, **21**(1), 27–58.

Hubert, M. and Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational statistics & data analysis*, **52**(12), 5186–5201.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press, London.

McNicholas, P. D. and Murphy, T. B. (2008). Parsimonious Gaussian mixture models. *Statistics and Computing*, **18**(3), 285–296.

McNicholas, P. D. and Murphy, T. B. (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics*, **26**(21), 2705–2712.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. *Statistics and Computing*, **10**(4), 339–348.

Punzo, A. and McNicholas, P. D. (2016). Parsimonious mixtures of multivariate contaminated normal distributions. *Biometrical Journal*, **58**(6), 1506–1537.

Punzo, A., Mazza, A., and McNicholas, P. D. (2018). ContaminatedMixt: An R package for fitting parsimonious mixtures of multivariate contaminated normal distributions. *Journal of Statistical Software*, **85**(10), 1–25.

Punzo, A., Blostein, M., and McNicholas, P. D. (2020). High-dimensional unsupervised classification via parsimonious contaminated mixtures. *Pattern Recognition*, **98**, 107031.

Qiu, W. and Joe, H. (2006). Separation index and partial membership for clustering. *Computational Statistics and Data Analysis*, **50**(3), 585–603.

Qiu, W. and Joe, H. (2015). *clusterGeneration: Random Cluster Generation (with Specified Degree of Separation)*. R package version 1.3.4.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, **8**(1), 205–233.

Tukey, J. W. (1977). *Exploratory data analysis*, volume 2. Reading, MA.

Ververidis, D. and Kotropoulos, C. (2008). Gaussian mixture modeling by exploiting the Mahalanobis distance. *IEEE Transactions on Signal Processing*, **56**(7), 2797–2811.

Wei, X. and Yang, Z. (2012). The infinite Student's t-factor mixture analyzer for robust clustering and classification. *Pattern Recognition*, **45**, 4346–4357.