On Fractionally-Supervised Classification: Weight Selection and Extension to the Multivariate t-Distribution

ON FRACTIONALLY-SUPERVISED CLASSIFICATION: WEIGHT SELECTION AND EXTENSION TO THE MULTIVARIATE T-DISTRIBUTION

BY

MICHAEL P.B. GALLAUGHER, 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 (2016) McMaster University

(Mathematics & Statistics) Hamilton, Ontario, Canada

TITLE:	On Fractionally-Supervised Classification: Weight Selection and Extension to the Multivariate t-Distribution
AUTHOR:	Michael P.B. Gallaugher B.Sc., (Mathematics and Statistics) McMaster University, Hamilton, Canada
SUPERVISOR:	Dr. Paul D. McNicholas
NUMBER OF PAGES:	viii, 50

To my parents, Eleanor and Brian

Abstract

Recent work on fractionally-supervised classification (FSC), an approach that allows classification to be carried out with a fractional amount of weight given to the unlabelled points, is extended in two important ways. First, and of fundamental importance, the question over how to choose the amount of weight given to the unlabelled points is addressed. Then, the FSC approach is extended to mixtures of multivariate t-distributions. The first extension is essential because it makes FSC more readily applicable to real problems. The second, although less fundamental, demonstrates the efficacy of FSC beyond Gaussian mixture models.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Paul McNicholas. His continual guidance, support, encouragement, and passion for research was crucial in the completion of this thesis.

Secondly, I would like to acknowledge the funds provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Milos Novotny Fellowship from McMaster University, and the Department of Mathematics and Statistics.

I would also like to show my appreciation to Dr. Roman Viveros, and Dr. Petar Jevtic who, along with Dr. McNicholas, were on my examination committee. I would like to thank all three of them for making the thesis defence an enjoyable experience.

I would like to thank my friends, both old and new, who I had the pleasure of working with throughout my time as a Masters student.

Finally, I would like to thank my parents, my grandparents, and my sister for all of their love, support and reminding me, that although hard work is important, there is much more to life than work.

Contents

List of Figures

Chapter 1

Introduction

In a typical classification application, some of the observations are unlabelled and the objective is to predict the labels of the unlabelled points, for details see McNicholas (2016a). In such situations, classification is generally semi-supervised or supervised (also called discriminant analysis). These two species of classification differ in whether any weight is given to the unlabelled points in the prediction of their labels. In semi-supervised classification, the labelled and unlabelled points are given equal weight; however, in supervised classification, the unlabelled points are given zero weight. Vrbik and McNicholas (2015) introduce a general approach, called fractionally-supervised classification (FSC), where classification can be carried out with a fractional amount of weight — anything between none and all — being given to the unlabelled points.

The approach of Vrbik and McNicholas (2015), which is rooted in the mixture model-based paradigm but can be applied more generally, is extended in two important ways herein. First, the question of how to choose the fraction, i.e., the amount of weight to give the unlabelled points, is addressed. Second, the FSC approach is

extended to mixtures of multivariate t-distributions.

Chapter 2

Background

2.1 Finite Mixture Models and Model-Based Clustering

The finite mixture model was first used for model-based clustering in Wolfe (1965) and has become one of the most common methods for model based clustering. A finite mixture model assumes that an observation $\mathbf{x}$ comes from a population with G subgroups. The density function of $\mathbf{x}$ is given by

$$f(\mathbf{x}|\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \pi_g f_g(\mathbf{x}|\boldsymbol{\theta}_g) \tag{2.1}$$

where $\pi_g > 0$, $\sum_{g=1}^{G} \pi_g = 1$, are called the mixing proportions, $f_g(\cdot)$ are the component densities, and $\boldsymbol{\vartheta} = (\pi_1, \pi_2, \ldots, \pi_G, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_G)$.

Because of its mathematical tractability, the Gaussian mixture model has been looked at extensively in the literature. In addition to Wolfe (1965), other examples of

earlier work in the area of model-based clustering using Gaussian mixtures include, Baum *et al.* (1970), Scott and Symons (1971) and Orchard and Woodbury (1972). For more details on the history of model based clustering, see McNicholas (2016b). More recently, there has also been a fair amount of work using non Gaussian mixtures such as the t-distribution (Peel and McLachlan, 2000) and skewed distributions (Vrbik and McNicholas, 2012, 2014; Franczak *et al.*, 2014; Dang *et al.*, 2015).

2.2 Three Species of Classification

Let the $N \times D$ matrix, $\mathbb{X} = (\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_N')'$, be our data matrix, where the $\mathbf{x}_i$ are D-dimensional vectors and N is the number of data points. We can then split $\mathbb{X}$ into two sub-matrices $\mathbb{X}_1$ and $\mathbb{X}_2$, where $\mathbb{X}_1 = (\mathbf{x}_{11}', \mathbf{x}_{12}', \ldots, \mathbf{x}_{1n_1}')'$ are data points with known labels, and $\mathbb{X}_2 = (\mathbf{x}_{21}', \mathbf{x}_{22}', \ldots, \mathbf{x}_{2n_2}')'$ are observations with unknown labels. We can then write $\mathbb{X} = (\mathbb{X}_1, \mathbb{X}_2)'$.

We can also define $\mathbb{Z} = (\mathbb{Z}_1, \mathbb{Z}_2)'$, to be a matrix of indicator vectors. Specifically, we define $\mathbb{Z}_1 = (\mathbf{z}_1^{(1)'}, \mathbf{z}_2^{(1)'}, \ldots, \mathbf{z}_{n_1}^{(1)'})'$, were $\mathbf{z}_i^{(1)}$ are G dimensional vectors with elements 0 or 1. For convenience, we will denote element g of $\mathbf{z}_j^{(1)}$ by $z_{jg}^{(1)}$ where

$$z_{jg}^{(1)} = \begin{cases} 1 & \text{if } \mathbf{x}_{1j} \text{ is in group } g \\ 0 & \text{otherwise} \end{cases}$$

We can likewise define $\mathbb{Z}_2$ in the same manner. Furthermore, $z_{jg}^{(2)}$ for $j = 1, 2, \ldots, n_2$ are analogous to $z_{jg}^{(1)}$ for the unlabelled observations.

We can then define $D_{\text{O}} = \{\mathbb{X}, \mathbb{Z}_1\}$ to be our set of observed data, and $D_{\text{C}} = \{\mathbb{X}, \mathbb{Z}\}$ to be our completed data. We can furthermore denote the observed data

corresponding to labelled observations by $D_{\text{L}} = \{\mathbb{X}_1, \mathbb{Z}_1\}$, and the data corresponding to unlabelled observations by $D_{\text{U}} = \{\mathbb{X}_2\}$.

Using the above notation, we can now look at the three species of classification. The first species is discriminant analysis, see McNicholas (2016a) for details. Discriminant analysis makes use of only labelled data to build a classifier. The likelihood function in the case of a discriminant analysis can be written as

$$\mathcal{L}_{\text{DA}}(\boldsymbol{\vartheta}|D_{\text{L}}) = \prod_{j=1}^{n_1}\prod_{g=1}^{G}\left[\pi_g f_g(\mathbf{x}_{1j}|\boldsymbol{\theta}_g)\right]^{z_{jg}^{(1)}}. \tag{2.2}$$

The second species is cluster analysis, and can take on one of two forms. The first form is the one that we will primarily consider, and makes use of only unlabelled data points and ignores the labelled points. In this case, the likelihood function is given by

$$\mathcal{L}_{\text{clust}}(\boldsymbol{\vartheta}|D_{\text{U}}) = \prod_{j=1}^{n_2}\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_{2j}|\boldsymbol{\theta}_g). \tag{2.3}$$

The second form of the cluster analysis is to utilize both labelled and unlabelled points, but treat the labelled points as unlabelled.

The third species is semi-supervised classification. This makes use of all of the observed data, D_{O}, and treats labelled and unlabelled points equally when building a classifier. The likelihood function for semi-supervised classification is given by the product of $\mathcal{L}_{\text{DA}}(\boldsymbol{\vartheta}|D_{\text{L}})$ and $\mathcal{L}_{\text{clust}}(\boldsymbol{\vartheta}|D_{\text{U}})$ to give

$$\mathcal{L}_{\text{semi}}(\boldsymbol{\vartheta}|D_{\text{O}}) = \prod_{j=1}^{n_1}\prod_{g=1}^{G}\left[\pi_g f_g(\mathbf{x}_{1j}|\boldsymbol{\theta}_g)\right]^{z_{jg}^{(1)}} \times \prod_{j=1}^{n_2}\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_{2j}|\boldsymbol{\theta}_g). \tag{2.4}$$

2.3 Fractionally Supervised Classification

Introduced by Vrbik and McNicholas (2015), FSC allows for a solution intermediate to the three species of classification. This is achieved by introducing the weight $\alpha_1 = \alpha$ to labelled observations, and $\alpha_2 = 1 - \alpha$ to unlabelled observations, where $0 \leq \alpha \leq 1$. Using these weights, the arguably most natural form of the weighted observed likelihood can be written as

$$\begin{aligned}\mathcal{L}_{\text{FSC}}(\boldsymbol{\vartheta}|D_{\text{O}}, \alpha) &= [\mathcal{L}_{\text{DA}}(\boldsymbol{\vartheta}|D_{\text{L}})]^{\alpha} \times [\mathcal{L}_{\text{clust}}(\boldsymbol{\vartheta}|D_{\text{U}})]^{1-\alpha} \\ &= \left[\prod_{j=1}^{n_1}\prod_{g=1}^{G}[\pi_g f_g(\mathbf{x}_{1j}|\boldsymbol{\theta}_g)]^{z_{jg}^{(1)}}\right]^{\alpha} \times \left[\prod_{j'=1}^{n_2}\sum_{h=1}^{H}\pi_h f_g(\mathbf{x}_{2j'}|\boldsymbol{\theta}_h)\right]^{1-\alpha}, \quad (2.5)\end{aligned}$$

where $z_{jg}^{(1)}$ is the gth element of $\mathbf{z}_j^{(1)}$. Although H does not necessarily have to equal G, we will make the assumption that $H = G$.

We can then write the complete-data log-likelihood function as

$$\ell(\boldsymbol{\vartheta}|D_{\text{c}}) = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{g=1}^{G}\alpha_i z_{jg}^{(i)}\left[\log(\pi_g) + \log(f_g(\mathbf{x}_{ij}|\boldsymbol{\theta}_g))\right]. \quad (2.6)$$

The expectation-maximization (EM) algorithm, Dempster *et al.* (1977) can then be used to maximize (2.6). The EM algorithm is an iterative algorithm that consists of an expectation step and the subsequent maximization of the expectation. We first initialize the parameters, and we denote this by $\boldsymbol{\vartheta}^{(0)}$. The $t+1$ iteration of the EM

algorithm proceeds as follows

$$\textbf{E Step: } \text{Calculate } Q\big(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)}\big) = \mathbb{E}_{\mathbb{Z}_2|\mathbb{X}}\big[\ell\left(\boldsymbol{\vartheta}|D_{\text{c}}\right)|D_{\text{o}}, \boldsymbol{\vartheta}^{(t)}\big] \tag{2.7a}$$

$$\textbf{M Step: } \text{Find } \underset{\boldsymbol{\vartheta}}{\arg\max}\ Q\big(\boldsymbol{\vartheta}|\boldsymbol{\vartheta}^{(t)}\big) \tag{2.7b}$$

Check for convergence. If the convergence criterion was not met, set $t = t + 1$ and repeat. (2.7c)

In the case of a Gaussian model, steps (2.7a) and (2.7b) simplify to

E Step: Update

$$\hat{z}_{jg}^{(2)} = \frac{\pi_g^{(t)}\phi(x_{2j}|\boldsymbol{\mu}_g^{(t)}, \boldsymbol{\Sigma}_g^{(t)})}{\sum_{g=1}^{G}\pi_g^{(t)}\phi(x_{2j}|\boldsymbol{\mu}_g^{(t)}, \boldsymbol{\Sigma}_g^{(t)})}. \tag{2.8}$$

Because the $z_{jg}^{(1)}$ are known, we set $\hat{z}_{jg}^{(1)} = z_{jg}^{(1)}$.

M Step: Update the estimates of π_g, $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$ by calculating

$$\pi_g^{(t+1)} = \frac{S_g}{\sum_{g=1}^{G} S_g} \tag{2.9a}$$

$$\boldsymbol{\mu}_g^{(t+1)} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\alpha_i\hat{z}_{jg}^{(i)}\mathbf{x}_{ij}}{S_g} \tag{2.9b}$$

$$\boldsymbol{\Sigma}_g^{(t+1)} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\alpha_i\hat{z}_{jg}^{(i)}(\mathbf{x}_{jg} - \boldsymbol{\mu}_g^{(t+1)})(\mathbf{x}_{jg} - \boldsymbol{\mu}_g^{(t+1)})'}{S_g} \tag{2.9c}$$

where $S_g = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\alpha_i\hat{z}_{jg}^{(i)}$.

We note that the three different species of classification fall out naturally as special cases of FSC. If $\alpha = 1$, then all of the weight is given to the labelled observations, and the unlabelled observations are ignored. In this case, we are performing discriminant analysis. If $\alpha = 0.5$, then the labelled and unlabelled observations are given equal

weight, and we are then performing semi-supervised classification. Finally, if $\alpha = 0$, then no weight is given to the labelled observations, and thus we are performing a cluster analysis.

One issue involved with FSC is the selection of the weight α. Vrbik and McNicholas (2015) looked at criteria to find an optimal weight. However, these criteria were considered undesirable as they would either always choose one of the three species, or were computationally expensive.

As the primary goal of the aforementioned paper was to look at clustering and classification performance, the authors considered taking candidate weights, and then choosing the optimal weight based on the adjusted Rand index (ARI; Hubert and Arabie, 1985). The ARI compares two different partitions of a dataset, and in the classification paradigm, a value of 1 would correspond to perfect classification, whereas a value of 0 indicates that the classification solution is essentially the same as randomly assigning the labels.

2.4 The Multivariate t-Distribution

An extension of Student's t-distribution, the p-dimensional t-distribution with ν degrees of freedom, location parameter $\boldsymbol{\mu}$ and scale matrix $\boldsymbol{\Sigma}$, arises from a special case of a normal scale mixture (Peel and McLachlan, 2000). Specifically, we can write the normal scale mixture as

$$\epsilon\phi(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) + (1 - \epsilon)\phi(\mathbf{x}|\boldsymbol{\mu}, \nu\boldsymbol{\Sigma}), \tag{2.10}$$

where $\phi(\cdot)$ denotes the multivariate Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and ϵ is small. We can then rewrite (2.10) as

$$\int \phi(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) dH(\omega),$$

where we take

$$H(\omega) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right)\left(\frac{\nu}{2}\right)^{\frac{\nu}{2}}} \omega^{\frac{\nu}{2}-1} \exp\left\{\frac{2\omega}{\nu}\right\}, \tag{2.11}$$

the probability density function of a gamma$(\nu/2, \nu/2)$ random variable, where $\Gamma(\cdot)$ is the gamma function. The resulting density for the multivariate t, then becomes

$$f_t(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{(\pi\nu)^{\frac{1}{2}p}\Gamma\left(\frac{\nu}{2}\right)\left[1 + \frac{\delta(\mathbf{x},\boldsymbol{\mu},\boldsymbol{\Sigma})}{\nu}\right]^{\frac{1}{2}(\nu+p)}}, \tag{2.12}$$

where $\delta(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is the squared Mahalanobis distance.

Maximum likelihood estimation for the t-distribution, in the context of model based clustering, utilizes the introduction of latent variables, W_{ig}. These variables are such that

$$W_{ig}|z_{ig} = 1 \sim \text{gamma}\left(\nu_g/2, \nu_g/2\right),$$

2.5 Parsimonious Models

The eigen decomposition of a matrix is a parametrization that is widely used in both mathematics and multivariate statistics. In the context of mixture models, we can write a covariance, or scale, matrix in the form

$$\boldsymbol{\Sigma}_g = \lambda_g \boldsymbol{\Lambda}_g \mathbf{D}_g \boldsymbol{\Lambda}_g{}',$$

where λ_g is a constant, $\mathbf{D}_g$ is a diagonal matrix with entries that are proportional to the eigenvalues, and $\mathbf{\Lambda}_g$ is a matrix of eigenvectors. We can then impose the following constraints

$$\lambda_g = \lambda,\ \mathbf{\Lambda}_g = \mathbf{\Lambda},\ \mathbf{\Lambda}_g = \mathbf{I},\ \mathbf{D}_g = \mathbf{D},\ \mathbf{D}_g = \mathbf{I},$$

where $\mathbf{I}$ is the identity matrix.

In Banfield and Raftery (1993), Celeux and Govaert (1995), and Fraley and Raftery (1998, 2002b), combinations of the above constraints were applied to the covariance matrices in a Gaussian mixture model to form a family of 14 different Gaussian parsimonious clustering models (GPCMs). Of these 14 models, 10 form the MCLUST family of models available in the R software (R Core Team, 2015) package `mclust` (Fraley and Raftery, 2002a; Fraley *et al.*, 2012).

Andrews and McNicholas (2012) considered an extension of the MCLUST family of models, to the t-distribution, called the tEIGEN family. Originally the tEIGEN family consisted of the same constraints for the scale matrix as the MCLUST family. These ten models, combined with either the equality or inequality of the degrees of freedom over each group led to twenty different models in the tEIGEN family. In Andrews and McNicholas (2012), the tEIGEN family was extended to include two additional scale matrix constraints, leading to a total of 24 different models. The current form of the tEIGEN package (Andrews *et al.*, 2015) in R supports all 14 GPCMs, and hence all 28 tEIGEN models. A table summarizing these 28 tEIGEN models is given in McNicholas (2016a), and shown in Table A.1 in Appendix A.

2.6 Model Selection Criteria

We now discuss a couple criteria that are commonly used to select an appropriate parsimonious model. The Bayesian information criterion (BIC; Schwarz, 1978) as an alternative to the Akaike information criterion (AIC; Akaike, 1974) for statistical model selection. The criterion is given by

$$\text{BIC} = 2\ell_{\text{obs}}(\boldsymbol{\vartheta}|D_{\text{O}}) - p\log N,$$

where ℓ_{obs} is the maximized observed likelihood, p is the number of free parameters, and N is the total number of data points. The BIC has been frequently used for parsimonious model selection such as in Fraley and Raftery (1998) and McNicholas and Murphy (2008).

Another criterion that is widely used is the integrated complete likelihood (ICL; Biernacki *et al.*, 2000), which penalizes the BIC for classification uncertainty. Approximated using the BIC, the ICL is given by

$$\text{ICL} \approx \text{BIC} - 2\sum_{i=1}^{n_g}\sum_{g=1}^{G}\text{MAP}(\hat{z}_{ig})\log(\hat{z}_{ig}),$$

where

$$\text{MAP}(\hat{z}_{ig}) = \begin{cases} 1 & \text{if } \arg\max_{h=1,\ldots,G}\{\hat{z}_{ih}\} = g, \text{and} \\ 0 & \text{otherwise.} \end{cases}.$$

Chapter 3

Methodology

3.1 Alternate Form of the Likelihood

We have already seen that the observed weighted likelihood can be written as

$$\mathcal{L}_{\text{FSC}}(\boldsymbol{\theta}|D_{\text{O}},\alpha) = \left[\prod_{j=1}^{n_1}\prod_{g=1}^{G}[\pi_g f_g(\mathbf{x}_{1j}|\boldsymbol{\theta}_g)]^{z_{jg}^{(1)}}\right]^{\alpha_1} \times \left[\prod_{j'=1}^{n_2}\sum_{g=1}^{G}\pi_g f_g(\mathbf{x}_{2j'}|\boldsymbol{\theta}_g)\right]^{\alpha_2},$$

and the complete weighted likelihood can be written as

$$\mathcal{L}_{\text{comp}}(\boldsymbol{\theta}|D_{\text{C}},\alpha) = \prod_{i=1}^{2}\left[\prod_{j=1}^{n_i}\prod_{g=1}^{G}[\pi_g f_g(\mathbf{x}_{ij}|\boldsymbol{\theta})]^{z_{jg}^{(i)}}\right]^{\alpha_i}. \tag{3.1}$$

One of the properties in Dempster *et al.* (1977) states that when integrating the complete likelihood over the space of unknown quantities, in our case $\mathbb{Z}_2$, the result is the observed likelihood. The observed likelihood as given in (3.1), however, does

not hold this property. Indeed,

$$\begin{aligned}\int_{\mathbb{Z}_2} \mathcal{L}_{\text{comp}}(\boldsymbol{\vartheta}|D_{\text{C}},\alpha)d\mathbf{z}_2 &= \int_{\mathbb{Z}_2} \left(\prod_{j=1}^{n_1}\prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{1j}|\boldsymbol{\theta}_g)]^{z_{jg}^{(1)}\alpha} \times \prod_{j=1}^{n_2}\prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{2j}|\boldsymbol{\theta}_g)]^{z_{jg}^{(2)}(1-\alpha)} \right) d\mathbf{z}_2 \\ &= \prod_{j=1}^{n_1}\prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{1j}|\boldsymbol{\theta})]^{z_{jg}^{(1)}\alpha} \prod_{j=1}^{n_2} \left(\sum_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{2j}|\boldsymbol{\theta})]^{(1-\alpha)} \right), \quad (3.2)\end{aligned}$$

Clearly, this is not the same as the form given in (3.1). We therefore propose, in order to maintain the relationship between the complete and incomplete weighted likelihood as presented in Dempster *et al.* (1977), using the form of the incomplete weighted likelihood as given in (3.2) and denote this by $\mathcal{L}_{\text{alt}}$. A mathematical derivation of (3.2) is given in Appendix B.

We do note that there are two extreme cases that should be considered separately. The first extreme case is when $\alpha = 0$. In this case,

$$\int_{\mathbb{Z}_2} \mathcal{L}_{\text{comp}}(\boldsymbol{\vartheta}|D_{\text{C}},\alpha=0)d\mathbf{z}_2 = \int_{\mathbb{Z}_2} \prod_{j=1}^{n_2}\prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{2j}|\boldsymbol{\theta}_g)]^{z_{jg}^{(2)}(1-\alpha)}\, d\mathbf{z}_2 = \prod_{j=1}^{n_2}\sum_{g=1}^{G} \pi_g f_g(\mathbf{x}_{2j}|\boldsymbol{\theta}),$$

which is equivalent to (3.2) when $\alpha = 0$.

The second extreme case, which turns out to be more interesting, is when $\alpha = 1$. In this case,

$$\int_{\mathbb{Z}_2} \mathcal{L}_{\text{comp}}(\boldsymbol{\vartheta}|D_{\text{C}},\alpha=1)d\mathbf{z}_2 = \int_{\mathbb{Z}_2} \prod_{j=1}^{n_1}\prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{1j}|\boldsymbol{\theta}_g)]^{z_{jg}^{(1)}} d\mathbf{z}_2 = \prod_{j=1}^{n_1}\prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{1j}|\boldsymbol{\theta}_g)]^{z_{jg}^{(1)}},$$

which is the same as $\mathcal{L}_{\text{DA}}$, the observed likelihood for a discriminant analysis. However,

in (3.2), when $\alpha = 1$,

$$\mathcal{L}_{\text{alt}}(\boldsymbol{\vartheta}|D_{\text{O}}) = n_2 \prod_{j=1}^{n_1} \prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{1j}|\boldsymbol{\theta}_g)]^{z_{jg}^{(1)}} = n_2 \mathcal{L}_{\text{DA}}(\boldsymbol{\vartheta}|D_{\text{L}}). \tag{3.3}$$

When $\alpha = 1$ we are performing a discriminant, analysis, and so the form of the observed and weighted likelihoods should be the same, which is clearly not the case. Therefore, we propose when $\alpha = 1$ to use $\mathcal{L}_{\text{DA}}$ for our observed likelihood.

For both the original and altered observed likelihood, the complete likelihood is identical. Therefore, if we were to take a Gaussian model, the updates in the M step would be the same as those given in Chapter 2.3 regardless of taking the original or altered likelihood However, the updates for $\hat{z}_{jg}^{(2)}$ in the E step, would become

$$\hat{z}_{jg}^{(2)} = \frac{\left[\pi_g^{(t)} \phi(\mathbf{x}_{2j}|\boldsymbol{\mu}_g^{(t)}, \boldsymbol{\Sigma}_g^{(t)})\right]^{(1-\alpha)}}{\sum_{g=1}^{G} \left[\pi_g^{(t)} \phi(\mathbf{x}_{2j}|\boldsymbol{\mu}_g^{(t)}, \boldsymbol{\Sigma}_g^{(t)})\right]^{(1-\alpha)}}.$$

3.1.1 Simulation Comparing the Original and Altered Likelihoods

We performed simulations to compare the performance of the original and the proposed altered likelihood. We simulated 100 datasets with 300 samples. 150 of these samples belonged to one group which followed a $\mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}_1)$, and the remaining 150 belonged to another group which followed a $\mathcal{N}_2(\boldsymbol{\Delta}, \boldsymbol{\Sigma}_2)$, where

$$\boldsymbol{\Delta} = [0, \Delta]',$$

and

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

We took $\Delta \in \{1, 5\}$ corresponding to different levels of clustering difficulty. For each dataset we looked at $p \in \{10, 20, \ldots, 80, 90\}$, where p is the percentage of labelled data.

To choose the weights for FSC, we looked at 11 different values of α. These values were taken to be $\alpha \in \boldsymbol{\alpha}_{\text{ari}}$ where $\boldsymbol{\alpha}_{\text{ari}} = \{0, 0.1, 0.2, \ldots, 1\}$. We then calculated the ARI for each of these weights for the 100 datasets and took the average ARI for each weight. We then choose the weight that had the highest average ARI. We will denote the resulting FSC solution for each weight α by FSC_α. Furthermore, for the FSC solution with the chosen weight resulting from the highest average ARI, we will denote by FSC_{ARI}. Finally, in the special cases corresponding to the three species of classification $\alpha = 0, 0.5, 1$, we denote the FSC solution by $\text{FSC}_{\text{clust}}$, $\text{FSC}_{\text{class}}$ and FSC_{DA} respectively.

In Figures 3.1 and 3.2, we show different line plots for $\Delta = 1$ and $\Delta = 5$ respectively. In each plot, we show the average ARI against the percentage of labelled data p. We also show in a dotted black line the result for FSC_{ARI} with the corresponding chosen weight shown above each point. The first row in each plot shows the results when using all the weights, and the second row singles out the three different species of classification and FSC_{ARI} The standard errors were calculated by taking the ARI for all 100 datasets of the chosen weight of FSC_{ARI} and calculating one (darker grey) and two (lighter grey) standard deviations from the mean ARI.

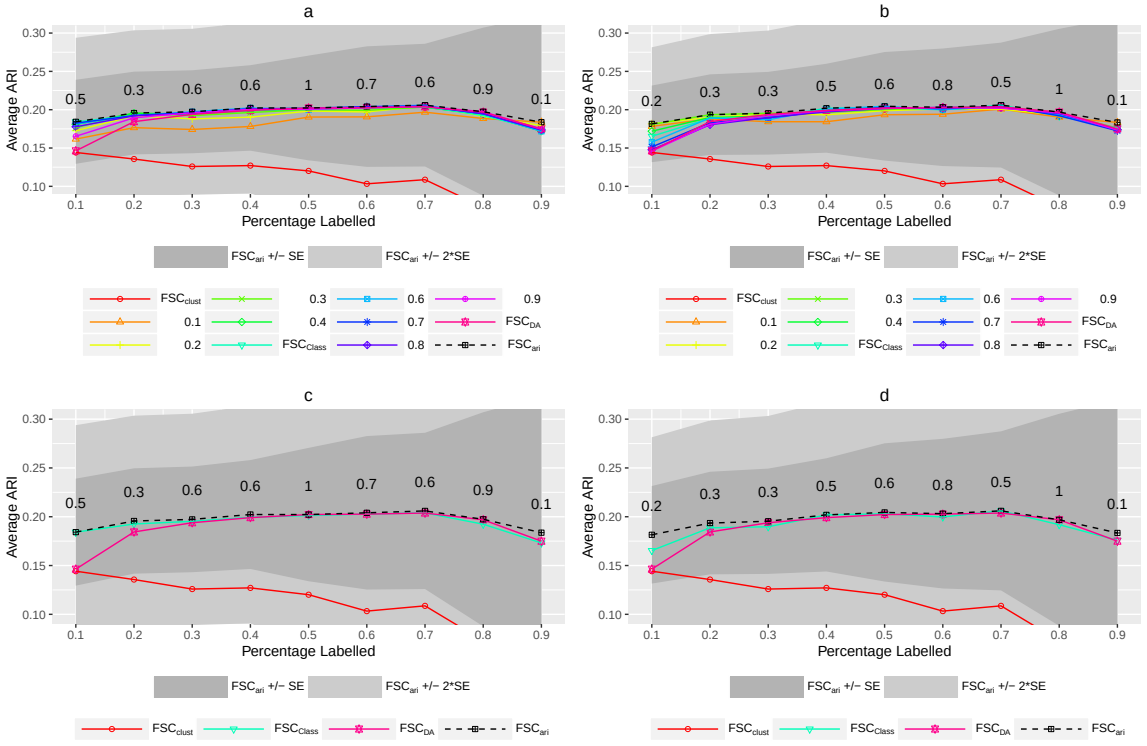

Figure 3.1: For $\Delta = 1$: (a) and (b) FSC_{α} and FSC_{ARI} ($\alpha \in \boldsymbol{\alpha}_{\text{ARI}}$) for the original and altered likelihood respectively. (c) and (d) $\text{FSC}_{\text{clust}}$, $\text{FSC}_{\text{class}}$, FSC_{DA} and FSC_{ARI} for the original and altered likelihood respectively.

We see that, in general, the overall classification performance between the altered and original likelihoods are similar. The chosen weights for FSC_{ARI}, however, do differ in general between the two forms of the likelihood. We notice for $\Delta = 1$, this difference is less pronounced than in the $\Delta = 5$ case. More specifically, for $\Delta = 1$, the difference between the weights for all but 10%, 30% and 50% differ by at most 0.1, if not exactly the same. In the $\Delta = 5$ case, however, the differences between the chosen weights are greater, and there are fewer proportions for which the difference is small. We also see that at lower percentages of labelled data, there is more variability in the average ARI between the different weights.

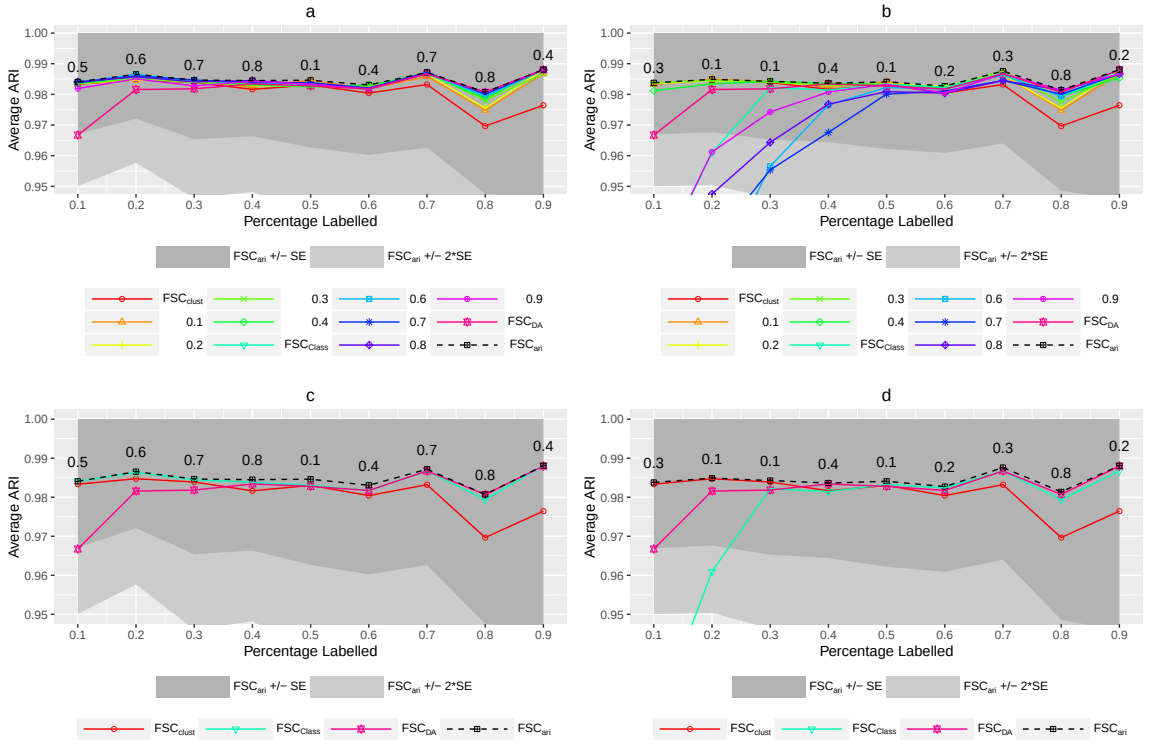

Figure 3.2: For $\Delta = 5$: (a) and (b) FSC_{α} and FSC_{ARI} ($\alpha \in \boldsymbol{\alpha}_{\text{ARI}}$) for the original and altered likelihood respectively. (c) and (d) $\text{FSC}_{\text{clust}}$, $\text{FSC}_{\text{class}}$, FSC_{DA} and FSC_{ARI} for the original and altered likelihood respectively.

In conclusion, although the choice of the weights are different between the two likelihoods, the overall classification performance when using the chosen weight in each case are the same. Moreover, the altered form is not technically a proper likelihood. We therefore will henceforth use the original form of the likelihood as it is the more natural form, and the altered form does not result in significantly higher classification performance.

3.2 Extension of FSC to the Multivariate t-Distribution

We now look at extending the concept of FSC to non Gaussian mixture models. There are many possible extensions, however we felt that the most natural extension would be the multivariate t-distribution. The main complication when compared to using a Gaussian mixture is the update for the degrees of freedom. This update, unfortunately, has no closed form and has to be calculated using numerical methods.

The incomplete weighted observed likelihood when using multivariate t component densities would be given by

$$\mathcal{L}_{\text{obs}}(\boldsymbol{\vartheta}|D_{\text{O}},\alpha)=\left[\prod_{j=1}^{n_1}\prod_{g=1}^{G}[\pi_g f_t(\mathbf{x}_{1j}|\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g,\nu_g)]\right]^{\alpha}\times\left[\prod_{j'=1}^{n_2}\sum_{g=1}^{G}\pi_g f_t(\mathbf{x}_{2j'}|\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)\right]^{1-\alpha},$$

where $f_t(\cdot)$ is the density for the multivariate t-distribution defined in (2.12). To find $\arg\max_{\boldsymbol{\theta}}\mathcal{L}_{\text{obs}}$, we use a multicycle ECM algorithm similar to Andrews and McNicholas (2012). After initializing $z_{jg}^{(i)}$ and $\omega_{jg}^{(i)}$, the $t+1$ iteration of the multicycle ECM algorithm would proceed as follows:

E Step: Update

$$\hat{z}_{jg}^{(2)}=\frac{\hat{\pi}_g f_t(\mathbf{x}_{2j}|\hat{\boldsymbol{\mu}}_g^{(t)},\hat{\boldsymbol{\Sigma}}_g^{(t)},\hat{\nu}_g^{(t)})}{\sum_{g=1}^{G}\hat{\pi}_g f_t(\mathbf{x}_{2j}|\hat{\boldsymbol{\mu}}_g^{(t)},\hat{\boldsymbol{\Sigma}}_g^{(t)},\hat{\nu}_g^{(t)})} \tag{3.4a}$$

$$\hat{\omega}_{jg}^{(i)}=\frac{\hat{\nu}_g^{(t)}+p}{\hat{\nu}_g^{(t)}+\delta(\mathbf{x}_{ij},\hat{\boldsymbol{\mu}}_g^{(t)},\hat{\boldsymbol{\Sigma}}_g^{(t)},\hat{\nu}_g^{(t)})} \tag{3.4b}$$

First CM Step: Update $\hat{\pi}_g,\hat{\boldsymbol{\mu}}_g$ and $\hat{\nu}_g$. The updates for $\hat{\pi}_g$, and $\hat{\boldsymbol{\mu}}_g$ are given in

closed form as

$$\hat{\pi}_g^{(t+1)} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\alpha_i\hat{z}_{jg}^{(i)}}{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{g=1}^{G}\alpha_i\hat{z}_{jg}^{(i)}} \tag{3.5}$$

and

$$\hat{\boldsymbol{\mu}}_g^{(t+1)} = \frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\alpha_i\hat{z}_{jg}^{(i)}\hat{\omega}_{jg}^{(i)}\mathbf{x}_{ij}}{\sum_{i=1}^{2}\sum_{j=1}^{n_i}\alpha_i\hat{z}_{jg}^{(i)}\hat{\omega}_{jg}^{(i)}} \tag{3.6}$$

The updates for the degrees of freedom ν_g, as mentioned before, do not have a closed form and have to be calculated using numerical methods. In the unconstrained case one has to solve (3.7) for $\hat{\nu}_g^{\text{new}}$.

$$\begin{aligned}-\Psi\left(\frac{1}{2}\hat{\nu}_g^{\text{new}}\right) + \log\left(\frac{1}{2}\hat{\nu}_g^{\text{new}}\right) - \Psi\left(\frac{\hat{\nu}_g + p}{2}\right) - \log\left(\frac{\hat{\nu}_g + p}{2}\right) + 1 \\ + \frac{1}{m_g}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\alpha_i\hat{z}_{jg}^{(i)}\left(\log\hat{\omega}_{jg}^{(i)} - \hat{\omega}_{jg}^{(i)}\right) = 0\end{aligned} \tag{3.7}$$

where

$$m_g = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\alpha_i\hat{z}_{jg}^{(i)}$$

and $\Psi(\cdot)$ is the digamma function. We then set $\hat{\nu}_g^{(t+1)} = \hat{\nu}_g^{\text{new}}$. We note that we used the `uniroot` function available in R to solve (3.7).

E Step: Update $\hat{z}_{jg}^{(2)}$ and $\hat{\omega}_{jg}^{(i)}$ using (3.4a) and (3.4b) with the current parameter estimates.

Second CM Step: Update $\boldsymbol{\Sigma}_g$. In the completely unconstrained case, the update

is given by

$$\hat{\boldsymbol{\Sigma}}_g^{(t+1)} = \frac{1}{m_g} \sum_{i=1}^{2} \sum_{j=1}^{n_i} \sum_{g=1}^{G} \alpha_i \hat{z}_{jg}^{(i)} \hat{\omega}_{jg}^{(i)} (\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_g^{(t+1)})(\mathbf{x}_{ij} - \hat{\boldsymbol{\mu}}_g^{(t+1)})' \tag{3.8}$$

We performed k-means clustering (MacQueen, 1967) with 50 random starts to initialize the ECM algorithm, and the Aitken acceleration procedure described in McNicholas *et al.* (2010) as our convergence criteria.

Because of the updates for the degrees of freedom, fitting FSC with a t mixture becomes more computationally expensive than fitting a Gaussian model. However, because of the heavier tails of the t-distribution, the t mixture would be more robust to outlying observations.

3.3 Weight Selection Criteria

We have already discussed using the ARI as a weight selection criteria. This, however, is only useful when exploring the overall performance of FSC in simulations and datasets where all the labels are known. In a general dataset, not all the labels would be known and hence the ARI would not be applicable. We therefore propose other criteria for weight selection.

The first criteria we considered was the BIC or ICL. We do not show the results here, but both of these criteria were seen to be monotone in α, and a boundary point would always be chosen. We looked at three different classification based criteria, the mean entropy and an alternate form of the entropy, Celeux and Soromenho (1996) and a U criterion, Bensmail *et al.* (1997).

In our case, the entropy E, can be written as:

$$E = \sum_{i=1}^{2}\sum_{j=1}^{n_i}\sum_{g=1}^{G} \text{MAP}(\hat{z}_{jg}^{(i)}) \log \hat{z}_{jg}^{(i)} = \sum_{j=1}^{n_2}\sum_{g=1}^{G} \text{MAP}(\hat{z}_{jg}^{(2)}) \log \hat{z}_{jg}^{(2)},$$

where

$$\text{MAP}(z_{jg}^{(i)}) = \begin{cases} 1 & \text{if } \hat{z}_{jg}^{(i)} = \max\limits_{g=1,2,\ldots,G}(\hat{z}_{jg}^{(i)}) \\ 0 & \text{otherwise} \end{cases}$$

and taking $0 \log 0 = 0$. This criterion is always negative, and no uncertainty in the clustering solution would result in an entropy of 0. Therefore when using this criterion, we would maximize E to choose the optimal weight.

An alternate form of the entropy is sometimes used that eliminates the MAP, and the resulting criterion, in our case, is given by

$$\text{AE} = \sum_{j=1}^{n_2}\sum_{g=1}^{G} \hat{z}_{jg}^{(2)} \log \hat{z}_{jg}^{(2)}.$$

Once again, we wish to maximize this criterion to find the optimal weight.

The last classification based criterion that we looked at was the U criterion. In our case, this is given by

$$U = \sum_{i=1}^{2}\sum_{j=1}^{n_i} \min_{g=1,2,\ldots,G}(1 - \hat{z}_{jg}^{(i)}) = \sum_{j=1}^{n_2} \min_{g=1,2,\ldots,G}(1 - \hat{z}_{jg}^{(2)}).$$

We observe that U is always positive and if there is no uncertainty in the classification solution, U would be 0. We would thus like to minimize U to find the optimal weight.

In addition to these three different classification based criteria, we considered two non parametric criteria. Before the BIC was introduced in the literature, the sum

of squares matrix was considered as a criteria to choose the number of groups in a model, for example Friedman and Rubin (1967). Assuming that our data matrix $\mathbf{X}$ has been partitioned into G groups, we can define the total sum of squares matrix to be

$$\mathbf{S} = \sum_{i=1}^{n_g} \sum_{g=1}^{G} (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)'.$$

Using a decomposition of $\mathbf{S}$ we can write

$$\mathbf{S} = \mathbf{W} + \mathbf{B},$$

where $\mathbf{W}$ is the within cluster sum of square matrix defined as

$$\mathbf{W} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)',$$

where $\bar{\mathbf{x}}_g$ is the sample mean of group g, and $\mathbf{B}$ is the between cluster sum of squares matrix defined as

$$\mathbf{B} = \sum_{g=1}^{G} (\bar{\mathbf{x}}_g - \bar{\mathbf{x}})(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})',$$

where $\bar{\mathbf{x}}$ is the grand mean.

Although the principle of using the sum of squares matrix was considered all the way back in the 1960's, it is still seen in the modern literature such as Andrews and McNicholas (2014). We therefore propose two different criteria using the within cluster sum of squares matrix $\mathbf{W}$. The first criterion is to take the trace of $\mathbf{W}$ and the second criterion is to take the determinant.

3.4 Specifying the Number of Groups

For the purposes of our simulations and data analyses, we assume that the number of groups are equal to the number of components present in the known labels. However, this could be potentially problematic as there could be a group present in the population that is not represented in the labelled data, for details see McNicholas (2016a). This is especially prevalent if only a small proportion of the data points are labelled. On the other hand, it is also possible for the true number of groups to be less than that indicated by the labels. This can occur when the labels identify an underlying subgroup structure within a group. The former case can be handled by fitting FSC with a different number of groups $H \geq G$ in the cluster analysis component of the likelihood, and then using a criterion such as the BIC or ICL to choose the number of groups. The latter case, however, would need to be treated more carefully.

Chapter 4

Analyses

4.1 Simulations

We performed simulations similar to those in Chapter 3.1.1 . We simulated two groups each with 200 samples. The first group followed a $t_2(\mathbf{0}, \mathbf{\Sigma}_1, \nu_1)$ distribution where $\nu_1 = 3$, and

$$\mathbf{\Sigma}_1 = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}.$$

The second group was taken from a $t_2(\mathbf{\Delta}, \mathbf{\Sigma}_2, \nu_2)$ distribution where $\mathbf{\Delta} = [0, \Delta]'$, $\nu_2 = 70$, and

$$\mathbf{\Sigma}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In this case, one group has a multivariate t-distribution, while the other group is approximately normal since ν_2 is quite large. This time, we took $\Delta \in \{1, 2, 3, 4, 5\}$, and we took the same percentages of labelled data, p, as we did previously. In Figure

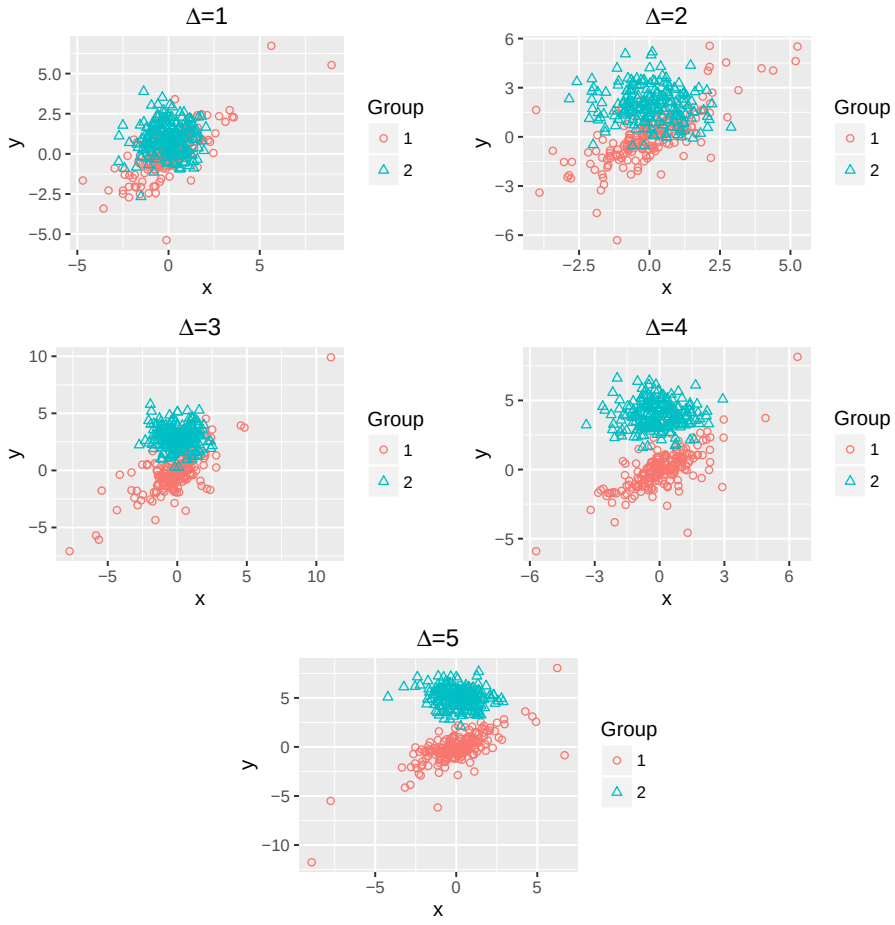

Figure 4.1: Typical datasets for each Δ.

4.1, we show example datasets for each Δ. The weight α for FSC_{ARI} was chosen in the same manner as in Section 3.1.

In Figure 4.2 we show line plots similar to what was shown in 3.1.1 for $\Delta = 1$. On the left hand side, we show the average ARI against the percentage of labelled data for all of the weights, the three species and FSC_{ARI}. On the right hand side, we isolate the three species of classification, and FSC_{ARI}.

For $\Delta = 1$, we notice that the line for $\text{FSC}_{\text{clust}}$ does not appear as the average ARI for each percentage of labelled data is quite small in comparison to the other weights.

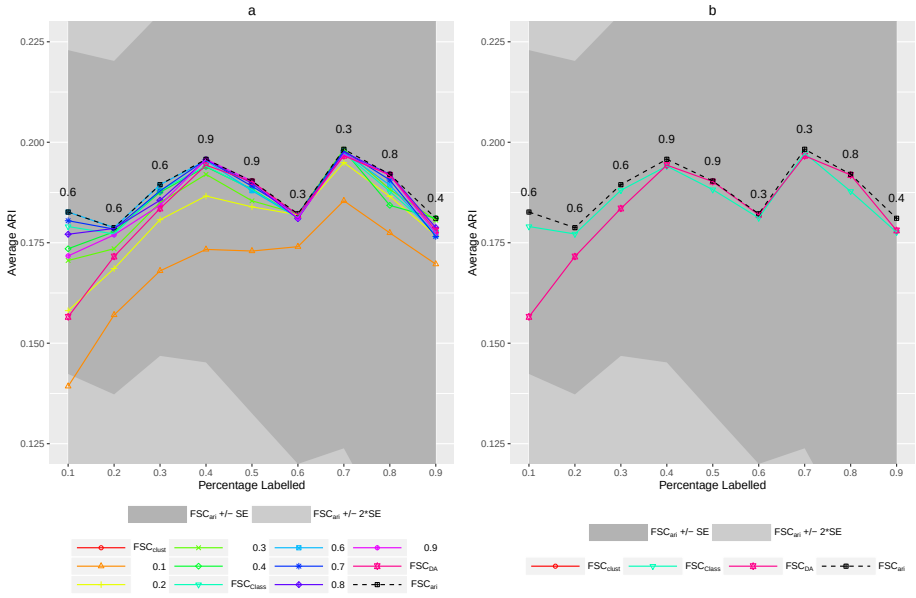

Figure 4.2: For $\Delta = 1$: a) FSC_{α} and FSC_{ARI} for $\alpha \in \boldsymbol{\alpha}_{\text{ARI}}$, b) $\text{FSC}_{\text{clust}}$,$\text{FSC}_{\text{class}}$, FSC_{DA} and FSC_{ARI}.

Furthermore, for all other values of Δ, $\text{FSC}_{\text{clust}}$ has the worst performance at higher percentages of labelled data, which is somewhat expected.

For $\Delta = 1$, we see that all of the chosen weights correspond to a non-species solution. Furthermore, it is interesting to point out that for lower percentages of labelled data, more weight is given to the labelled points, and at higher percentages, with the exception of 80%, less weight is given to the labelled observations.

For the remaining values of Δ, of the 36 different cases the chosen weight corresponds to a species of classification only 9 times. Of these 9 occurrences, 8 of them correspond to semi-supervised classification, one corresponds to a discriminant analysis, and none of them correspond to a cluster analysis.

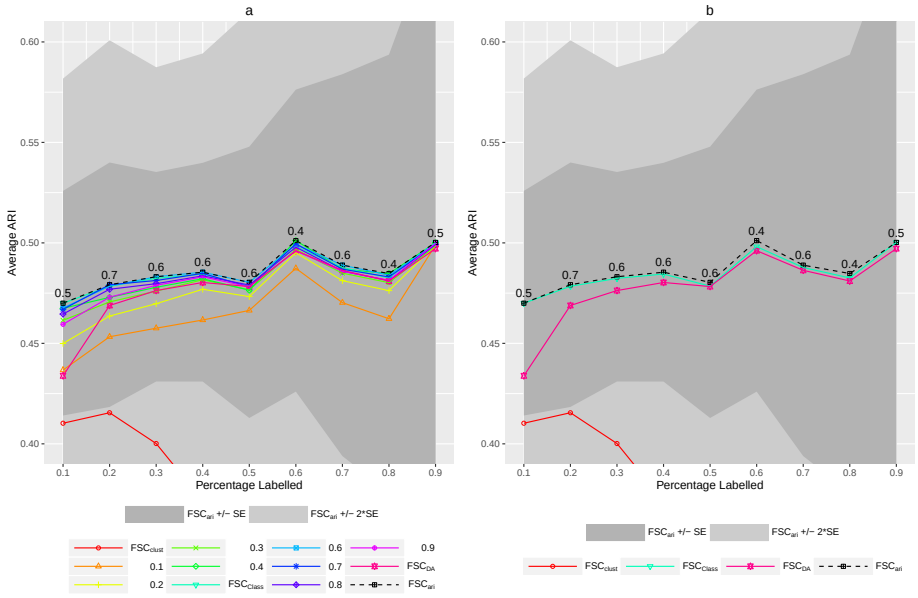

Figure 4.3: For $\Delta = 2$: a) FSC_{α} and FSC_{ARI} for $\alpha \in \boldsymbol{\alpha}_{\text{ARI}}$, b) $\text{FSC}_{\text{clust}}$,$\text{FSC}_{\text{class}}$, FSC_{DA} and FSC_{ARI}.

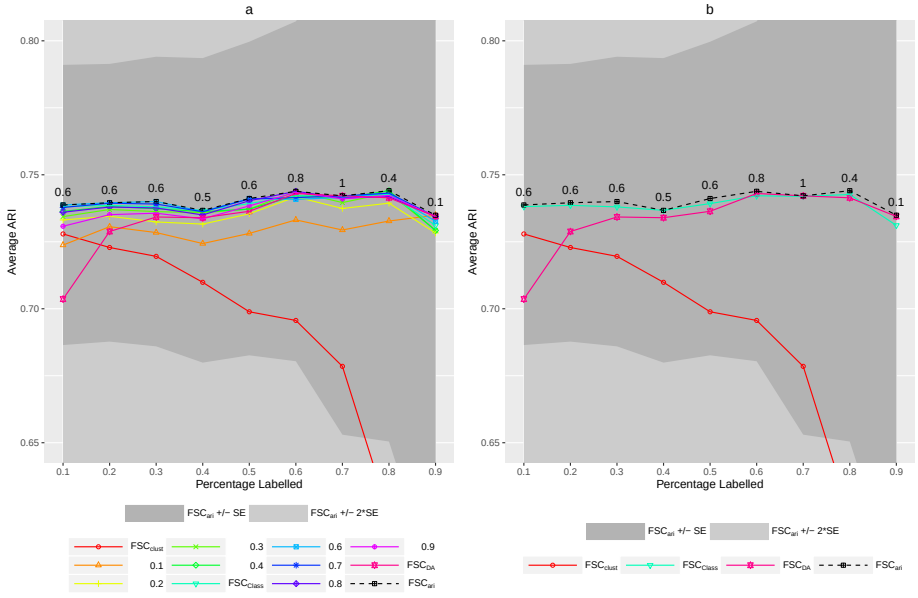

Figure 4.4: For $\Delta = 3$: a) FSC_{α} and FSC_{ARI} for $\alpha \in \boldsymbol{\alpha}_{\text{ARI}}$, b) $\text{FSC}_{\text{clust}}$,$\text{FSC}_{\text{class}}$, FSC_{DA} and FSC_{ARI}.

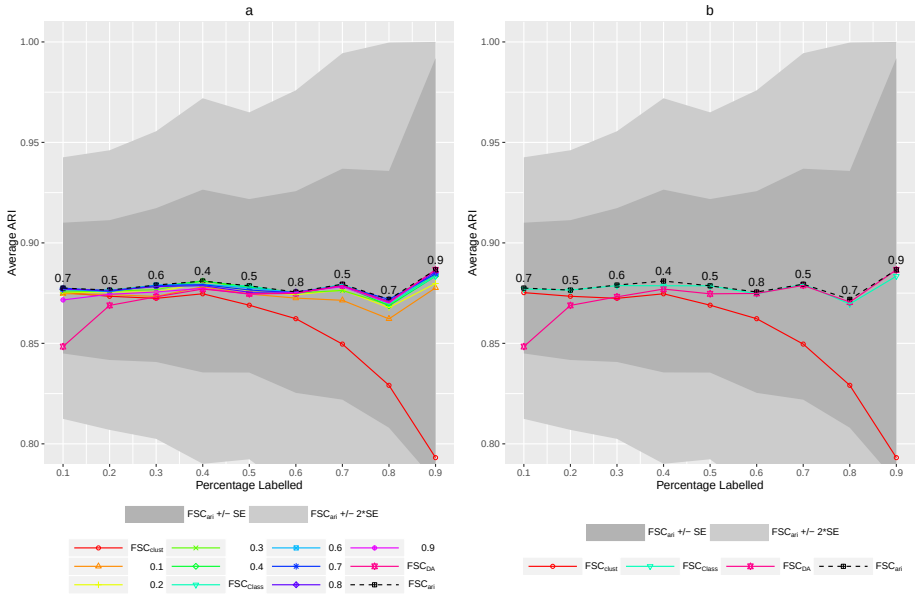

Figure 4.5: For $\Delta = 4$: a) FSC_{α} and FSC_{ARI} for $\alpha \in \boldsymbol{\alpha}_{\text{ARI}}$, b) $\text{FSC}_{\text{clust}}$,$\text{FSC}_{\text{class}}$, FSC_{DA} and FSC_{ARI}.

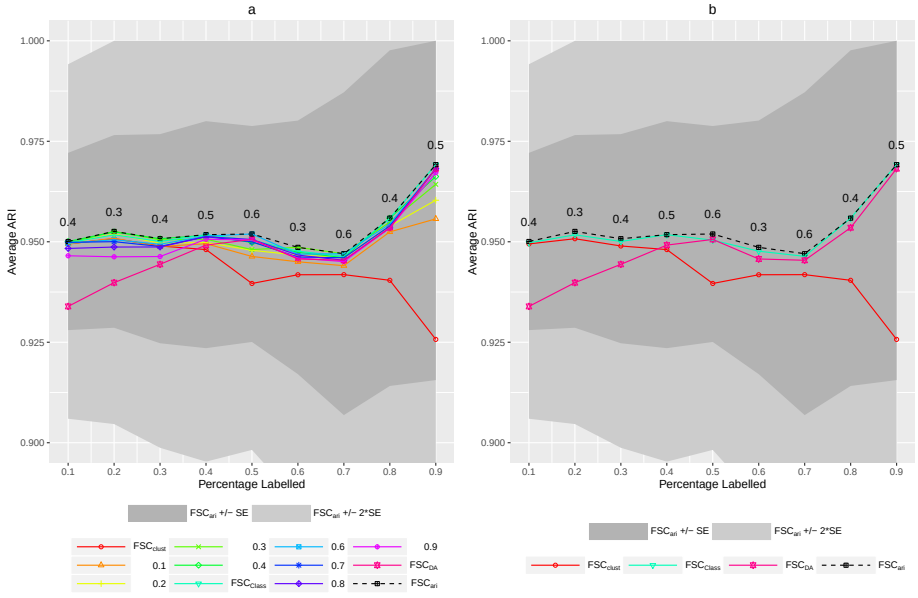

Figure 4.6: For $\Delta = 5$: a) FSC_{α} and FSC_{ARI} for $\alpha \in \boldsymbol{\alpha}_{\text{ARI}}$, b) $\text{FSC}_{\text{clust}}$,$\text{FSC}_{\text{class}}$, FSC_{DA} and FSC_{ARI}.

Estimation

In addition to classification performance, we also considered the accuracy of the estimates. We look at the parameter estimates for FSC_{ARI} from our previous simulation, for p =20%, 50%, and 80% of points labelled, and $\Delta = 3$. These results are seen in Table 4.1. We observe that in all cases the estimates are very close to the actual values. We do note that there is a lot of variability in the estimate for ν_2. However, this is to be expected because the t-distribution is asymptotically Gaussian, and because the second component is approximately Gaussian, we can expect a large amount of variability in this estimate.

Table 4.1: Average parameter estimates for $\Delta = 3$ for 20%, 50% and 80% of points labelled with component wise standard deviations in brackets

20% ($\alpha = 0.6$)					
Group 1			Group 2		
ν_1 (sd)	$\boldsymbol{\mu}_1$ (sd)	$\boldsymbol{\Sigma}_1$ (sd)	ν_2 (sd)	$\boldsymbol{\mu}_2$ (sd)	$\boldsymbol{\Sigma}_2$ (sd)
3.21 (0.766)	$\begin{bmatrix} -0.00698 \\ -0.00352 \end{bmatrix}$ $\left(\begin{bmatrix} 0.100 \\ 0.100 \end{bmatrix}\right)$	$\begin{bmatrix} 1.01 & 0.703 \\ 0.703 & 1.01 \end{bmatrix}$ $\left(\begin{bmatrix} 0.200 & 0.154 \\ 0.154 & 0.184 \end{bmatrix}\right)$	63.2 (57.0)	$\begin{bmatrix} 0.00535 \\ 2.99 \end{bmatrix}$ $\left(\begin{bmatrix} 0.0772 \\ 0.0845 \end{bmatrix}\right)$	$\begin{bmatrix} 0.988 & -0.00720 \\ -0.00720 & 0.978 \end{bmatrix}$ $\left(\begin{bmatrix} 0.133 & 0.0881 \\ 0.0881 & 0.138 \end{bmatrix}\right)$
50% ($\alpha = 0.6$)					
3.19 (0.742)	$\begin{bmatrix} 0.00186 \\ 0.00476 \end{bmatrix}$ $\left(\begin{bmatrix} 0.0956 \\ 0.0913 \end{bmatrix}\right)$	$\begin{bmatrix} 1.03 & 0.716 \\ 0.716 & 1.03 \end{bmatrix}$ $\left(\begin{bmatrix} 0.195 & 0.143 \\ 0.143 & 0.170 \end{bmatrix}\right)$	67.3 (57.4)	$\begin{bmatrix} -0.00270 \\ 3.00 \end{bmatrix}$ $\left(\begin{bmatrix} 0.0760 \\ 0.0799 \end{bmatrix}\right)$	$\begin{bmatrix} 0.990 & 0.000940 \\ 0.000940 & 0.980 \end{bmatrix}$ $\left(\begin{bmatrix} 0.127 & 0.0811 \\ 0.0811 & 0.140 \end{bmatrix}\right)$
80% ($\alpha = 0.4$)					
3.20 (0.716)	$\begin{bmatrix} -0.00242 \\ 0.00122 \end{bmatrix}$ $\left(\begin{bmatrix} 0.0951 \\ 0.0906 \end{bmatrix}\right)$	$\begin{bmatrix} 1.02 & 0.712 \\ 0.712 & 1.02 \end{bmatrix}$ $\left(\begin{bmatrix} 0.194 & 0.149 \\ 0.149 & 0.170 \end{bmatrix}\right)$	67.1 (52.7)	$\begin{bmatrix} -0.00254 \\ 3.00 \end{bmatrix}$ $\left(\begin{bmatrix} 0.0730 \\ 0.0737 \end{bmatrix}\right)$	$\begin{bmatrix} 0.995 & -0.00215 \\ -0.00215 & 0.978 \end{bmatrix}$ $\left(\begin{bmatrix} 0.124 & 0.0804 \\ 0.0804 & 0.120 \end{bmatrix}\right)$

4.2 Applications to Datasets

We now look at a few datasets and compare the performance of using a t mixture and a Gaussian mixture to model the data. We took 100 random splits for each dataset for each percentage of labelled data, $p \in \{10, 20, \ldots, 80, 90\}$. We used the same criterion (ARI) as in the simulations to choose the optimal weight. As with the simulations we used a completely unconstrained model for both the covariance structure and, in the case of the t-distribution, the degrees of freedom. We note that for lower percentages for some of the datasets, we were not able to perform a discriminant analysis, and for higher percentages we were not able to perform a cluster analysis.

Iris Data

We first looked at the Anderson Iris data in the R package `datasets`. This dataset looked at four different attributes of three different species of iris. The measurements (in centimetres) were the sepal length and width, and the petal length and width. In Figure 4.7 we show the line plot. On the left hand side we show the results for the t mixture, and on the right hand side we show the results for the Gaussian. Comparing these two plots, we see that the overall classification performance is similar between the t mixture and the Gaussian mixture. Moreover, except at $p = 60\%$, the weights chosen for both the t and Gaussian mixtures are very similar if not exactly the same.

Crabs Data

The next dataset that we looked at was the crabs dataset in the R package `MASS`, Venables and Ripley (2002). It consisted of 5 measurements on four different types of

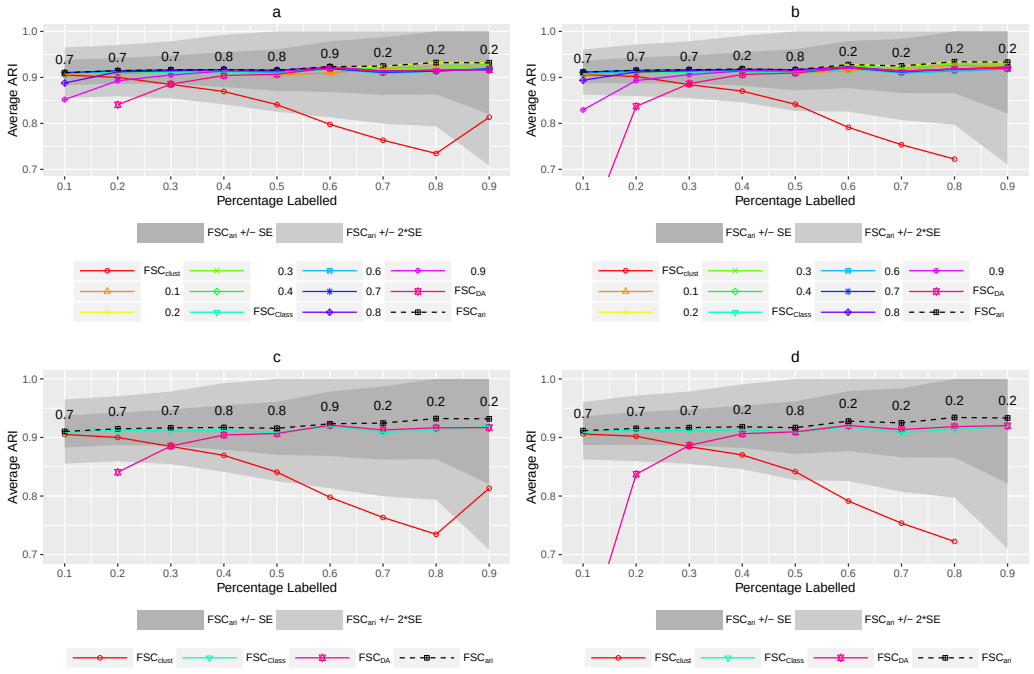

Figure 4.7: Iris Data: FSC_α for $\alpha \in \boldsymbol{\alpha}_{\text{ARI}}$ and FSC_{ARI} for a) the t mixture and b) for the Gaussian Mixture. $\text{FSC}_{\text{clust}}$,$\text{FSC}_{\text{class}}$, FSC_{DA} and FSC_{ARI} for c) the t mixture, and d) the Gaussian mixture.

rock crabs (two species, male and female in each species). These measurements were the frontal lobe size, carapace length and width, and the rear length and width. In Figure 4.8, we show similar plots to Figure 4.7. We see, like with the iris data, that the classification performance for the t and Gaussian mixtures are similar. Moreover, the weights chosen are very similar. It is interesting to note that almost all the weights are around 0.5

Wine Data

The third dataset we looked at was the 13 variable wine dataset in the R package `gclus` (Hurley, 2004). This dataset looked 13 characteristics of three different classes

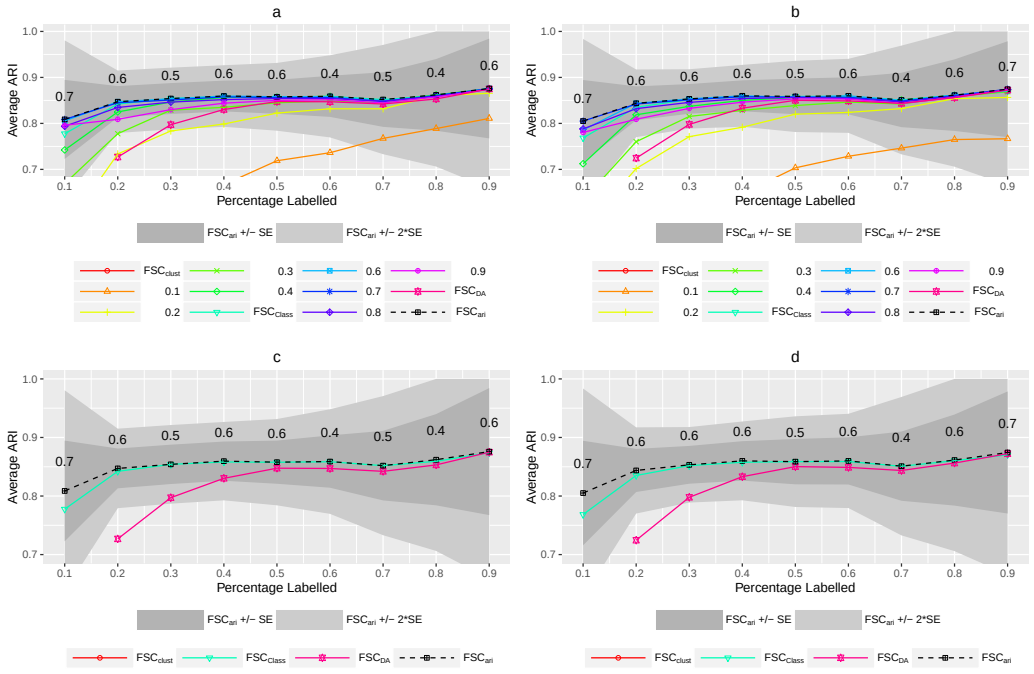

Figure 4.8: Crab Data: FSC_α for $\alpha \in \boldsymbol{\alpha}_{\text{ARI}}$ and FSC_{ARI} for a) the t mixture and b) for the Gaussian Mixture. $\text{FSC}_{\text{clust}}$,$\text{FSC}_{\text{class}}$, FSC_{DA} and FSC_{ARI} for c) the t mixture, and d) the Gaussian mixture.

of wine. We show the familiar line plots in Figure 4.9. One interesting aspect to note is that in general, (until one gets to the higher proportions of labelled data), the t mixture performs slightly better than the Gaussian. Another thing to note is that, like the crabs data the cluster analysis does not perform well at all in comparison to the other values of α. Finally, the chosen weights for the t and Gaussian models are fairly similar and tend to choose larger weights for the labelled observations at all proportions.

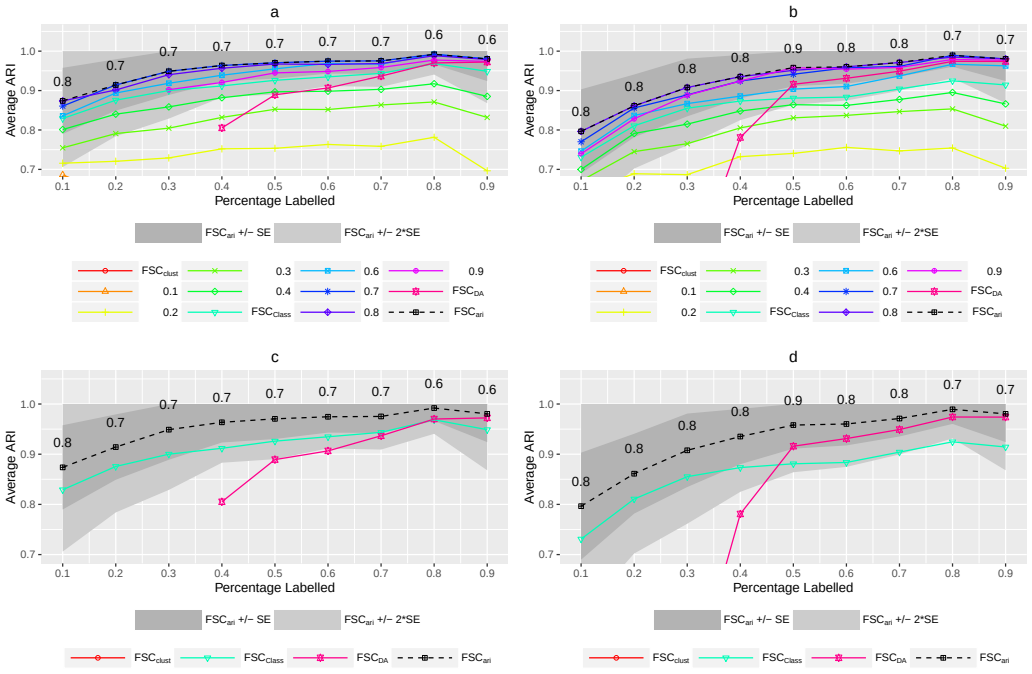

Figure 4.9: Wine Data: FSC_α for $\alpha \in \boldsymbol{\alpha}_{\text{ARI}}$ and FSC_{ARI} for a) the t mixture and b) for the Gaussian Mixture. $\text{FSC}_{\text{clust}}$,$\text{FSC}_{\text{class}}$, FSC_{DA} and FSC_{ARI} for c) the t mixture, and d) the Gaussian mixture.

Bankruptcy Data

The last dataset we looked at was the bankruptcy data found in the R package `MixGHD` (Tortora *et al.*, 2015). This dataset looked at the financial situation of 66 American firms. Each firm was labelled as either bankrupt or financially sound. We show the plots in Figure 4.10. These results show the greatest difference between the t and Gaussian mixtures when compared to the other datasets we looked at. The first item to note are the chosen weights. The weights chosen using a t mixture are very different than those chosen when using the Gaussian mixture. The second item to note is, that like the wine data, the t mixture results in better classification performance at lower percentages of labelled points. Finally, we note the difference in variability. For the

Gaussian mixture, at lower percentages, we see a lot more variability in the error bars than for the t mixture. Also, in general, there is more variability between the different weights for the Gaussian mixture. This could suggest that the selection of the weight should be treated a bit more carefully for the Gaussian mixture in this case, as the selection of a non optimal weight could result decreased classification performance. This is especially true, once again, at lower percentages.

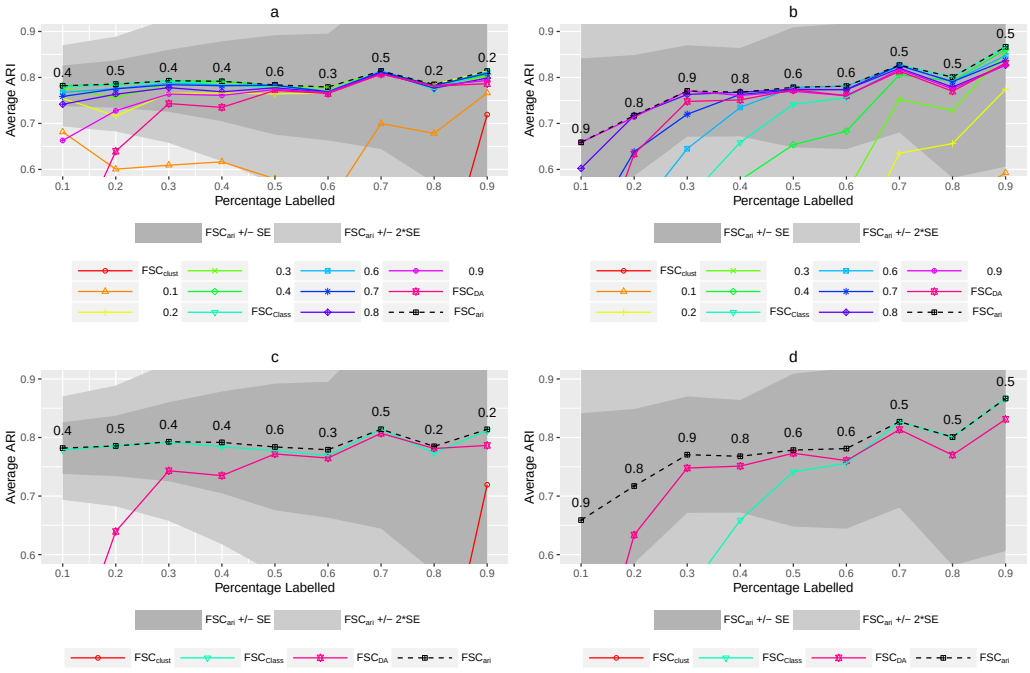

Figure 4.10: Bankruptcy Data: FSC_α for $\alpha \in \boldsymbol{\alpha}_{\text{ARI}}$ and FSC_{ARI} for a) the t mixture and b) for the Gaussian Mixture. $\text{FSC}_{\text{clust}}$,$\text{FSC}_{\text{class}}$, FSC_{DA} and FSC_{ARI} for c) the t mixture, and d) the Gaussian mixture.

4.3 Weight Selection Criteria for Parsimonious Models

We previously discussed parsimonious models, as well as five different weight selection criteria. We now compare the performance of the criteria. To do this, we looked at the Wine, Bankruptcy, Crabs and Iris datasets. We took 50 different splits for each dataset, with 80% of data labelled and used a mixture of multivariate t-distributions. We took the same candidate weights as we did before. For each candidate weight, we choose the model using the BIC, and then calculated each of weight selection criteria mentioned earlier. We then choose the optimal weight based on each of the selection criteria, and calculated the ARI. Also, we looked at the highest ARI of all the weights after choosing the model to evaluate the overall performance of each of the criteria. In Figure 4.11, we show box plots of the resulting ARI values using each of the criteria, as well as the box plot for the distribution of the highest ARI.

The distributions of the ARI values for the three classification based criteria show that the resulting ARI from the chosen weight is generally much lower than if we were to use the highest ARI. Moreover, the variability is generally much higher. Specifically, for the bankruptcy and crabs data, the chosen ARI values cover practically all possible values for the ARI.

On the other hand, the trace of the within sum of squares matrix, performs well in comparison to the three classification based criteria for the wine and bankruptcy data. Furthermore, in the case of the bankruptcy data, it performs the best of all 5 criteria, when comparing the medians, and has a distribution closest to that of the highest ARI. However, in the case of the crabs data, it performs very poorly, and

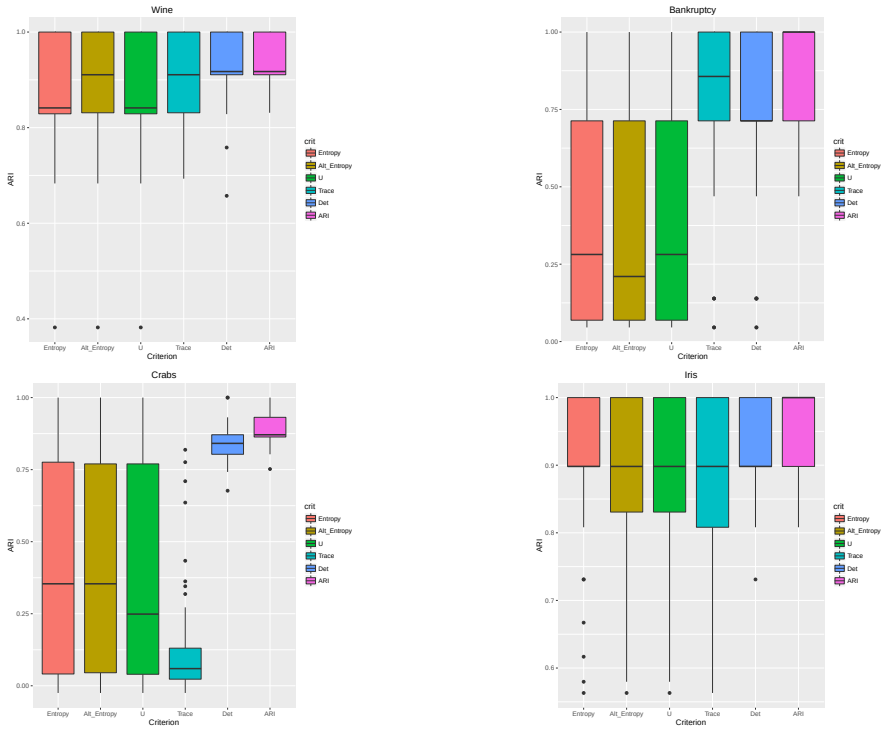

Figure 4.11: Distribution of ARI values for each of the criteria as well as the distribution of the highest ARI for the four datasets. The BIC was used to choose the model.

has the worst performance of the other criteria. For the iris data, the performance is similar to the alternate entropy and the U criterion.

Finally, we see that the determinant in general performs well for all of the datasets. In the case of the wine data, except for a couple outliers, the distribution is very similar to that for the highest ARI. Furthermore, it performs the best of all of the proposed criteria in all of the datasets except for the bankruptcy data. In this case, the trace performs better, but the distributions are still similar, and the determinant does not have the same problems as the trace does for the rest of the datasets. We therefore propose the determinant of the within sum of squares matrix to be a possible criterion to select the weight α.

The Determinant as a Model Selection Criterion

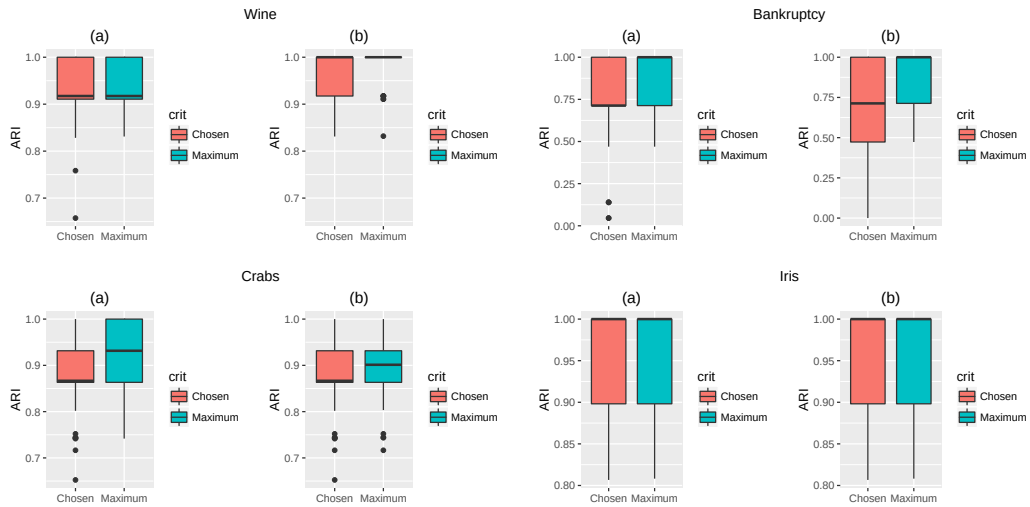

Figure 4.12: Distribution of ARI values for (a) the first procedure, and (b) the second procedure for each of the four datasets.

We have already seen that the determinant of $\mathbf{W}$ appears to be a fairly good selection criterion for the weight in FSC, and we now look at the possibility of using this criterion for selecting the model. To further explore this idea, we once again consider the four datasets, and performing 50 random splits with 80% of the data points having known labels. This time, we looked at two different procedures. In the first procedure we proceed as before and choose the model based on the BIC, and then the weight using the determinant of $\mathbf{W}$. In the second procedure we choose the model based on the determinant, and then the weight also based on the determinant. We once again took the ARI values after choosing the model and the weight using one of these two procedures, and we took the maximum ARI value amongst all of the weights.

In Figure 4.12, we show the histograms of the distribution of the ARI values. In

(a) we show the results for procedure 1, and in (b) we show the results for procedure 2.

There are a few interesting items to note. First, for the wine dataset, we see that when using the determinant to choose the model, the distribution of the maximum ARI has a lot less variability. Also, these maximum ARI values are generally larger after using the determinant to choose the model. One final note on the Wine dataset, is that the median of the ARI values using procedure 2 is higher than those from procedure 1. For the Bankruptcy data, we see that the distribution of the maximum ARI is the same regardless of using the BIC or determinant to choose the model. However, after choosing the weight, we see that the distribution of the ARI values for procedure 2 shows more variability than procedure 1. For the crabs data, we see the opposite, procedures 1 and 2 produce approximately the same distribution, however, the distribution of the maximum ARI values is different favouring using the BIC to select the model. Finally, for the Iris data, all of the distributions are very similar.

From these results, we conclude that neither procedure greatly out performs the other, and even the differences that we do see are not extremely pronounced. We therefore conclude that using the determinant as a criterion for model selection is a possibility that could be looked at in future work.

4.4 Justification for a Cluster Analysis

If some of the points are labelled, it may not be immediately clear as to why a cluster analysis should even be considered. However, there are situations in which performing a cluster analysis is just as good, if not better, than putting more weight on the labelled observations.

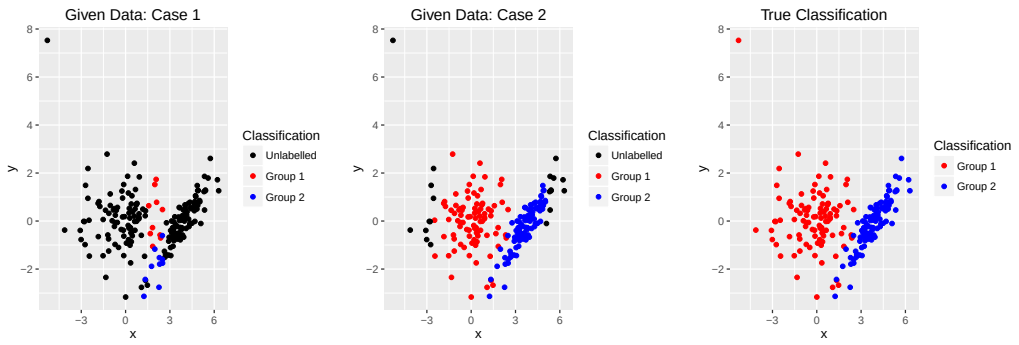

Figure 4.13: Two different possible datasets with different organizations of labelled points with the true classification.

In Figure 4.13, we show two different situations where this would be the case. In Table 4.2, we look at the ARI and the determinant of $\mathbf{W}$ for each of the weights for the two different cases. In the first case only 10% of the points are labelled, and all labelled points are around the intersection of the two clusters. In this case, we see from the ARI and determinant values that we would only want give very little weight, or no weight to the labelled observations.

Table 4.2: ARI and determinant values for each candidate weight for both of the two cases in Figure 4.13.

Weight	Case 1		Case 2	
	ARI	Det	ARI	Det
0.0	0.9341	81006	1	82849
0.1	0.9341	81006	1	82849
0.2	0.9341	81006	1	82849
0.3	0.9126	81984	1	82849
0.4	0.9126	81984	1	82849
0.5	0.9126	81984	1	82849
0.6	0.8914	84250	1	82849
0.7	0.8914	84250	1	82849
0.8	0.8914	84250	1	82849
0.9	0.0075	178858	1	82849
1.0	-0.0016	187192	1	82849

In this case, we see that a cluster analysis would actually be better than using

higher weights, and just as good as using smaller weights. In the second case, 90% of the points are labelled, and the unlabelled points lie on the outside of the two clusters. We see from the ARI and determinant values, in Table 4.2, that all weights give perfect classification, including a cluster analysis, and thus a cluster analysis would perform just as well as the other weights in this case.

Chapter 5

Conclusions and Future Work

A major contribution of this thesis is the use of the determinant of the within sum of squares matrix as a weight selection criterion in FSC. Although quite an old idea, this criteria is shown to outperform alternatives such as the near-ubiquitous BIC. Moreover, the determinant criterion, unlike the ARI used previously, can be used for a general dataset. The FSC approach is also shown to be effective for mixtures of multivariate t-distributions. For example, we saw in our simulations that in very few cases did the selected weight correspond to one of the three species of classification. Furthermore, in our real data analyses, the use of a mixture of multivariate t-distributions was shown to either perform just as well or, in the case of the Wine and Bankruptcy datasets, better than the mixture of multivariate Gaussian distributions. This is partly due to the t-distribution being more robust to outliers than the Gaussian distribution.

Future work will investigate using the determinant of the within sum of squares matrix as an alternative to the BIC for model selection in model-based clustering, classification, and discriminant analysis. Using the FSC approach in a wider range

of situations will also be explored. For example, FSC could be applied in the area of item response theory.

Appendix A

tEIGEN Models

Table A.1: Model nomenclature and number of free covariance parameters of tEIGEN models with constrained (C), unconstrained (U) and identity (I) elements.

Model	$\lambda_g = \lambda$	$\mathbf{\Lambda}_g = \mathbf{\Lambda}$	$\mathbf{D}_g = \mathbf{D}$	$\nu_g = \nu$	No. of Free Covariance Parameters
CIIC	C	I	I	C	1+1
CIIU	C	I	I	U	1+G
UIIC	U	I	I	C	(G-1)+1
UIIU	U	I	I	U	(G-1)+G
CICC	C	I	C	C	p+1
CICU	C	I	C	U	p+G
UICC	U	U	C	C	p+(G-1)+1
UICU	U	I	C	U	p+(G-1)+G
CIUC	C	I	U	C	Gp-(G-1)+1
CIUU	C	I	U	U	Gp-(G-1)+G
UIUC	U	I	U	C	Gp+1
UIUU	U	I	U	U	Gp+G
CCCC	C	C	C	C	[p(p+1)/2]+1
CCCU	C	C	C	U	[p(p+1)/2]+G
UCCC	U	C	C	C	[p(p+1)/2]+(G-1)+1
UCCU	U	C	C	U	[p(p+1)/2]+(G-1)+G
CUCC	C	U	C	C	G[p(p+1)/2]-(G-1)(p)+1
CUCU	C	U	C	U	G[p(p+1)/2]-(G-1)(p)+G
UUCC	U	U	C	C	G[p(p+1)/2]-(G-1)(p-1)+1
UUCU	U	U	C	U	G[p(p+1)/2]-(G-1)(p-1)+G
CCUC	C	C	U	C	[p(p+1)/2]+(G-1)(p-1)+1
CCUU	C	C	U	U	[p(p+1)/2]+(G-1)(p-1)+G
CUUC	C	U	U	C	G[p(p+1)/2]-(G-1)+1
CUUU	C	U	U	U	G[p(p+1)/2]-(G-1)+G
UCUC	U	C	U	C	G[p(p+1)/2]+(G-1)(p)+1
UCUU	U	C	U	U	G[p(p+1)/2]+(G-1)(p)+G
UUUC	U	U	U	C	G[p(p+1)/2]+1
UUUU	U	U	U	U	G[p(p+1)/2]+G

Appendix B

Mathematical Derivation

Recall that we derived an altered likelihood in that satisfies the property given in Dempster *et al.* (1977). Herein, we present the derivation of the altered likelihood.

$$
\begin{aligned}
\int_{\mathbb{Z}_2} \mathcal{L}_{\text{comp}}(\boldsymbol{\vartheta}|D_{\text{C}},\alpha)d\mathbf{z}_2 &= \int_{\mathbb{Z}_2} \left(\prod_{j=1}^{n_1}\prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{1j}|\boldsymbol{\theta}_g)]^{z_{jg}^{(1)}\alpha} \times \prod_{j=1}^{n_2}\prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{2j}|\boldsymbol{\theta}_g)]^{z_{jg}^{(2)}(1-\alpha)} \right) d\mathbf{z}_2 \\
&= \prod_{j=1}^{n_1}\prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{1j}|\boldsymbol{\theta}_g)]^{z_{jg}^{(1)}\alpha} \int_{\mathbb{Z}_2} \prod_{j=1}^{n_2}\prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{2j}|\boldsymbol{\theta}_g)]^{z_{jg}^{(2)}(1-\alpha)} \, d\mathbf{z}_2 \\
&= \prod_{j=1}^{n_1}\prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{1j}|\boldsymbol{\theta}_g)]^{z_{jg}^{(1)}\alpha} \prod_{j=1}^{n_2} \left(\int_{\mathbb{Z}_2} \prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{2j}|\boldsymbol{\theta}_g)]^{z_{jg}^{(2)}(1-\alpha)} \, d\mathbf{z}_2 \right) \\
&= \prod_{j=1}^{n_1}\prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{1j}|\boldsymbol{\theta})]^{z_{jg}^{(1)}\alpha} \prod_{j=1}^{n_2} \left(\sum_{\mathbf{z}_j \in \mathfrak{B}} \prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{2j}|\boldsymbol{\theta})]^{z_{jg}^{(2)}(1-\alpha)} \right) \\
&= \prod_{j=1}^{n_1}\prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{1j}|\boldsymbol{\theta})]^{z_{jg}^{(1)}\alpha} \prod_{j=1}^{n_2} \left(\sum_{g=1}^{G} [\pi_g f_g(\mathbf{x}_{2j}|\boldsymbol{\theta})]^{(1-\alpha)} \right),
\end{aligned}
$$

where

$$\mathfrak{B} = \left\{ \mathbf{z}_j = \left(z_{j1}^{(2)}, z_{j2}^{(2)}, \ldots, z_{jG}^{(2)} \right) \;\middle|\; z_{jg}^{(2)} \in \{0,1\}, \forall g \in \{1,2,\ldots,G\}, \sum_{g=1}^{G} z_{jg}^{(2)} = 1 \right\}.$$

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Andrews, J. L. and McNicholas, P. D. (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions: The tEIGEN family. *Statistics and Computing*, **22**(5), 1021–1029.

Andrews, J. L. and McNicholas, P. D. (2014). Variable selection for clustering and classification. *Journal of Classification*, **31**(2), 136–153.

Andrews, J. L., Wickins, J. R., Boers, N. M., and McNicholas, P. D. (2015). *teigen: Model-Based Clustering and Classification with the Multivariate t Distribution*. R package version 2.1.0.

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

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, **41**, 164–171.

Bensmail, H., Celeux, G., Raftery, A., and Robert, C. (1997). Inference in model-based cluster analysis. *Statistics and Computing*, **7**, 1–10.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(7), 719–725.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, **28**(5), 781–793.

Celeux, G. and Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, **13**, 195–212.

Dang, U. J., Browne, R. P., and McNicholas, P. D. (2015). Mixtures of multivariate power exponential distributions. *Biometrics*, **71**(4), 1081–1089.

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.

Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering methods? Answers via model-based cluster analysis. *The Computer Journal*, **41**(8), 578–588.

Fraley, C. and Raftery, A. E. (2002a). MCLUST: Software for model-based clustering, density estimation, and discriminant analysis. Technical Report 415, University of Washington, Department of Statistics.

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

Fraley, C., Raftery, A. E., Murphy, T. B., and Scrucca, L. (2012). mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report 597, Department of Statistics, University of Washington, Seattle, WA.

Franczak, B. C., Browne, R. P., and McNicholas, P. D. (2014). Mixtures of shifted asymmetric Laplace distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **36**(6), 1149–1157.

Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, **62**, 1159–1178.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, **2**(1), 193–218.

Hurley, C. (2004). Clustering visualizations of multivariate data. *Journal of Computational and Graphical Statistics*, **13**(4), 788–806.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297, Berkeley. University of California Press.

McNicholas, P. D. (2016a). *Mixture Model-Based Classification*. Chapman & Hall/CRC Press, Boca Raton.

McNicholas, P. D. (2016b). Model-based clustering. *Journal of Classification*, **33**.

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

McNicholas, P. D., Murphy, T. B., McDaid, A. F., and Frost, D. (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis*, **54**(3), 711–723.

Orchard, T. and Woodbury, M. A. (1972). A missing information principle: Theory and applications. In L. M. Le Cam, J. Neyman, and E. L. Scott, editors, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Theory of Statistics*, pages 697–715. University of California Press, Berkeley.

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

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, **27**, 387–397.

Tortora, C., Browne, R. P., Franczak, B. C., and McNicholas, P. D. (2015). *MixGHD: Model Based Clustering, Classification and Discriminant Analysis Using the Mixture of Generalized Hyperbolic Distributions*. R package version 1.8.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

Vrbik, I. and McNicholas, P. D. (2012). Analytic calculations for the EM algorithm for multivariate skew-t mixture models. *Statistics and Probability Letters*, **82**(6), 1169–1174.

Vrbik, I. and McNicholas, P. D. (2014). Parsimonious skew mixture models for model-based clustering and classification. *Computational Statistics and Data Analysis*, **71**, 196–210.

Vrbik, I. and McNicholas, P. D. (2015). Fractionally-supervised classification. *Journal of Classification*, **32**(3), 359–381.

Wolfe, J. H. (1965). A computer program for the maximum likelihood analysis of types. Technical Bulletin 65-15, U.S. Naval Personnel Research Activity.