MODEL-BASED CLASSIFICATION VIA MIXTURES OF
MULTIVARIATE t -FACTOR ANALYZERS

A Thesis

Presented to

The Faculty of Graduate Studies

of

The University of Guelph

by

MICHELLE A. STEANE

In partial fulfilment of requirements

for the degree of

Master of Science

November, 2009

© Michelle A. Steane, 2009

Library and Archives
Canada

Published Heritage
Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et
Archives Canada

Direction du
Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-56783-8
Our file Notre référence
ISBN: 978-0-494-56783-8

NOTICE:

The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protègent cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

ABSTRACT

MODEL-BASED CLASSIFICATION VIA MIXTURES OF MULTIVARIATE t -FACTOR ANALYZERS

Michelle A. Steane
University of Guelph, 2009

Advisors:
Dr. Paul D. McNicholas

A new model-based classification technique is developed, based on the mixtures of multivariate t -factor analyzers model and a variation of the EM algorithm is used to estimate model parameters. A convergence criteria based on Aitken's acceleration is used and model selection is carried out using the Bayesian information criteria. This new classification technique is applied to physical and chemical measurements of red wine samples from Italy and to fatty acid measurements on olive oil samples from Italy. These results are discussed and compared to another well-established classification technique. The new method outperformed the traditional technique in virtually all analyses performed.

For my grandmothers.

Acknowledgments

First, I would like to thank Dr. Paul McNicholas for giving me the opportunity to complete a master's degree, and for spending so much time providing advice, guidance and motivation whenever I needed it. I would also like to thank Dr. Rickey Yada for being on my advisory committee and providing valuable insight and viewpoints from someone outside the field of statistics. Thank you Dr. Ayesha Ali and Dr. O. Brian Allen for taking the time to be on my examining committee.

I would also like to thank my friends and family who encouraged and supported me throughout this last year. Thank you for listening to my frustrations and complaints, providing opportunities for well-needed relaxation, reminding me that this would be well worth it in the end and for motivating me to strive for the best.

Table of Contents

List of Tables	v
List of Figures	viii
1 Introduction and Background	1
1.1 Mixture Models	1
1.1.1 Model-Based Clustering and Finite Mixture Models	1
1.1.2 Clustering, Classification and Food Authenticity	2
1.1.3 Gaussian Mixture Models	2
1.1.4 Mixtures of Multivariate t -Distributions	3
1.2 The Expectation-Maximisation Algorithm	4
1.2.1 The EM Algorithm	4
1.2.2 Examples of EM Algorithm Application	5
1.2.3 The Expectation-Conditional Maximization Algorithm	6
1.2.4 The Alternating Expectation-Conditional Maximization Algorithm	7
1.3 MCLUST: Software for Model-Based Cluster Analysis	8
1.4 Variable Selection	10
1.5 Parsimonious Gaussian Mixture Models	11
1.5.1 Latent Gaussian Models	11
1.5.2 Mixtures of Factor Analyzers	12
1.5.3 Parsimonious Gaussian Mixture Models	13
1.5.4 Applications of Parsimonious Gaussian Mixture Models	13
1.6 Mixture Model Selection and Comparison	14
1.6.1 The Bayesian Information Criterion	14
1.6.2 The Rand and Adjusted Rand Indices	15
2 The AECM Algorithm for a Mixture of Multivariate t-Distributions	17
2.1 Introduction	17
2.2 The Mixtures of Multivariate t -Factor Analyzers Model	17
2.3 The Likelihood Function	19
2.4 E-Step	20
2.5 CM-Steps	21
3 Model-Based Classification	25
3.1 Discriminant Analysis	25

3.2	Model-Based Classification <i>via</i> a Mixture of Multivariate t -Factor Analyzers	27
3.3	AECM Algorithm for Model-Based Classification <i>via</i> a Mixture of Multivariate t -Factor Analyzers	29
3.4	Model Selection for Model-Based Classification <i>via</i> a Mixture of Multivariate t -Factor Analyzers	31
4	Data Analysis: Italian Wines	34
4.1	The Data	34
4.2	Proportion of Data Known: 60%	35
4.2.1	Classification <i>via</i> MMtFAs	35
4.2.2	Classification <i>via</i> LDA	37
4.3	Proportion of Data Known: 70%	37
4.3.1	Classification <i>via</i> MMtFAs	37
4.3.2	Classification <i>via</i> LDA	38
4.4	Proportion of Data Known: 80%	39
4.4.1	Classification <i>via</i> MMtFAs	39
4.4.2	Classification <i>via</i> LDA	39
4.5	Proportion of Data Known: 90%	41
4.5.1	Classification <i>via</i> MMtFAs	41
4.5.2	Classification <i>via</i> LDA	42
4.6	Method Comparisons	42
5	Data Analysis: Italian Olive Oils	45
5.1	The Data	45
5.2	Proportion of Data Known: 60%	46
5.2.1	Classification <i>via</i> MMtFAs	46
5.2.2	Classification <i>via</i> LDA	49
5.3	Proportion of Data Known: 70%	50
5.3.1	Classification <i>via</i> MMtFAs	50
5.3.2	Classification <i>via</i> LDA	52
5.4	Proportion of Data Known: 80%	53
5.4.1	Classification <i>via</i> a Mixture of Multivariate t -Factor Analyzers	53
5.4.2	Classification <i>via</i> LDA	55
5.5	Proportion of Data Known: 90%	56
5.5.1	Classification <i>via</i> LDA	58
5.6	Method Comparisons	59
6	Conclusions	63
6.1	Discussion	63
6.2	Future Work	64
	Bibliography	66

List of Tables

1.1	Covariance structures, Σ_g , available using <code>mclust</code> , along with the corresponding number of covariance parameters.	9
1.2	Parsimonious covariance structures derived from the mixtures of factor analyzers model along with the number of covariance parameters. . .	13
4.1	Thirteen of the chemical and physical properties of wine reported by Forina et al. (1986).	34
4.2	True mixing proportions of wine reported by Forina et al. (1986). . .	35
4.3	Estimated classifications for UUU, $G = 3$ and $q = 2$, for the wine data with 60% known.	36
4.4	Best model estimated degrees of freedom, for the wine data with 60% known.	36
4.5	Best model estimated mixing proportions, for the wine data with 60% known.	36
4.6	LDA predicted classifications, for the wine data with 60% known. . .	37
4.7	Estimated classifications for UUU, $G = 3$ and $q = 2$, for the wine data with 70% known.	38
4.8	Best model estimated degrees of freedom, for the wine data with 70% known.	38
4.9	Best model estimated mixing proportions, for the wine data with 70% known.	38
4.10	LDA predicted classifications, for the wine data with 70% known. . .	39
4.11	Estimated classifications for UUU, $G = 3$ and $q = 2$, for the wine data with 80% known.	40
4.12	Best model estimated degrees of freedom, for the wine data with 80% known.	40
4.13	Best model estimated mixing proportions, for the wine data with 80% known.	40
4.14	LDA predicted classifications, for the wine data with 80% known. . .	40
4.15	Estimated classifications for UUU, $G = 3$ and $q = 2$, for the wine data with 90% known.	41
4.16	Best model estimated degrees of freedom, for the wine data with 90% known.	41
4.17	Best model estimated mixing proportions, for the wine data with 90% known.	41
4.18	Wine 90%: LDA predicted classifications.	42

4.19	Mixtures of multivariate t -factor analyzers summary for the wine data.	43
4.20	Linear discriminant analysis summary for the wine data.	43
4.21	Estimated degrees of freedom summary for the wine data.	43
4.22	Mixing proportions summary for the wine data.	44
5.1	Eight fatty acids of olive oil reported by Forina et al., 1983.	46
5.2	True mixing proportions for the olive oil data's three group structure.	46
5.3	True mixing proportions for the olive oil data's nine group structure.	46
5.4	Estimated classifications for $G = 3$, $q = 5$ and UUU, for the olive oil data with 60% known.	48
5.5	Best model estimated degrees of freedom for $G = 3$, for the olive oil data with 60% known.	48
5.6	Best model estimated mixing proportions for $G = 3$, for the olive oil data with 60% known.	48
5.7	Estimated classifications for $G = 9$, $q = 4$ and UUU, for the olive oil data with 60% known.	48
5.8	Best model estimated degrees of freedom for $G = 9$, for the olive oil data with 60% known.	49
5.9	Best model estimated mixing proportions for $G = 9$, for the olive oil data with 60% known.	49
5.10	Olive Oil 60%: LDA Estimated classifications for $G = 3$	49
5.11	Olive Oil 60%: LDA Estimated classifications for $G = 9$	50
5.12	Estimated classifications for $G = 3$, $q = 5$ and UUU, for the olive oil data with 70% known.	51
5.13	Best model estimated degrees of freedom for $G = 3$, for the olive oil data 70% known.	51
5.14	Best model estimated mixing proportions for $G = 3$, for the olive oil data 70% known.	51
5.15	Estimated classifications for $G = 9$, $q = 4$ and UUU, for the olive oil data with 70% known.	51
5.16	Best model estimated degrees of freedom for $G = 9$, for the olive oil data with 70% known.	52
5.17	Best model estimated mixing proportions for $G = 9$, for the olive oil data with 70% known.	52
5.18	Olive Oil 70%: LDA Estimated classifications for $G = 3$	52
5.19	Olive Oil 70%: LDA Estimated classifications for $G = 9$	53
5.20	Estimated classifications for $G = 3$, $q = 5$ and UUU, for the olive oil data with 80% known.	54
5.21	Best model estimated degrees of freedom for $G = 3$, for the olive oil data with 80% known.	54
5.22	Best model estimated mixing proportions for $G = 3$, for the olive oil data with 80% known.	54
5.23	Estimated classifications for $G = 9$, $q = 5$ and UUU, for the olive oil data with 80% known.	54

5.24	Best model estimated degrees of freedom for $G = 9$, for the olive oil data with 80% known.	55
5.25	Best model estimated mixing proportions for $G = 9$, for the olive oil data with 80% known.	55
5.26	LDA Estimated classifications for the olive oil data with 80% known.	55
5.27	LDA Estimated classifications for $G = 9$, for the olive oil data with 80% known.	56
5.28	Estimated classifications for $G = 3$, $q = 5$ and UUU, for the olive oil data with 90% known.	57
5.29	Best model estimated degrees of freedom, for the olive oil data with 90% known.	57
5.30	Best model estimated mixing proportions, for the olive oil data with 90% known.	57
5.31	Estimated classifications for $G = 9$, $q = 4$ and UUU, for the olive oil data with 90% known.	57
5.32	Best model estimated degrees of freedom for $G = 9$, for the olive oil data with 90% known.	58
5.33	Best model estimated mixing proportions for $G = 9$, for the olive oil data with 90% known.	58
5.34	LDA Estimated classifications for the olive oil data with 90% known.	58
5.35	LDA Estimated classifications for $G = 9$, for the olive oil data with 90% known.	59
5.36	Olive oil data MMtFAs summary for $G = 3$	60
5.37	Olive oil data LDA summary for $G = 3$	60
5.38	Estimated degrees of freedom summary for the olive oil data for $G = 3$	60
5.39	Mixing proportions summary for the olive oil data for $G = 3$	60
5.40	Olive oil data MMFtAs summary for $G = 9$	61
5.41	Olive oil data LDA summary for $G = 9$	61
5.42	Estimated degrees of freedom summary for the olive oil data for $G = 9$	61
5.43	Mixing proportions summary for the olive oil data for $G = 9$	62
6.1	Possible covariance structures derived from the mixtures of factor analyzers model along with the number of covariance parameters.	65

List of Figures

1.1	Cluster shapes that correspond to the covariance structures given in Table 1.1.	9
4.1	Pairs plot of physical and chemical measurements from the Italian wine data.	35
5.1	Pairs plot of fatty acid measurements from the Italian olive oil data. .	47

Chapter 1

Introduction and Background

1.1 Mixture Models

1.1.1 Model-Based Clustering and Finite Mixture Models

Model-based clustering techniques first appeared at least as far back as the year 1963 in a master's thesis (Wolfe, 1963). In recent years, model-based clustering has appeared in the statistics literature with increased frequency (Fraley and Raftery, 1998; McLachlan et al., 2003; Raftery and Dean, 2006; McLachlan et al., 2007; McNicholas and Murphy, 2008). Typically, some pre-assumed mixture modelling structure is used to cluster the data and some variant of the expectation-maximisation algorithm (Dempster et al., 1977) is used to estimate the parameters associated with these models.

Finite mixture models assume that data are collected from a population with a finite number of sub-populations and that some standard statistical model can be used to model the data within each sub-population (McNicholas and Murphy, 2008). Extensive reviews of finite mixture models can be found in McLachlan and Peel (2000) and Frühwirth-Schnatter (2006).

1.1.2 Clustering, Classification and Food Authenticity

An authentic food is one that is precisely what it claims to be. One of the principal statistical problems in food authenticity is the development of accurate classifiers that can be used to 'automatically' tell us if a food is authentic or not. Consider three types of Italian red wines that fall into very different price ranges, and therefore it would be detrimental to mislabel one type as another. If we had no previous information on these wines, that is we have no samples with known origins, a clustering approach could be taken to estimate the type of wine present. In a real world setting it is likely that some information about the different types of wines would be available, and samples with known types could be incorporated into the analysis. In this situation a classification approach would be taken. For either case, it may be desirable to compare the estimated classifications of the wines to the supposed type as a test of authenticity.

1.1.3 Gaussian Mixture Models

The Gaussian mixture model is particularly popular within the literature. Each sub-population is assumed to have a Gaussian distribution and the model density is of the form

$$f(\mathbf{x} \mid \vartheta) = \sum_{g=1}^G \pi_g \phi(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \quad (1.1)$$

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 a multivariate Gaussian probability density with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$, ϑ is the full parameter space, p is the number of variables and π_g is the probability of belonging to sub-population g so that $\pi_g \in [0, 1] \forall g$ and $\sum_{g=1}^G \pi_g = 1$.

1.1.1.4 Mixtures of Multivariate t -Distributions

Mixtures of multivariate t -distributions have also been explored as an approach to model-based clustering (McLachlan and Peel, 1998; McLachlan et al., 2007). Each sub-population is assumed to have a multivariate t -distribution and the model density is given by

$$f(\mathbf{x}) = \sum_{g=1}^G \pi_g f(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g), \quad (1.2)$$

where π_g is the probability of sub-population membership, and $f(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g)$ is the density of a multivariate t -distribution with mean $\boldsymbol{\mu}_g$, covariance matrix $\boldsymbol{\Sigma}_g$, and degrees of freedom ν_g . That is,

$$f(\mathbf{x} \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g) = \frac{\Gamma\left(\frac{\nu_g + p}{2}\right) |\boldsymbol{\Sigma}_g|^{-1/2}}{(\pi \nu_g)^{\frac{1}{2}p} \Gamma\left(\frac{\nu_g}{2}\right) \left\{1 + \frac{1}{\nu_g} (\mathbf{x} - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x} - \boldsymbol{\mu}_g)\right\}^{\frac{1}{2}(\nu_g + p)}},$$

where p is the number of variables and $\Gamma(\cdot)$ is the gamma function defined by

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt.$$

Multivariate t -distributions provide a more robust approach to model-based cluster analysis as they have thicker tails than their Gaussian counterparts. When performing Gaussian cluster analysis, outliers in the data can impact the parameter estimates. However, when analyzing with the t -distribution, less weight is placed on observations with large Mahalanobis distances and therefore these observations have

a smaller effect on density parameter estimates. Note that the Mahalanobis distance is defined as $\delta(\mathbf{x}, \boldsymbol{\mu}_g \mid \boldsymbol{\Sigma}_g) = (\mathbf{x} - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x} - \boldsymbol{\mu}_g)$.

A thorough review of multivariate t -distributions, their application to cluster analysis, and a demonstration of how applying t -components as opposed to Gaussian components results in less extreme estimates of posterior probabilities of group membership can be found in McLachlan and Peel (1998) and McLachlan et al. (2007).

1.2 The Expectation-Maximisation Algorithm

1.2.1 The EM Algorithm

The EM algorithm (Dempster et al., 1977) is an iterative method for finding maximum likelihood estimates (MLEs) where the data is incomplete, missing or assumed to be incomplete. Though the data does not necessarily need to be incomplete, treating it as such can lead to efficient solutions through the use of the EM algorithm.

In the E-step, current estimates of the model parameters and the ‘complete-data’ vector, the vector of observed data plus missing data, are used to calculate the expected value of the log-likelihood. In the M-step, updated model parameters are estimated by maximizing the expected value found in the E-step with respect to the model parameters we wish to estimate.

These E-steps and M-steps are repeated consecutively, iteratively until convergence is reached. Convergence can be quantified in a variety of ways. One popular approach is the lack of progress approach in which convergence is said to have been reached when the difference in successive estimates of the log-likelihood is sufficiently

small. An in depth overview of EM algorithms can be found in McLachlan and Krishnan (2008) and some other convergence criteria are discussed in Lindsay (1995), McLachlan and Krishnan (2008) and McNicholas et al. (2009).

1.2.2 Examples of EM Algorithm Application

An application of the EM algorithm to a mixture model classification problem can be found in Dean et al. (2006). This study compares EM, classification EM (CEM) algorithms, and model-based discriminant analysis in their abilities to classify various meat samples and Greek olive-oils correctly using their visible and near infra-red spectra. The study methodology was based on the assumption that the data arose from a Gaussian mixture model with parsimonious covariance structure within the MCLUST framework (see Section 1.3).

The EM and CEM algorithms differ only in the manner in which they estimate group membership. The EM algorithm allows the expected values of the unknown group labels, $\hat{z}_{ig}$, to take on any value between zero and one. However, the CEM algorithm makes a discrete classification at each iteration, where each object is allocated to a unique group based on the rule

$$\hat{z}_{ig}^{(k+1)} = \begin{cases} 1 & \text{if } w_{ig}^{(k+1)} > w_{ig'}^{(k+1)} \text{ for all } g' \neq g, \\ 0 & \text{otherwise,} \end{cases},$$

where

$$w_{ig}^{(k+1)} = \frac{\hat{\pi}_g^{(k)} f(y_i | \hat{\theta}_g^{(k)})}{\sum_{g'=1}^G \hat{\pi}_{g'}^{(k)} f(y_i | \hat{\theta}_{g'}^{(k)})}$$

for $g = 1, 2, \dots, G$ and $i = 1, 2, \dots, n$. For example, if the EM algorithm produces ex-

pected values of the unknown labels of $\hat{z}_i^{(k+1)} = (0.14, 0.60, 0.26)$, then the equivalent step in the CEM algorithm would transform them into $\hat{z}_i^{(k+1)} = (0, 1, 0)$.

Dean et al. (2006) analyzed two cases for each food type; one where the number of known sample labels was equal to that of unknown sample labels and another where the number of known sample labels was very small in comparison to that of unknown sample labels. When analyzing the meat sample data by type, EM and CEM procedures have only slightly higher classification rates than discriminant analysis when the number of known and unknown labels were equal. However, when the number of known labels was very small, then the EM and CEM procedures had a much higher correct classification rate than the model-based discriminant analysis. Also, the EM algorithm had slightly better classification rates than the CEM algorithm in this case. Analysis of Greek olive-oils by location resulted in similar conclusions. When the number of known sample labels is small, EM and CEM procedures have a distinct advantage over ordinary discriminant analysis in terms of correct classification rates.

1.2.3 The Expectation-Conditional Maximization Algorithm

Meng and Rubin (1993) extended the expectation maximization (EM) algorithm by replacing the M-step with several computationally simpler conditional maximization (CM) steps. Each CM-step maximizes the function found in the E-step subject to constraints on φ , the full parameter space, which is partitioned into sub-spaces.

In general, the full parameter space is partitioned into S sub-spaces, and the

sth CM-step of the $(k + 1)^{\text{th}}$ iteration maximizes the function found in the E-step with respect to φ_s , the sth sub-space, while all other $(S - 1)$ sub-spaces are held constant at their current estimates.

When considering mixtures of t -distributions (McLachlan and Peel, 2000) φ is partitioned into $(\varphi_1^T, \varphi_2^T)^T$ where $\varphi_1 = (\pi_1, \dots, \pi_{G-1}, \xi^T)^T$ and $\varphi_2 = (\nu_1, \dots, \nu_G)^T$ for $g = 1, \dots, G$. The matrix ξ contains the elements of μ_g and components of the covariance matrix Σ_g . The E-step is the same as the EM algorithm, however the M-step is replaced with the following two CM-steps.

CM-Step 1: $\varphi_1^{(k+1)}$ is calculated by maximizing the function found in the E-step with φ_2 held constant at $\varphi_2^{(k)}$.

CM-Step 2: $\varphi_2^{(k+1)}$ is calculated by maximizing the function found in the E-step with φ_1 held constant at $\varphi_1^{(k)}$.

1.2.4 The Alternating Expectation-Conditional Maximization Algorithm

The alternating expectation-conditional maximization (AECM) algorithm (Meng and van Dyk, 1997) is an extension of the ECM algorithm that specifies different missing data at each stage. During the first stage of the algorithm, which consists of the first E-step and the first CM-step, we estimate μ_g and π_g . For this stage, we let the unobserved group labels be $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n)$. During the second stage of the algorithm, which consists of the second E-step and the second CM-step, we estimate the components of the covariance structure, Λ_g the factor loading matrix and Ψ_g the error variance matrix, and we let $\mathbf{z}$ and $\mathbf{u}$ be the missing data. The AECM algorithm updates the parameter estimates iteratively until they converge to the maximum like-

likelihood estimates. At convergence, the $\hat{z}_{ig}$ estimate the posterior probability of group membership for each observation. These probability estimates can then be used to cluster observations into sub-populations.

1.3 MCLUST: Software for Model-Based Cluster Analysis

MCLUST is probably the most popular model-based clustering technique in the literature. The eigenvalue decomposition of the covariance matrix (Banfield and Raftery, 1993; Celeux and Govaert, 1995; Fraley and Raftery, 1998, 2002b) is of the form

$$\Sigma_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g', \quad (1.3)$$

where λ_g is a constant, $\mathbf{D}_g$ is a matrix containing the eigenvectors from the decomposition of Σ_g , and $\mathbf{A}_g$ is a diagonal matrix with entries proportional to the eigenvalues from the decomposition of Σ_g .

A variety of constraints can be applied to the model-based clustering techniques that have been developed using the covariance structure given in Equation 1.3. Each of the covariance structure components may be constrained to be equal across groups. The matrix $\mathbf{D}_g$ or the matrices $\mathbf{D}_g$ and $\mathbf{A}_g$ may also be set equal to $\mathbf{I}$. This leads to a family of ten models.

The `mclust` software, which is available as a package in the software R (R Development Core Team, 2009), is described by Fraley and Raftery (2002a, 2003). The work of Fraley and Raftery (1998, 1999, 2002b) is incorporated into the `mclust` package and allows for efficient model-based clustering. Table 1.1, which is taken

from Dean et al. (2006), demonstrates the various covariance decompositions that are available using the `mclust` software. Visual representations of these covariance decompositions is available in Figure 1.1, which is similar to a figure in Dean et al. (2006).

Table 1.1: Covariance structures, Σ_g , available using `mclust`, along with the corresponding number of covariance parameters.

ID	Volume	Shape	Orientation	Covariance Decomposition
EII	Equal	Spherical	—	$\lambda \mathbf{I}$
VII	Variable	Spherical	—	$\lambda_g \mathbf{I}$
EEI	Equal	Equal	Axis-aligned	$\lambda \mathbf{A}$
VEI	Variable	Equal	Axis-aligned	$\lambda_g \mathbf{A}$
EVI	Equal	Variable	Axis-aligned	$\lambda \mathbf{A}_g$
VVI	Variable	Variable	Axis-aligned	$\lambda_g \mathbf{A}_g$
EEE	Equal	Equal	Equal	$\lambda \mathbf{D} \mathbf{A} \mathbf{D}'$
EEV	Equal	Equal	Variable	$\lambda \mathbf{D}_g \mathbf{A} \mathbf{D}_g'$
VEV	Variable	Equal	Variable	$\lambda_g \mathbf{D}_g \mathbf{A} \mathbf{D}_g'$
VVV	Variable	Variable	Variable	$\lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g'$

Figure 1.1: Cluster shapes that correspond to the covariance structures given in Table 1.1.

It has been demonstrated by Fraley and Raftery (2002b) that by using these parsimonious mixture models in a model-based clustering context one can achieve

excellent results in a variety of applications. Dean et al. (2006) demonstrate the application of these models to provide classification rules in food authenticity studies as described earlier.

1.4 Variable Selection

A variable selection method based on the use of Bayes factors is proposed by Raftery and Dean (2006). Their approach is that which would be taken to a model selection problem. Consider using Bayes factors to compare two models, M_1 and M_2 , for data X . The Bayes factor, B_{12} , for model M_1 versus model M_2 , is defined as

$$B_{12} = \frac{p(X | M_1)}{p(X | M_2)}, \quad (1.4)$$

where

$$p(X | M_k) = \int p(X | \boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k | M_k) d\boldsymbol{\theta}_k,$$

for $k = 1, 2$ and $\boldsymbol{\theta}_k$ is the parameter vector for model M_k and $p(\boldsymbol{\theta}_k | M_k)$ is the prior distribution of M_k (Kass and Raftery, 1995). Variable selection is then based on which model is the ‘best’. The `clustvarsel` package (Dean and Raftery, 2006) in R can be used to carry out variable selection. The user only needs to pre-specify the maximum number of sub-populations and `clustvarsel` automatically selects the variables. Unfortunately, variable selection does not always outperform `mclust`. A practical example of this phenomenon is given by McNicholas and Murphy (2008).

Approximated Bayes factors were introduced by Kass and Raftery (1995). This approximation allows for a simpler and easier method for calculating the Bayes

factor for a model M_1 against another model M_2 , and is given by

$$2\log(B_{12}) \approx \Delta_{\text{BIC}},$$

where B_{12} is the Bayes factor for M_1 versus M_2 and Δ_{BIC} is the difference in BIC for M_1 and M_2 . This method assumes that the prior model probabilities for M_1 and M_2 are equal.

1.5 Parsimonious Gaussian Mixture Models

1.5.1 Latent Gaussian Models

The latent Gaussian model underlies both factor analysis and probabilistic principal components analysis. The term ‘latent’ in a statistical context literally means underlying, or not visible.

Factor analysis (Spearman, 1904) is one data reduction technique that supposes that variability in data can be attributed to unobserved factors. The model is $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\mathbf{U} + \mathbf{e}$, where $\mathbf{X}$ is a p -dimensional random vector that is modeled by $\mathbf{U}$, a q -dimensional vector of unobserved latent factors, and $q \ll p$. Also, $\boldsymbol{\Lambda}$ is a $p \times q$ matrix of factor loadings, $\mathbf{e} \sim N(\mathbf{0}, \boldsymbol{\Psi})$ where $\boldsymbol{\Psi} = \text{diag}(\psi_1, \psi_2, \dots, \psi_p)$, and $\mathbf{U} \sim N(\mathbf{0}, \mathbf{I}_q)$. The random vector $\mathbf{X}$ is distributed marginally Gaussian with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$.

Probabilistic principal component analysis (Tipping and Bishop, 1999) is simply a special case of factor analysis that assumes the distribution of errors are isotropic such that $\boldsymbol{\Psi} = \psi\mathbf{I}_p$. The previously mentioned EM algorithm can be used to find max-

imum likelihood estimates (MLEs) for $\mathbf{\Lambda}$ and $\mathbf{\Psi}$, where the complete-data is $(\mathbf{x}, \mathbf{u})$.

The MLE of $\boldsymbol{\mu}$ is easily found to be $\bar{\mathbf{x}}$.

1.5.2 Mixtures of Factor Analyzers

Single-factor analysis models represent the reduced data using a global linear model. Due to this linearity, they are limited in their ability to reveal group structure in a data set. This limitation leads to a non-linear approach in which a finite mixture of linear sub-models is proposed as the distribution of the full observation vector $\mathbf{x}$ (Ghahramani and Hinton, 1997; McLachlan and Peel, 2000). This can be modelled as

$$\mathbf{x}_i = \boldsymbol{\mu}_g + \mathbf{\Lambda}_g \mathbf{u}_{ig} + \mathbf{e}_{ig} \quad (1.5)$$

with probability π_g for $g = 1, \dots, G$ and $i = 1, \dots, n$, where $\mathbf{u}_{ig}$ is a q -dimensional vector of unobserved latent factors ($q \ll p$), and $\mathbf{\Lambda}_g$ is a $p \times q$ matrix of factor loadings.

The distributions of $\mathbf{u}_{ig}$ and $\mathbf{e}_{ig}$ are the same as in the single-factor case.

Therefore, the mixture of factor analyzers model is given by

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

where the covariance matrix now has the form $\boldsymbol{\Sigma}_g = \mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \mathbf{\Psi}_g$. The mixture of factor analyzers model has density

$$f(\mathbf{x}_i) = \sum_{g=1}^G \frac{\pi_g}{[(2\pi)^p |\mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \mathbf{\Psi}_g|]^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_g)' (\mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \mathbf{\Psi}_g)^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g) \right\},$$

where π_g is the probability of membership in group g , also known as the mixing proportions.

1.5.3 Parsimonious Gaussian Mixture Models

The mixture of factor analyzers model can be extended by allowing constraints across groups on the Λ_g and Ψ_g matrices, and on whether or not $\Psi_g = \psi_g \mathbf{I}_p$. By including constraints on the factor loadings matrices Λ_g , the number of covariance parameters is notably reduced and parsimonious models are obtained (McNicholas and Murphy, 2008). All possible combinations of these constraints result in eight different parsimonious Gaussian mixture models (PGMMs), which are given in Table 1.2.

Table 1.2: Parsimonious covariance structures derived from the mixtures of factor analyzers model along with the number of covariance parameters.

ModelID	Loading Matrix	Error Variance	Isotropic	Covariance Structure
CCC	Constrained	Constrained	Constrained	$\Sigma_g = \Lambda\Lambda' + \psi\mathbf{I}_p$
CCU	Constrained	Constrained	Unconstrained	$\Sigma_g = \Lambda\Lambda' + \Psi$
CUC	Constrained	Unconstrained	Constrained	$\Sigma_g = \Lambda\Lambda' + \psi_g\mathbf{I}_p$
CUU	Constrained	Unconstrained	Unconstrained	$\Sigma_g = \Lambda\Lambda' + \Psi_g$
UCC	Unconstrained	Constrained	Constrained	$\Sigma_g = \Lambda_g\Lambda_g' + \psi\mathbf{I}_p$
UCU	Unconstrained	Constrained	Unconstrained	$\Sigma_g = \Lambda_g\Lambda_g' + \Psi$
UUC	Unconstrained	Unconstrained	Constrained	$\Sigma_g = \Lambda_g\Lambda_g' + \psi_g\mathbf{I}_p$
UUU	Unconstrained	Unconstrained	Unconstrained	$\Sigma_g = \Lambda_g\Lambda_g' + \Psi_g$

1.5.4 Applications of Parsimonious Gaussian Mixture Models

Examples of the use of PGMMs include analysis of physical and chemical properties of red wines from Italy, chemical properties of coffee (McNicholas and Murphy, 2008), and biological measurements of crabs (McNicholas et al., 2009).

A data set of twenty-eight physical and chemical properties of three types of red wine (Barolo, Grignolino, and Barbera) from the Piedmont region of Italy was collected (Forina et al., 1986). McNicholas and Murphy (2008) performed PGMMs, MCLUST and variable selection analyses on twenty-seven of the twenty-eight collected

variables, as well as a subset of thirteen of the variables. When twenty-seven variables were considered, all three methods of analysis discovered the correct number of groups, however, the PGMM method classified more wines into the correct sub-populations, and therefore is more accurate than the MCLUST or variable selection methods. When the subset of thirteen variables was considered, all three methods were not able to produce good results.

A data set on chemical properties of coffee samples collected world wide (Streuli, 1973) was examined using PGMM, MCLUST, and variable selection analyses on 12 of the chemical variables. The actual group structure contained only two clusters and the PGMM analysis was the only one to correctly identify this structure. MCLUST found three groups and the variable selection method found eight. As with the wine data, the PGMM method provided the most accurate analysis and sub-population structure. Note that variable selection failed to outperform MCLUST on these data.

1.6 Mixture Model Selection and Comparison

1.6.1 The Bayesian Information Criterion

One method often used to select an appropriate mixture model is the Bayesian information criterion (BIC Schwartz, 1978). The `mclust` software is an example where this is used. For a model with parameters Φ , the BIC is given by

$$\text{BIC} = 2l(x, \hat{\Phi}) - m \log n,$$

where $l(x, \hat{\Phi})$ is the maximised log-likelihood, $\hat{\Phi}$ is the MLE of Φ , n is the number of observations and m is the number of free parameters in the model. An asymptotic approximation of the log-posterior probability of the models motivates the use of the BIC (Kass and Raftery, 1995).

In general, mixture models do not satisfy the regularity conditions for the asymptotic approximation that were used in the formation of the BIC. Leroux (1992) demonstrated that, asymptotically, the BIC does not underestimate the actual number of sub-populations in the model and Keribin (1998, 2000) demonstrated that the BIC gives consistent estimates of the number of sub-populations. Additionally, Fraley and Raftery (1998, 2002b) provide evidence that in a mixture model context the BIC performs well as a model selection criterion.

It should be noted that there are alternatives to the BIC for mixture model selection such as the integrated completed likelihood (Biernacki et al., 2000) which subtracts the estimated mean entropy from the BIC.

1.6.2 The Rand and Adjusted Rand Indices

The ability of multiple mixture models to classify unknown observations can be compared using the Rand index (Rand, 1971) and the adjusted Rand index (Hubert and Arabie, 1985). Both of these indices are based on a cross-tabulation of the true group memberships and the maximum *a posteriori* (MAP) classifications. The Rand index is given by

$$\frac{\text{number of pair agreements}}{\text{number of pair agreements} + \text{number of pair disagreements}} \quad (1.6)$$

The adjusted Rand index penalizes the Rand index for classifications that are correct by chance by subtracting the expected value of the Rand index. That is, it takes into account that if all of the observations were randomly classified, some of these classifications would be correct. The Rand index can take on values in the range of $[0,1]$ where an index of 0 indicates that none of the estimated classifications are correct and an index of 1 indicates that all of the estimated classifications are correct. The adjusted Rand index can take on negative values as well. In general, higher Rand and adjusted Rand indices are desirable model characteristics.

Chapter 2

The AECM Algorithm for a Mixture of Multivariate t -Distributions

2.1 Introduction

Traditionally, Gaussian distributions have been employed in model-based clustering analysis (Titterton et al., 1985; McLachlan and Basford, 1988; Celeux and Govaert, 1995; McLachlan and Peel, 2000; McNicholas and Murphy, 2008; McNicholas et al., 2009). However, in more recent years, these methods have been expanded to incorporate multivariate t -distributions (McLachlan and Peel, 1998; McLachlan et al., 2007). The following alternating expectation-conditional maximization (AECM) algorithm, introduced by McLachlan et al. (2007), estimates unknown parameters of a multivariate t -distribution mixture model with a factor analysis covariance structure of $\Sigma_g = \Lambda_g \Lambda_g' + \Psi_g$.

2.2 The Mixtures of Multivariate t -Factor Analyzers Model

Recall that the mixtures of factor analysis method assumes a model of

$$\mathbf{X}_i = \boldsymbol{\mu}_g + \Lambda_g \mathbf{U}_{ig} + \mathbf{e}_{ig},$$

with probability π_g for $g = 1, \dots, G$ and $i = 1, \dots, n$, where $\mathbf{X}$ is a p -dimensional random vector that is modeled by a q -dimensional vector of unobserved latent factors $\mathbf{U}_{ig}$, and where $q \ll p$. Also, the $p \times q$ matrix of factor loadings is denoted by $\mathbf{\Lambda}_g$. The distributions of $\mathbf{u}_{ig}$ and $\mathbf{e}_{ig}$, given $z_{ig} = 1$, are as follows:

$$\mathbf{U}_{ig} \mid z_{ig} = 1 \sim t_q(\mathbf{0}, \mathbf{I}_q, \nu_g),$$

and

$$\mathbf{e}_{ig} \mid z_{ig} = 1 \sim t_p(\mathbf{0}, \mathbf{\Psi}_g, \nu_g).$$

To incorporate the multivariate t -distribution, the following mixtures of t -factor analyzers model is employed

$$f(\mathbf{x}_i \mid \pi_g, \boldsymbol{\mu}_g, \mathbf{\Lambda}_g, \mathbf{\Psi}_g, \nu_g) = \sum_{g=1}^G \pi_g f(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \mathbf{\Sigma}_g, \nu_g), \quad (2.1)$$

where

$$\mathbf{\Sigma}_g = \mathbf{\Lambda}_g \mathbf{\Lambda}_g' + \mathbf{\Psi}_g$$

for $g = 1, \dots, G$ and

$$f(\mathbf{x} \mid \boldsymbol{\mu}_g, \mathbf{\Sigma}_g, \nu_g) = \frac{\Gamma\left(\frac{\nu+p}{2}\right) |\mathbf{\Sigma}_g|^{-1/2}}{(\pi \nu_g)^{\frac{1}{2}p} \Gamma\left(\frac{\nu_g}{2}\right) \left\{1 + \frac{1}{\nu_g} (\mathbf{x} - \boldsymbol{\mu}_g)' \mathbf{\Sigma}_g^{-1} (\mathbf{x} - \boldsymbol{\mu}_g)\right\}^{\frac{1}{2}(\nu_g+p)}},$$

where

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt.$$

Each group is assumed to have a multivariate t -distribution with mean $\boldsymbol{\mu}_g$ and covariance matrix $\mathbf{\Sigma}_g$, where π_g is the probability of membership of sub-population g .

2.3 The Likelihood Function

It follows from Equation 2.1 that the complete-data log-likelihood of $\mathbf{x} = (\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_n)'$, the component indicator variables z_{ig} , the weights w_i and the latent factors $\mathbf{u}_{ig}$ is given by

$$\begin{aligned} \log L_c(\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g, \nu_g \mid \mathbf{x}, z_{ig}, \mathbf{u}_{ig}, w_i) \\ = \sum_{g=1}^G \sum_{i=1}^n z_{ig} \log a_{ig}, \end{aligned}$$

where

$$a_{ig} = \pi_g f_G \left(w_{ig} \mid \frac{1}{2} \nu_g, \frac{1}{2} \nu_g \right) \phi \left(\mathbf{u}_{ig} \mid \mathbf{0}, \frac{1}{w_i} \mathbf{I}_q \right) \phi \left(\mathbf{x}_i \mid \boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{u}_{ig}, \frac{1}{w_{ig}} \boldsymbol{\Psi}_g \right). \quad (2.2)$$

The observations weights, w_{ig} , are a measure of how far an observation lies from the group mean. The smaller an observation's weight, the less impact that observations has on density parameter estimates. Conversely, the larger an observations weight, the more influence it has on density parameter estimates.

The distributions in Equation 2.2 are as follows

$$W_{ig} \sim \text{gamma} \left(\frac{\nu_g}{2}, \frac{\nu_g}{2} \right), \quad (2.3)$$

$$\mathbf{U}_{ig} \mid w_i, z_{ig} = 1 \sim N \left(\mathbf{0}, \frac{1}{w_i} \mathbf{I}_q \right), \quad (2.4)$$

and

$$\mathbf{X}_i \mid \mathbf{u}_{ig}, w_i, z_{ig} = 1 \sim N \left(\boldsymbol{\mu}_g + \boldsymbol{\Lambda}_g \mathbf{u}_{ig}, \frac{1}{w_i} \boldsymbol{\Psi}_g \right). \quad (2.5)$$

From Equations 2.3, 2.4, and 2.5, $\log a_{ig}$ can also be expressed as

$$\log a_{ig} = \sum_{d=1}^4 a_{dig}, \quad (2.6)$$

where

$$\begin{aligned} a_{1ig} &= \log \pi_g, \\ a_{2ig} &= -\log \Gamma\left(\frac{\nu_g}{2}\right) + \frac{\nu_g}{2} \log\left(\frac{1}{2}\nu_g\right) + \frac{1}{2}\nu_g(\log w_i - w_i) - \log w_i, \\ a_{3ig} &= -\frac{1}{2}q \log(2\pi) + \frac{1}{2}q \log w_i + \frac{1}{2}\left(\frac{1}{w_i} \mathbf{u}'_{ig} \mathbf{u}_{ig}\right), \end{aligned}$$

and

$$\begin{aligned} a_{4ig} &= -\frac{1}{2}p \log(2\pi) + \frac{1}{2}p \log w_i - \frac{1}{2} \log |\Psi_g| \\ &\quad - \frac{1}{2} w_i (\mathbf{x}_i - \boldsymbol{\mu}_g - \Lambda_g \mathbf{u}_{ig})' \Psi_g^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g - \Lambda_g \mathbf{u}_{ig}). \end{aligned}$$

2.4 E-Step

In order to calculate the required conditional expectations, we need to make use of the following distributions

$$(\mathbf{X}'_i \mathbf{U}'_{ig})' \mid w_i, z_{ig} = 1 \sim N_{p+q}\left(\boldsymbol{\mu}_g^*, \frac{\boldsymbol{\xi}_g}{w_i}\right), \quad (2.7)$$

where

$$\boldsymbol{\mu}_g^* = (\boldsymbol{\mu}'_g, \mathbf{0}')'$$

and

$$\boldsymbol{\xi}_g = \begin{pmatrix} \Lambda_g \Lambda'_g + \Psi_g & \Lambda_g \\ \Lambda'_g & \mathbf{I}_q \end{pmatrix},$$

as well as

$$\mathbf{U}_{ig} \mid \mathbf{x}_i, w_{ig} \sim N\left(\gamma'_g(\mathbf{x}_i - \boldsymbol{\mu}_g), \frac{\Omega_g}{w_i}\right), \quad (2.8)$$

where

$$\boldsymbol{\gamma}_g = (\boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g)^{-1} \boldsymbol{\Lambda}_g \quad \text{and} \quad \boldsymbol{\Omega}_g = \mathbf{I}_q - \boldsymbol{\gamma}_g' \boldsymbol{\Lambda}_g.$$

From Equations 2.7, and 2.8 we can calculate the required conditional expectations, given by

$$E[Z_{ig} W_{ig} \mathbf{U}_{ig} \mid \mathbf{x}_i, w_{ig}] = w_{ig} \boldsymbol{\gamma}_g' (\mathbf{x}_i - \boldsymbol{\mu}_g),$$

$$E[Z_{ig} W_{ig} \mathbf{U}_{ig} \mathbf{U}_{ig}' \mid \mathbf{x}_i, w_{ig}] = \boldsymbol{\Omega}_g + w_{ig} \boldsymbol{\gamma}_g' (\mathbf{x}_i - \boldsymbol{\mu}_g) (\mathbf{x}_i - \boldsymbol{\mu}_g)' \boldsymbol{\gamma}_g,$$

and

$$E[W_{ig} \mid \mathbf{x}_i, z_{ig} = 1] = w_i(\mathbf{x}_i \mid \pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g, \nu_g),$$

where

$$w_i = \frac{\nu_g + p}{\nu_g + (\mathbf{x}_i - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_g)}.$$

The conditional expectation of the component indicator variable given $\mathbf{x}_i$ is given by the posterior probability that an observed value $\mathbf{x}_i$ belongs to the i th group. This posterior probability can be expressed as

$$\hat{z}_{ig} = \frac{\pi_g f(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g)}{\sum_{h=1}^G \pi_h f(\mathbf{x}_i \mid \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h, \nu_h)},$$

for $g = 1, \dots, G$ and $i = 1, \dots, n$, where $f(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \nu_g)$ is the density of a multivariate t -distribution with mean $\boldsymbol{\mu}_g$, covariance matrix $\boldsymbol{\Sigma}_g$ and degrees of freedom ν_g .

2.5 CM-Steps

To perform the CM-steps of the algorithm, the unknown parameter vector is divided into two sub-vectors. The first sub-vector, $\boldsymbol{\Theta}_1$, contains the mixing proportions π_g , the elements of $\boldsymbol{\mu}_g$ and the degrees of freedom ν_g . The second sub-vector,

Θ_2 , contains the elements of the factor loadings matrix Λ_g and the diagonal matrix Ψ_g .

On the first cycle of the $(k + 1)$ st iteration, the missing data consists of the component indicator variables z_{ig} and the weights w_{ig} . From this, the expected value of the complete-data log likelihood Q is calculated by substituting the previously-calculated expected values for z_{ig} and w_i into the aforementioned complete-data log-likelihood equation.

Differentiating Q with respect to π_g and setting the derivative equal to zero results in the following updated mixing proportions

$$\pi_g^{(k+1)} = \sum_{i=1}^n \frac{\hat{z}_{ig}}{n}. \quad (2.9)$$

Differentiating Q with respect to μ_g and setting the derivative equal to zero results in an updated mean of

$$\mu_g^{(k+1)} = \frac{\sum_{i=1}^n \hat{z}_{ig} \hat{w}_{ig}^{(k)} \mathbf{x}_i}{\sum_{i=1}^n \hat{z}_{ig} \hat{w}_{ig}^{(k)}}, \quad (2.10)$$

where

$$\hat{w}_{ig}^{(k)} = \frac{\nu_g^{(k)} + p}{\nu_g^{(k)} + (\mathbf{x}_i - \mu_g^{(k)})'(\Sigma^{(k)})^{-1}(\mathbf{x}_i - \mu_g^{(k)})}.$$

If the degrees of freedom are not specified before running the algorithm, then $\nu_g^{(k+1)}$ is the solution to the equation

$$\begin{aligned} & -\varphi\left(\frac{1}{2}\nu_g^{(k+1)}\right) + \log\left(\frac{1}{2}\nu_g^{(k+1)}\right) + 1 + \frac{1}{n_g^{(k)}} \sum_{i=1}^n z_{ig}^{(k)} \log w_{ig}^{(k)} \\ & + \varphi\left(\frac{\nu_g^{(k)} + p}{2}\right) - \log\left(\frac{\nu_g^{(k)} + p}{2}\right) = 0, \end{aligned} \quad (2.11)$$

where

$$z_{ig}^{(k)} = z_g(\mathbf{x}_i \mid \pi_g^{(k)}, \boldsymbol{\mu}_g^{(k)}, \boldsymbol{\Lambda}_g^{(k)}, \boldsymbol{\Psi}_g^{(k)}, \nu_g^{(k)}),$$

$$n_g^{(k)} = \sum_{i=1}^n z_{ig}^{(k)},$$

and $\varphi(\cdot)$ is the digamma function. That is,

$$\varphi(\cdot) = \frac{d}{dx} \ln \Gamma(\cdot).$$

Once the first cycle is complete, the estimate of the parameter space Θ is updated to $\Theta^{(k+\frac{1}{2})} = (\Theta_1^{(k+1)'}, \Theta_2^{(k)'})$.

On the second cycle of the $(k+1)$ st iteration, the missing data is expanded to include the unobserved latent factors $\mathbf{u}_{ig}$ associated with the $\mathbf{x}_i$. After substituting the expected values for $\mathbf{u}_{ig}$, $\mathbf{u}_{ig}\mathbf{u}_{ig}'$, z_{ig} and w_i into the complete-data log likelihood, the update for $\boldsymbol{\Lambda}_g$ can be calculated by differentiating Q with respect to $\boldsymbol{\Lambda}_g$ and setting equal to zero. That is

$$\boldsymbol{\Lambda}_g^{(k+1)} = \mathbf{V}_g^{(k+\frac{1}{2})} (\boldsymbol{\gamma}_g^{(k)'} \mathbf{V}_g^{(k+\frac{1}{2})} \boldsymbol{\gamma}_g^{(k)} + \boldsymbol{\Omega}_g^{(k)})^{-1}, \quad (2.12)$$

where

$$\mathbf{V}_g^{(k+\frac{1}{2})} = \frac{\sum_{i=1}^n z_g(\mathbf{x}_i \mid \pi_g^{(k+1)}, \boldsymbol{\mu}_g^{(k+1)}, \boldsymbol{\Lambda}_g^{(k)}, \boldsymbol{\Psi}_g^{(k)}, \nu_g^{(k+1)}) w_{ig}^{(k+\frac{1}{2})} (\mathbf{x}_i - \boldsymbol{\mu}_g^{(k+1)}) (\mathbf{x}_i - \boldsymbol{\mu}_g^{(k+1)})'}{\sum_{i=1}^n z_g(\mathbf{x}_i \mid \pi_g^{(k)}, \boldsymbol{\mu}_g^{(k)}, \boldsymbol{\Lambda}_g^{(k)}, \boldsymbol{\Psi}_g^{(k)}, \nu_g^{(k+1)})},$$

$$\boldsymbol{\gamma}_g^{(k)} = (\boldsymbol{\Lambda}_g^{(k)} \boldsymbol{\Lambda}_g^{(k)'} + \boldsymbol{\Psi}_g^{(k)})^{-1} \boldsymbol{\Lambda}_g^{(k)},$$

$$\boldsymbol{\Omega}_g^{(k)} = \mathbf{I}_q - \boldsymbol{\gamma}_g^{(k)'} \boldsymbol{\Lambda}_g^{(k)},$$

and $w_{ig}^{(k+\frac{1}{2})}$ is partially updated by using $\Theta^{(k+\frac{1}{2})}$ for Θ .

The update $\Psi_g^{(k+1)}$ is obtained by differentiating the expected complete-data log likelihood function with respect to Ψ_g and setting equal to zero. That is

$$\begin{aligned}\Psi_g^{(k+1)} &= \text{diag} \left\{ \mathbf{V}_g^{(k+\frac{1}{2})} - \Lambda_g^{(k+1)} \mathbf{H}_g^{(k+\frac{1}{2})} \Lambda_g^{(k+1)'} \right\} \\ &= \text{diag} \left\{ \mathbf{V}_g^{(k+\frac{1}{2})} - \mathbf{V}_g^{(k+\frac{1}{2})} \gamma_g^{(k)} \Lambda_g^{(k+1)'} \right\},\end{aligned}\tag{2.13}$$

where

$$\mathbf{H}_g^{(k+\frac{1}{2})} = \gamma_g^{(k)'} \mathbf{V}_g^{(k+\frac{1}{2})} \gamma_g^{(k)} + \Omega_g^{(k)}.$$

Chapter 3

Model-Based Classification

3.1 Discriminant Analysis

Classically, classification has been implemented through the use of methods such as discriminant analysis (Fisher, 1936; Bensmail and Celeux, 1996; Fraley and Raftery, 2002b). Discriminant analysis assumes that some subset of the observations have known classifications, and also assumes that the true number of groups, G , is also known. The known observations are then used to classify the remaining unknown observations.

Many methods of discriminant analysis are based on the assumption that the observations in each group are generated by a probability distribution. Each group can be modeled by its own probability distribution $f_g(\cdot)$. Using Bayes's theorem, the posterior probability that an observation $\mathbf{x}$ belongs to group g is

$$P(\mathbf{x} \in \text{group}_g) = \frac{\pi_g f_g(\mathbf{x})}{\sum_{g=1}^G \pi_g f_g(\mathbf{x})},$$

where π_g is the proportion of observations that belong to group g . Each observation is then assigned to the group with the highest posterior probability, which minimizes the expected misclassification rate. This technique is called Bayes classification.

Linear discriminant analysis (LDA; Fisher, 1936) is another method commonly used in classification applications (Friedman, 1989; Fukunaga, 1990), in which the classification rule is based on a linear combination of the explanatory variables. That is, the probability of observation $\mathbf{x}$ being put into group g is a function of a linear combination of explanatory variables from the observations whose classifications are known. This linear rule is then used to classify unknown observations. It is assumed that the $P(\mathbf{x} \mid g = 1), P(\mathbf{x} \mid g = 2), \dots, P(\mathbf{x} \mid g = G)$ are normally distributed, and that groups have equal covariance structures. This method is available as the `lda` function in the `MASS` package (Venables and Ripley, 2002) in R.

Bayes classification, linear discriminant analysis, and other discriminant analysis methods do not incorporate the unknown observations in the development of classification rules. However, in other statistical literature (Titterton, 1976; McLachlan and Ganesalingam, 1982; Dean et al., 2006; Toher et al., 2007), techniques have been developed that incorporate the observations with unknown classifications in updating the classification rules. Dean et al. (2006) developed a classification technique that updates the discriminant analysis rule by using the observations for which group memberships are unknown. This is done by using both known and unknown observations to estimate parameters of the mixture model, and the resulting finite mixture model provides the rules by which the unlabeled data can be classified. An EM algorithm, or a variant thereof, can be used to estimate the model parameters.

3.2 Model-Based Classification *via* a Mixture of Multivariate *t*-Factor Analyzers

The model for classification through the use of a mixture of *t*-factor analyzers is the same as in Equation 2.1, however some aspects of the likelihood and parameter estimates are affected. Consider $\mathbf{x}$, a set of n observations, the first m of which have known classifications, and $n - m$ of which have unknown classifications. The complete-data likelihood for the first stage of the AECM algorithm is then

$$\begin{aligned} L_1(\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g, \nu_g \mid \mathbf{x}, \mathbf{z}, w_i) \\ = \prod_{i=1}^m \prod_{g=1}^G \left[\pi_g f(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g, \nu_g) \phi(w_i \mid \frac{\nu_g}{2}, \frac{\nu_g}{2}) \right]^{z_{ig}} \\ \times \prod_{i=m+1}^n \prod_{g=1}^G \left[\pi_g f(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g, \nu_g) \phi(w_i \mid \frac{\nu_g}{2}, \frac{\nu_g}{2}) \right]^{\hat{z}_{ig}}. \end{aligned}$$

For the following analyses, the parameter estimates for π_g and $\boldsymbol{\mu}_g$ obtained by maximizing the expected value of the previous complete-data likelihood are

$$\hat{\boldsymbol{\mu}}_g^{(k+1)} = \frac{\sum_{i=1}^m z_{ig} w_{ig} \mathbf{x}_i + \sum_{i=m+1}^n \hat{z}_{ig} w_{ig} \mathbf{x}_i}{n_g w_{ig}} \text{ and } \hat{\pi}_g^{(k+1)} = \frac{n_g}{n}.$$

The estimate for ν_g can be found by solving the equation

$$\begin{aligned} -\varphi\left(\frac{1}{2}\nu_g^{(k+1)}\right) + \log\left(\frac{1}{2}\nu_g^{(k+1)}\right) + 1 + \frac{1}{n_g^{(k)}} \sum_{i=1}^m z_{ig}^{(k)} \log w_{ig}^{(k)} \\ + \frac{1}{n_g^{(k)}} \sum_{i=m+1}^n \hat{z}_{ig}^{(k)} (\log w_{ig}^{(k)}) + \varphi\left(\frac{\nu_g^{(k)} + p}{2}\right) - \log\left(\frac{\nu_g^{(k)} + p}{2}\right) = 0. \end{aligned}$$

The complete-data likelihood at the second stage of the algorithm is

$$\begin{aligned}
& L_2(\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g, \nu_g \mid \mathbf{x}, \mathbf{z}, \mathbf{u}_{ig}, w_i) \\
&= \prod_{i=1}^m \prod_{g=1}^G \left[\pi_g f(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g, \nu_g) f(\mathbf{u}_{ig} \mid \mathbf{0}, \frac{\mathbf{I}_q}{w_i}) \phi(w_i \mid \frac{\nu_g}{2}, \frac{\nu_g}{2}) \right]^{z_{ig}} \\
&\quad \times \prod_{i=m+1}^n \prod_{g=1}^G \left[\pi_g f(\mathbf{x}_i \mid \boldsymbol{\mu}_g, \boldsymbol{\Lambda}_g, \boldsymbol{\Psi}_g, \nu_g) f(\mathbf{u}_{ig} \mid \mathbf{0}, \frac{\mathbf{I}_q}{w_i}) \phi(w_i \mid \frac{\nu_g}{2}, \frac{\nu_g}{2}) \right]^{\hat{z}_{ig}}.
\end{aligned}$$

However, the parameter estimates for $\boldsymbol{\Lambda}_g$, and $\boldsymbol{\Psi}_g$, are of the same form as in equations 2.12 and 2.13 as they do not directly depend on the true or estimated values of z_{ig} .

This method will also consider an additional covariance structure as well as the one used by McLachlan et al. (2007). The covariance structure previously introduced by McLachlan and Peel (2000) in the mixtures of factor analyzers model, $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}_g$, will be denoted as the UUU structure. The additional covariance structure will apply the constraint that $\boldsymbol{\Psi}_g$ must be equal across all groups such that $\boldsymbol{\Psi}_g = \boldsymbol{\Psi}$ and $\boldsymbol{\Sigma}_g = \boldsymbol{\Lambda}_g \boldsymbol{\Lambda}_g' + \boldsymbol{\Psi}$. This additional structure will be denoted as the UCU structure. If the UCU covariance structure is implemented, then the update for $\boldsymbol{\Psi}$ is

$$\boldsymbol{\Psi}^{(k+1)} = \frac{\sum_{g=1}^G \pi_g^{(k+1)} \boldsymbol{\Psi}_g^{(k+1)}}{G},$$

where G is the true number of groups. The other updates remain the same with $\boldsymbol{\Psi}$ in place of $\boldsymbol{\Psi}_g$.

From these likelihood functions and parameter updates, it is apparent that this approach provides a joint framework for model-based classification. Specifically, both the information from the known observations and group memberships and the

estimated information from the unknown observations are incorporated at every step of the algorithm.

3.3 AECM Algorithm for Model-Based Classification *via* a Mixture of Multivariate t -Factor Analyzers

A summary of the AECM algorithm for model-based classification *via* a mixture of multivariate t -factor analyzers is as follows. Note that this algorithm can be applied to test cases, that is data sets for which we know all of the true classifications. Also, the initializations for Λ_g and Ψ_g are the same as those used in McNicholas and Murphy (2008). Specifically, Λ_g is initialized with elements $\hat{\lambda}_{ij} = \sqrt{d_j} \rho_{ij}$ where d_j is the j^{th} largest eigenvalue of $\tilde{\mathbf{S}}_g$ and ρ_{ij} is the j^{th} element of the eigenvector corresponding to the j^{th} largest eigenvalue of $\tilde{\mathbf{S}}_g$ for $i = (1, 2, \dots, p)$ and $j = (1, 2, \dots, q)$.

if a test case

define what proportion of observations have known classifications

randomly choose which observations have known classifications

end if

initialize unknown $\hat{\mathbf{z}}$ as `mclust` classifications

initialize $\hat{n}_g, \hat{\pi}_g, \hat{\mu}_g, \tilde{\mathbf{S}}$ as per updates

initialize $\hat{\nu}_g = 100$

run Householder's algorithm with QL reduction on $\tilde{\mathbf{S}}$ to get $\tilde{\mathbf{S}} = \mathbf{PDP}^{-1}$

initialize $\hat{\Lambda}$

initialize $\hat{\psi}$ with $\hat{\Psi}_{\mathbf{g}} = \text{diag}(\tilde{\mathbf{S}}_g - \hat{\Lambda}_g \hat{\Lambda}'_g)$

initialize $\hat{w}_{ig}$

while not converged:

CM-step 1 $\left\{ \begin{array}{l} \text{update } \hat{\pi}_g \\ \text{update } \hat{\mu}_g \\ \text{update } \hat{\nu}_g \end{array} \right.$

E-step 2 $\left\{ \begin{array}{l} \text{if not first iteration:} \\ \quad \text{update } \hat{w}_{ig} \\ \quad \text{update unknown } \hat{\mathbf{z}} \\ \text{end if} \end{array} \right.$

update $\hat{\mathbf{V}}, \hat{\gamma}, \hat{\Omega}$

CM-step 2 $\left\{ \begin{array}{l} \text{update } \hat{\Lambda} \\ \text{update } \hat{\Psi} \end{array} \right.$

E-step 1 $\left\{ \begin{array}{l} \text{update } \hat{w}_{ig} \\ \text{update unknown } \hat{\mathbf{z}} \end{array} \right.$

```

compute log-likelihood

check for convergence using Aitken's Acceleration

set  $\hat{\nu}_g^{(k+1)} = \hat{\nu}_g^{(k)}$ 

end while

compute BIC

```

3.4 Model Selection for Model-Based Classification *via* a Mixture of Multivariate *t*-Factor Analyzers

In the subsequent data analyses, the ‘best’ model will be chosen by analyzing various numbers of factors associated with a fixed value of G . Two covariance structures are considered. These include UUU, where there are no constraints applied to the components of the covariance matrix, and UCU, where Ψ_g is constrained to be equal for all groups such that $\Psi_g = \Psi$. The $\hat{z}$ matrix is initialized using the classification results from the `mclust` package (Fraley and Raftery, 2002a, 2003) in R. That is, MCLUST is run on the data set and the classifications from the best MCLUST model are used as the initial $\hat{z}$ values for unknown observations. Initially, the group memberships of the unknown observations were assigned randomly and the resulting classifications were very similar to those obtained using the MCLUST starting values, however the algorithm took significantly longer to converge. For this reason, initializing unknown z_{ig} values using MCLUST classifications was used to decrease the algorithm convergence time. The observations whose group memberships are considered known will influence the final parameter estimates and classifications, and therefore multiple

subsets of known observations will be considered. First, the data is analyzed with G fixed at the true number of groups and the number of factors q ranging from one to five. Both the UUU and UCU covariance structures are used. The BIC for each of these models is calculated once convergence is reached. For the UUU models, the BIC is given by

$$2l(\mathbf{x}, \hat{\Phi}) - \left\{ (G - 1) + Gp + \left[G \left(pq - \frac{q(q - 1)}{2} \right) \right] + Gp + G \right\} \log n,$$

and for the UCU models, the BIC is given by

$$2l(\mathbf{x}, \hat{\Phi}) - \left\{ (G - 1) + Gp + \left[G \left(pq - \frac{q(q - 1)}{2} \right) \right] + p + G \right\} \log n.$$

For test cases, this process is repeated five times with different randomly selected subsets of known observations. To obtain these random subsets of known observations, each observation is assigned a random number between one and one hundred. If this assigned number is less than or equal to the pre-specified known proportion then the observation is treated as being known. If the assigned number is greater than the pre-specified known proportion, then that observation is treated as unknown.

From the five starts, the number of factors associated with the highest BIC of all considered models, and for all five random starts is deemed the best q value. The covariance structure associated with the highest BIC of all considered models, and for all five random starts is also deemed the most appropriate for the data.

This best q value and covariance structure are then run again, with a different random subset of known observations. The parameter estimates and classifications associated with this final model are chosen as the best estimates for the data.

Convergence is tested using Aitken's Acceleration. That is, the algorithm is said to converge when (Lindsay, 1995; McNicholas et al., 2009)

$$l_{\infty}^{(k+1)} - l^{(k)} < \varepsilon,$$

where

$$l_{\infty}^{(k+1)} = l^{(k)} + \frac{1}{1 - a_k} (l^{(k+1)} - l^{(k)}),$$

$$a_k = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}},$$

where $l_{\infty}^{(k+1)}$ is the asymptotic estimate of the log-likelihood at the current iteration, $l^{(k)}$ is the log-likelihood at the previous iteration, and ε was pre-specified to equal 0.1.

Chapter 4

Data Analysis: Italian Wines

4.1 The Data

Twenty-eight physical and chemical properties of 178 samples of three types of Italian wine, Barolo, Grignolino, and Barbera, were reported by Forina et al. (1986). A subset of thirteen of these variables are available through the `gclus` package (Hurley, 2004) in R as the `wine` data set. The thirteen variables analyzed to estimate classification are displayed in Table 4.1 and the true mixing proportions of this data set are displayed in Table 4.2. A pairs plot of this variable subset is shown in Figure 4.1. Classification of the data by wine type is performed using mixtures of multivariate *t*-factor analyzers and linear discriminant analysis. These proportions of data treated as being known are 60% (Section 4.2), 70% (Section 4.3), 80% (Section 4.4), and 90% (Section 4.5). These two classification techniques are then compared in Section 4.6.

Table 4.1: Thirteen of the chemical and physical properties of wine reported by Forina et al. (1986).

Alcohol	Flavonoids	Color intensity	Malic acid
Ash	Nonflavonoid phenols	Hue	Proline
Alcalinity of ash	Magnesium	Total phenols	Proanthocyanins
OD ₂₈₀ /OD ₃₁₅ of diluted wines			

Table 4.2: True mixing proportions of wine reported by Forina et al. (1986).

Barolo	Grignolino	Barbera
0.3315	0.3989	0.2700

Figure 4.1: Pairs plot of physical and chemical measurements from the Italian wine data.

4.2 Proportion of Data Known: 60%

4.2.1 Classification via MMtFAs

The data was analyzed with five different subsets of known data for $q = 1, 2, \dots, 5$ and $G = 3$ where 60% of the classifications were considered known. The

overall maximum BIC was -5304.474 with $G = 3$, $q = 2$ and the UUU covariance structure. This combination was run again with a new sample of known observations.

This model had a BIC of -5303.628 and produced the classifications in Table 4.3, with a Rand index of 0.9792, an adjusted Rand index of 0.9529, and an error rate of 1.56%. Selected associated estimated model parameters are displayed in tables 4.4 and 4.5.

Table 4.3: Estimated classifications for UUU, $G = 3$ and $q = 2$, for the wine data with 60% known.

	1	2	3
Barolo	17		
Grignolino	1	25	
Barbera			21

Table 4.4: Best model estimated degrees of freedom, for the wine data with 60% known.

Barolo	Grignolino	Barbera
68.61	8.21	130.57

Table 4.5: Best model estimated mixing proportions, for the wine data with 60% known.

Barolo	Grignolino	Barbera
0.3382	0.3921	0.2697

Note that the degrees of freedom for each group, ν_g , was allowed a maximum value of 200. From Table 4.4 we can see that the Barolo and Barbera groups have high values for their estimated degrees of freedom, indicating that they could be reasonably modeled using Gaussian distributions. However, the Grignolino group has estimated degrees of freedom of 8.21, and it might not be reasonable to model

this with a Gaussian distributions. This is especially important since the Grignolino group comprises the largest proportion of the data, and therefore modeling it with a Gaussian distribution may have a substantial impact on the estimated classifications.

4.2.2 Classification *via* LDA

Linear discriminant analysis was also performed on the wine data. The training set consisted of the same observations that were selected as known for the final model when using the mixture of multivariate t -factors. The classification rule developed by analyzing the training set was then applied to the remaining observations, also known as the test set, and the classifications of these observations were predicted. This resulted in the group memberships displayed in Table 4.6, with a Rand index of 0.9568, an adjusted Rand index of 0.9024 and an error rate of 3.12%.

Table 4.6: LDA predicted classifications, for the wine data with 60% known.

	1	2	3
Barolo	17		
Grignolino	1	24	1
Barbera			21

4.3 Proportion of Data Known: 70%

4.3.1 Classification *via* MMtFAs

As in the previous section, the data were analyzed for $q = 1, 2, \dots, 5$, $G = 3$, and both the UUU and UCU covariance structures, for five random samples of known observations. The overall maximum BIC was -5299.005, with $G = 3$, $q = 3$ and the

UUU covariance structure. This combination was run again with a new randomly selected sample of known observations.

The BIC for the best combination of covariance structure, G and q was -5318.501 , and produced the classifications in Table 4.7, with a Rand index of 0.9755, an adjusted Rand index of 0.9463, and an error rate of 1.54%. Selected associated estimated model parameters are displayed in tables 4.8, and 4.9.

Table 4.7: Estimated classifications for UUU, $G = 3$ and $q = 2$, for the wine data with 70% known.

	1	2	3
Barolo	23		
Grignolino	1	28	
Barbera			13

Table 4.8: Best model estimated degrees of freedom, for the wine data with 70% known.

Barolo	Grignolino	Barbera
200.00	10.46	200.00

Table 4.9: Best model estimated mixing proportions, for the wine data with 70% known.

Barolo	Grignolino	Barbera
0.3350	0.3953	0.2697

4.3.2 Classification via LDA

Subsequently, linear discriminant analysis was performed using the known observations from the model in the previous section as the training set, and the remaining unknown observations as the test set. The predicted classifications are

displayed in Table 4.10, and had an associated Rand index of 0.9492, an adjusted Rand index of 0.8845 and an error rate of 3.77%.

Table 4.10: LDA predicted classifications, for the wine data with 70% known.

	1	2	3
Barolo	23		
Grignolino	2	26	1
Barbera			13

4.4 Proportion of Data Known: 80%

4.4.1 Classification *via* MMtFAs

The data was analyzed for $G = 3$, $q = 1, 2, \dots, 5$ and for both the UUU and UCU covariance structures. Five different randomly chosen subsets were chosen for the 80% of data being treated as having known group memberships. The highest BIC for these five starts was -5304.516 with $G = 3$, $q = 2$ and the UUU covariance structure. This combination was run with another subset of known observations.

The model had a BIC of -5305.113. The associated classifications are shown in Table 4.11. These estimated classifications have a Rand index of 1.000, an adjusted Rand index of 1.000 and an error rate of 0.00%. Selected associated model parameters are displayed in tables 4.12 and 4.13.

4.4.2 Classification *via* LDA

The subset that was treated as known in the previous MMtFA model was also analyzed using linear discriminant analysis. The predicted classifications of the

Table 4.11: Estimated classifications for UUU, $G = 3$ and $q = 2$, for the wine data with 80% known.

	1	2	3
Barolo	13		
Grignolino		13	
Barbera			9

Table 4.12: Best model estimated degrees of freedom, for the wine data with 80% known.

Barolo	Grignolino	Barbera
86.62	8.44	143.24

Table 4.13: Best model estimated mixing proportions, for the wine data with 80% known.

Barolo	Grignolino	Barbera
0.3326	0.3977	0.2697

remaining unknown observations are displayed in Table 4.14. These classifications had a Rand index of 0.9580, an adjusted Rand index of 0.9040 and an error rate of 2.86%.

Table 4.14: LDA predicted classifications, for the wine data with 80% known.

	1	2	3
Barolo	13		
Grignolino	1	12	
Barbera			9

4.5 Proportion of Data Known: 90%

4.5.1 Classification *via* MMtFAs

The AEEM algorithm was run for $G = 3$, $q = 1, 2, \dots, 5$, the UUU and UCU covariance structures and 90% of the observations were treated as having known classifications. This 90% sample was randomly selected for each of the five starts. The highest BIC was -5305.241, with $G = 3$, $q = 2$ and the UUU covariance structure.

The BIC for the final model was -5305.266, and the associated classifications are in Table 4.15, with a Rand index of 1.0000, an adjusted Rand index of 1.0000, and an error rate of 0.00%. Selected estimated model parameters are displayed in tables 4.16, and 4.17.

Table 4.15: Estimated classifications for UUU, $G = 3$ and $q = 2$, for the wine data with 90% known.

	1	2	3
Barolo	4		
Grignolino		7	
Barbera			2

Table 4.16: Best model estimated degrees of freedom, for the wine data with 90% known.

Barolo	Grignolino	Barbera
90.16	8.42	143.22

Table 4.17: Best model estimated mixing proportions, for the wine data with 90% known.

Barolo	Grignolino	Barbera
0.3329	0.3974	0.2697

4.5.2 Classification *via* LDA

Linear discriminant analysis was performed on the data, using the known observations from the final model as the training set, and the remaining observations as the test set. The resulting classifications are displayed in Table 4.18, and had an associated Rand index of 0.8974, an adjusted Rand index of 0.7699, and an error rate of 7.69%.

Table 4.18: Wine 90%: LDA predicted classifications.

	1	2	3
Barolo	4		
Grignolino		6	1
Barbera			2

4.6 Method Comparisons

The number of factors chosen for the model was between 2 when 60%, 80% and 90% of the data was known, and 3 when 70% of the data was known. This makes sense when referring to the pairs plot of the data in Figure 4.1. Because there are not many strong linear relationships between the variables, one would expect that multiple linear combinations of these variables would be required to account for most of the variability in the original data set.

Summaries of the various models considered are displayed in tables 4.19, 4.20, 4.22 and 4.21. At all four proportions of known data the mixtures of multivariate t -factor analyzers outperform the linear discriminant analysis by correctly classifying more of the unknown observations. As the proportion of known data increases, the

mixture of multivariate t -factor analyzers technique's ability to correctly classify unknown observations appears to have an increasing trend, whereas there do not appear to be any trends in LDA's ability to do the same.

Table 4.19: Mixtures of multivariate t -factor analyzers summary for the wine data.

Proportion known	Rand index	Adjusted Rand index	Error rate (%)
60%	0.9791	0.9529	1.56
70%	0.9755	0.9463	1.54
80%	1.0000	1.0000	0.00
90%	1.0000	1.0000	0.00

Table 4.20: Linear discriminant analysis summary for the wine data.

Proportion known	Rand index	Adjusted Rand index	Error rate (%)
60%	0.9568	0.9024	3.12
70%	0.9332	0.8530	4.62
80%	0.9580	0.9040	2.86
90%	0.8974	0.7699	7.69

For all levels of known data that were tested, the Barolo and Barbera groups had relatively high degrees of freedom, and therefore could be adequately modelled using the Gaussian distribution. However, the Grignolino group had very low degrees of freedom indicating that the t -distribution is a better statistical model for this group than the Gaussian distribution, and therefore the MMtFAs method is more suited to

Table 4.21: Estimated degrees of freedom summary for the wine data.

Proportion known	Barolo	Grignolino	Barbera
60%	68.61	8.21	130.57
70%	200.00	10.46	200.00
80%	86.62	8.44	143.24
90%	90.16	8.42	143.22

Table 4.22: Mixing proportions summary for the wine data.

	Barolo	Grignolino	Barbera
True	0.3315	0.3989	0.2700
60% known	0.3382	0.3921	0.2697
70% known	0.3350	0.3953	0.2697
80% known	0.3326	0.3977	0.2697
90% known	0.3329	0.3974	0.2697

this data than a mixture of Gaussians.

At all four levels of known data, the UUU covariance structure was chosen over the UCU covariance structure *via* the BIC even though the UUU model requires the estimation of a larger number of free parameters. This indicates that the wine data is better described by a different Ψ matrix for each of the three groups and that these matrices are substantially different such that the estimation of a common noise matrix does not adequately describe the groups.

Chapter 5

Data Analysis: Italian Olive Oils

5.1 The Data

Forina et al. (1983) described a data set of 572 olive oil samples from three different areas in Italy. This data is available from the GGobi software (Swayne et al., 2006). The percent composition for eight different fatty acids were reported for each sample. These variables used in the analysis are listed in Table 5.1 and the true mixing proportions are displayed in tables 5.2 and 5.3. A pairs plot of these eight variables is shown in Figure 5.1. Classification of the data by geographic origin is performed using mixtures of multivariate t -factor analyzers and linear discriminant analysis. The geographic origin of the oils can be classified in two ways. Broadly, the samples come from three areas, Southern Italy, Sardinia and Northern Italy. These areas can be further divided into nine regions: in Southern Italy there is North Apulia, Calabria, South Apulia and Sicily; Sardinia is broken into Inland Sardinia and Coastal Sardinia; and Northern Italy comprises Umbria, East Liguria and West Liguria. The observations were classified into both the three group and nine group classification structures. The proportions of data treated as being known are 60% (Section 5.2), 70% (Section 5.3), 80% (Section 5.4, and 90% (Section 5.5). These two classification

Table 5.1: Eight fatty acids of olive oil reported by Forina et al., 1983.

Palmitic acid	Palmitoleic acid	Stearic acid	Oleic acid
Linoleic acid	Linolenic acid	Arachidic acid	Eicosenoic acid

techniques are then compared in Section 5.6.

Table 5.2: True mixing proportions for the olive oil data's three group structure.

Southern Italy	Sardinia	Northern Italy
0.5647	0.1713	0.2640

Table 5.3: True mixing proportions for the olive oil data's nine group structure.

North Apulia	Calabria	South Apulia	Sicily	Inland Sardinia
0.0437	0.0979	0.3601	0.0629	0.1136
Coastal Saridinia	East Liguria	West Liguria	Umbria	
0.0577	0.0874	0.0874	0.0892	

5.2 Proportion of Data Known: 60%

5.2.1 Classification *via* MMtFAs

The olive oil data was analyzed for $q = 1, 2, \dots, 5$, the three and nine group structures and both the UUU and UCU covariance structures, with 60% of the data considered to have known classifications. The sample of known observations was chosen randomly for each of five starts. When considering the three group structure the highest BIC of these five starts was -5700.370 with $q = 5$ and the UUU covariance structure.

The final model had $q = 5$ and the UUU covariance structure with a BIC of

Figure 5.1: Pairs plot of fatty acid measurements from the Italian olive oil data.

-5617.726. The associated classifications, and selected model parameters are displayed in tables 5.4, 5.5 and 5.6. This model had a Rand index of 0.9962, an adjusted Rand index of 0.9922 and an error rate of 0.44%.

When the nine group structure was considered the highest BIC for the initial five random starts was -5607.712 with $q = 4$ and the UUU covariance structure. The

Table 5.8: Best model estimated degrees of freedom for $G = 9$, for the olive oil data with 60% known.

North Apulia	Calabria	South Apulia	Sicily	Inland Sardinia
100.00	100.00	13.06	7.38	15.62
Coastal Saridinia	East Liguria	West Liguria	Umbria	
100.00	5.87	100.00	1.92	

Table 5.9: Best model estimated mixing proportions for $G = 9$, for the olive oil data with 60% known.

North Apulia	Calabria	South Apulia	Sicily	Inland Sardinia
0.0436	0.1039	0.3596	0.0576	0.1136
Coastal Saridinia	East Liguria	West Liguria	Umbria	
0.0577	0.0892	0.0874	0.0874	

5.2.2 Classification *via* LDA

Linear discriminant analysis was also performed on the olive oil data for both group structures using the randomly selected known observations as the training set, and the remaining unknown observations as the test set. The classifications produced for the three group structure are shown in Table 5.10, and had an associated Rand index of 0.9888, an adjusted Rand index of 0.9770 and an error rate of 1.33%. The classifications produced for the nine group structure are in Table 5.11, and had an associated Rand index of 0.9806, an adjusted Rand index of 0.9332 and an error rate of 6.57%.

Table 5.10: Olive Oil 60%: LDA Estimated classifications for $G = 3$.

	1	2	3
Southern Italy	128		
Sardinia		34	
Northern Italy		3	61

Table 5.11: Olive Oil 60%: LDA Estimated classifications for $G = 9$.

	1	2	3	4	5	6	7	8	9
North Apulia	8	1							
Calabria		15		1					
South Apulia			74	1					
Sicily	3	5		9					
Inland Sardinia					27				
Coastal Sardinia						12			
East Liguria							21		2
West Liguria								15	
Umbria							1		18

5.3 Proportion of Data Known: 70%

5.3.1 Classification *via* MMtFAs

The olive oil data was run again in the same manner as the previous section, however, this time 70% of the was randomly selected and treated as being known. The AECM algorithm was run five times for $q = 1, 2, \dots, 5$, the three group and nine group structures and both the UUU and UCU covariance structures, with a different randomly selected subset of known data each time. When $G = 3$ was considered, the highest BIC was -5588.015 with $q = 5$ and the UUU covariance structure. This best combination of G , q and covariance structure was then run again and the BIC value for this model was -5582.762. The associated classifications with a rand index of 1.0000, an adjusted Rand index of 1.0000 and an error rate of 0.00% are displayed in Table 5.12. Selected model parameters are displayed in tables 5.13 and 5.14.

When the nine group structure was considered the highest BIC of the initial five random starts was -5632.552 with $q = 4$ and the UUU covariance structure. This

Table 5.12: Estimated classifications for $G = 3$, $q = 5$ and UUU, for the olive oil data with 70% known.

	1	2	3
Southern Italy	97		
Sardinia		30	
Northern Italy			39

Table 5.13: Best model estimated degrees of freedom for $G = 3$, for the olive oil data 70% known.

Southern Italy	Sardinia	Northern Italy
100.00	7.63	6.57

Table 5.14: Best model estimated mixing proportions for $G = 3$, for the olive oil data 70% known.

Southern Italy	Sardinia	Northern Italy
0.5647	0.1712	0.2641

combination was subsequently run another time, resulting in a BIC of -5610.445 and the estimated classifications from this model are shown in Table 5.15. These classifications had a Rand index of 0.9878, an adjusted Rand index of 0.9568 and an error rate of 2.79%. Other selected model parameters are displayed in tables 5.16 and 5.17.

Table 5.15: Estimated classifications for $G = 9$, $q = 4$ and UUU, for the olive oil data with 70% known.

[illegible]

Table 5.16: Best model estimated degrees of freedom for $G = 9$, for the olive oil data with 70% known.

North Apulia	Calabria	South Apulia	Sicily	Inland Sardinia
100.00	100.00	13.31	7.02	100.00
Coastal Saridinia	East Liguria	West Liguria	Umbria	
100.00	6.01	6.98	2.00	

Table 5.17: Best model estimated mixing proportions for $G = 9$, for the olive oil data with 70% known.

North Apulia	Calabria	South Apulia	Sicily	Inland Sardinia
0.0435	0.0958	0.3623	0.0631	0.1119
Coastal Saridinia	East Liguria	West Liguria	Umbria	
0.0594	0.0910	0.0873	0.0857	

5.3.2 Classification *via* LDA

Linear discriminant analysis was performed using the 70% sample of known observations from the final model as the training set, and the remaining observations as the test set. When considering the three group structure, the classifications in Table 5.18 were produced with a Rand index of 0.9714, an adjusted Rand index of 0.9415 and an error rate of 3.01%.

Table 5.18: Olive Oil 70%: LDA Estimated classifications for $G = 3$.

	1	2	3
Southern Italy	96		1
Sardinia		30	
Northern Italy		4	35

When the nine group structure was considered LDA produced the estimated classifications in Table 5.19 with a Rand index of 0.9800, an adjusted Rand index of 0.9302 and an error rate of 5.59%.

Table 5.19: Olive Oil 70%: LDA Estimated classifications for $G = 9$.

	1	2	3	4	5	6	7	8	9
North Apulia	8			1					
Calabria		15							
South Apulia			61						
Sicily		2	1	10					
Inland Sardinia					19				
Coastal Sardinia					2	6			
East Liguria							15		3
West Liguria								17	
Umbria									18

5.4 Proportion of Data Known: 80%

5.4.1 Classification *via* a Mixture of Multivariate t -Factor Analyzers

The olive oil data was analyzed for 80% of the data being treated as known. Again, the algorithm was run five times for both the three group and nine group structures for $q = 1, 2, \dots, 5$ and both the UUU and UCU covariance structures, with five different randomly selected subsets of known observations. When considering the three group structure, the highest BIC was -5572.298 with $q = 5$ and the UUU covariance structure.

This model was subsequently run another time and resulted in a BIC value of -5577.531 and produced the classifications in Table 5.20, and have a Rand index of 1.000, an adjusted Rand index of 1.000 and an error rate of 0.00%. Selected model parameters are displayed in tables 5.21 and 5.22.

When considering the nine group structure the highest BIC of the initial five starts was -5657.505 with $q = 5$ and the UUU covariance structure. This combination

Table 5.20: Estimated classifications for $G = 3$, $q = 5$ and UUU, for the olive oil data with 80% known.

	1	2	3
Southern Italy	64		
Sardinia		16	
Northern Italy			26

Table 5.21: Best model estimated degrees of freedom for $G = 3$, for the olive oil data with 80% known.

Southern Italy	Sardinia	Northern Italy
100.00	7.68	6.47

Table 5.22: Best model estimated mixing proportions for $G = 3$, for the olive oil data with 80% known.

Southern Italy	Sardinia	Northern Italy
0.5647	0.1712	0.2641

was run another time and the resulting model had a BIC value of -5683.820. This best model produced the classifications in Table 5.23 with a Rand index of 0.9815, an adjusted Rand index of 0.9355 and an error rate of 3.31%. Selected associated model parameters are given in tables 5.24 and 5.25.

Table 5.23: Estimated classifications for $G = 9$, $q = 5$ and UUU, for the olive oil data with 80% known.

[illegible]

Table 5.24: Best model estimated degrees of freedom for $G = 9$, for the olive oil data with 80% known.

North Apulia	Calabria	South Apulia	Sicily	Inland Sardinia
100.00	100.00	14.63	12.4	100.00
Coastal Saridinia	East Liguria	West Liguria	Umbria	
100.00	8.39	100.00	2.01	

Table 5.25: Best model estimated mixing proportions for $G = 9$, for the olive oil data with 80% known.

North Apulia	Calabria	South Apulia	Sicily	Inland Sardinia
0.0434	0.0994	0.3567	0.0632	0.1136
Coastal Saridinia	East Liguria	West Liguria	Umbria	
0.0577	0.0892	0.0874	0.0874	

5.4.2 Classification *via* LDA

Linear discriminant analysis was also performed using the subset of data that was treated as known in the final model from the MMtFAs method as the training set, and the remaining observations as the test set. For the three group structure this resulted in the classifications in Table 5.26 with a Rand index of 0.9768, and adjusted Rand index of 0.9529 and an error rate of 1.89%. For the nine group structure this resulted in the classifications in Table 5.27 with a Rand index of 0.9709, an adjusted Rand index of 0.8980 and an error rate of 7.44%.

Table 5.26: LDA Estimated classifications for the olive oil data with 80% known.

	1	2	3
Southern Italy	63		1
Sardinia		16	
Northern Italy		1	25

Table 5.27: LDA Estimated classifications for $G = 9$, for the olive oil data with 80% known.

	1	2	3	4	5	6	7	8	9
North Apulia	5								
Calabria		8		1					
South Apulia		1	41	1					
Sicily	3			6					
Inland Sardinia					16				1
Coastal Sardinia						8			
East Liguria							11		1
West Liguria								8	
Umbria							1		9

5.5 Proportion of Data Known: 90%

Classification via MMtFAs

Finally, the olive oil data was analyzed again specifying a random sample of 90% of the observations as being known, and the remaining observations as having unknown classifications. The algorithm was run five times with different times for $G = 3$, $q = 1, 2, \dots, 5$ and both the UUU and UCU covariance structures. The highest BIC was -5565.242 with $G = 3$, $q = 5$ and the UUU covariance structure. This best combination was run another time with different subsets of data being treated as having known classifications.

The BIC for this final model was -5563.565 and the associated model classifications are displayed in Table 5.28, with a Rand index of 1.000, an adjusted Rand index of 1.000 and an error rate of 0.00%. Selected model parameters are displayed in tables 5.29 and 5.30.

When considering the nine group structure, the highest BIC of the initial five runs was -5635.864 with $q = 4$ and the UUU covariance structure. The following

Table 5.28: Estimated classifications for $G = 3$, $q = 5$ and UUU, for the olive oil data with 90% known.

	1	2	3
Southern Italy	39		
Sardinia		7	
Northern Italy			15

Table 5.29: Best model estimated degrees of freedom, for the olive oil data with 90% known.

Southern Italy	Sardinia	Northern Italy
100.00	8.21	6.55

Table 5.30: Best model estimated mixing proportions, for the olive oil data with 90% known.

Southern Italy	Sardinia	Northern Italy
0.5647	0.1713	0.2640

run of this model had a BIC value of -5668.827 and produced the classifications in Table 5.31, as well as the parameters in tables 5.32 and 5.33. This model had a Rand index of 0.9800, and adjusted Rand index of 0.9466 and an error rate of 2.90%.

Table 5.31: Estimated classifications for $G = 9$, $q = 4$ and UUU, for the olive oil data with 90% known.

[illegible]

Table 5.32: Best model estimated degrees of freedom for $G = 9$, for the olive oil data with 90% known.

North Apulia	Calabria	South Apulia	Sicily	Inland Sardinia
100.00	100.00	13.87	8.86	100.00
Coastal Saridinia	East Liguria	West Liguria	Umbria	
100.00	6.00	100.00	1.70	

Table 5.33: Best model estimated mixing proportions for $G = 9$, for the olive oil data with 90% known.

North Apulia	Calabria	South Apulia	Sicily	Inland Sardinia
0.0437	0.0979	0.3585	0.0646	0.1119
Coastal Saridinia	East Liguria	West Liguria	Umbria	
0.0594	0.0874	0.0874	0.0892	

5.5.1 Classification via LDA

Linear discriminant analysis was also performed on the same data set, using the 90% sample of known observations from the final model as the training set, and the remaining unknown observations as the test set. When the three group structure was considered, the classifications had a Rand index of 0.9885, an adjusted Rand index of 0.9770 and an error rate of 1.64%. These classifications are displayed in Table 5.34. When considering the nine group structure, the classifications in Table 5.35. These classifications had a Rand index of 0.9834, an adjusted Rand index of 0.9559 and an error rate of 2.90%.

Table 5.34: LDA Estimated classifications for the olive oil data with 90% known.

	1	2	3
Southern Italy	39		
Sardinia		7	
Northern Italy		1	14

Table 5.35: LDA Estimated classifications for $G = 9$, for the olive oil data with 90% known.

	1	2	3	4	5	6	7	8	9
North Apulia	2								
Calabria		3							
South Apulia		1	31						
Sicily	1			3					
Inland Sardinia					12				
Coastal Sardinia						1			
East Liguria							6		
West Liguria								3	
Umbria									6

5.6 Method Comparisons

When considering the three group structure, for all four levels of known observations, $q = 5$ was consistently chosen as the best number of factors for the model. This makes sense when referring to Figure 5.1, the pairs plot of the variables. There are no strong linear relationships between any of the eight variables. Therefore it is logical that it would require multiple latent variables to account for most of the variability of the original eight. Summaries of the different classification techniques' abilities to correctly classify unknown observations are given in tables 5.36 and 5.37. Summaries of degrees of freedom and mixing proportion estimates are given in tables 5.38 and 5.39. For all proportions of observations treated as having known classifications, LDA had a lower correct classification rate, as well as lower Rand and adjusted Rand indices.

When considering the nine group structure, $q = 4$ or $q = 5$ was chosen as the best number of factors, which is consistent with the non-linear relationships between the eight variables visible in Figure 5.1. Summaries of the two techniques' classifica-

Table 5.36: Olive oil data MMtFAs summary for $G = 3$.

Proportion known	Rand index	Adjusted Rand index	Error rate (%)
60%	0.9962	0.9922	0.44
70%	1.0000	1.0000	0.00
80%	1.0000	1.0000	0.00
90%	1.0000	1.0000	0.00

Table 5.37: Olive oil data LDA summary for $G = 3$.

Proportion known	Rand index	Adjusted Rand index	Error rate (%)
60%	0.9888	0.9770	1.33
70%	0.9714	0.9415	3.01
80%	0.9768	0.9529	1.89
90%	0.9885	0.9770	1.64

Table 5.38: Estimated degrees of freedom summary for the olive oil data for $G = 3$.

Proportion known	Southern Italy	Sardinia	Northern Italy
60%	100.00	100.00	5.81
70%	100.00	7.63	6.57
80%	100.00	7.68	6.47
90%	100.00	8.21	6.55

Table 5.39: Mixing proportions summary for the olive oil data for $G = 3$.

	Southern Italy	Sardinia	Northern Italy
True	0.5647	0.1713	0.2640
60% known	0.5647	0.1691	0.2662
70% known	0.5647	0.1712	0.2641
80% known	0.5647	0.1712	0.2641
90% known	0.5647	0.1713	0.2640

tion abilities are given in tables 5.40 and 5.41. Summaries of the estimated degrees of freedom and mixing proportions are given in tables 5.42 and 5.43. In the 60%, 70%, and 80% cases the MMtFA model had higher Rand and adjusted Rand indices as well as lower error rates. In the 90% the MMtFA and LDA techniques misclassified

Table 5.40: Olive oil data MMFtAs summary for $G = 9$.

Proportion known	Rand index	Adjusted Rand index	Error rate (%)
60%	0.9814	0.9360	3.76
70%	0.9878	0.9568	2.79
80%	0.9815	0.9355	3.31
90%	0.9800	0.9466	2.90

Table 5.41: Olive oil data LDA summary for $G = 9$.

Proportion known	Rand index	Adjusted Rand index	Error rate (%)
60%	0.9806	0.9332	6.57
70%	0.9800	0.9302	5.59
80%	0.9709	0.8980	7.44
90%	0.9834	0.9559	2.90

the same number of observations and therefore had the same error rate. The specific observations that were misclassified were slightly different for the two techniques and it is for this reason that the MMtFA classifications had slightly lower Rand and adjusted Rand indices than the LDA classifications.

Table 5.42: Estimated degrees of freedom summary for the olive oil data for $G = 9$.

Proportion known	North Apulia	Calabria	South Apulia
60%	100.00	100.00	13.06
70%	100.00	100.00	13.31
80%	100.00	100.00	14.63
90%	100.00	100.00	13.87
Proportion known	Sicily	Inland Sardinia	Coastal Saridinia
60%	7.38	15.62	100.00
70%	7.02	100.00	100.00
80%	12.4	100.00	100.00
90%	8.86	100.00	100.00
Proportion known	East Liguria	West Liguria	Umbria
60%	5.87	100.00	1.92
70%	6.01	6.98	2.00
80%	8.39	100.00	2.01
90%	6.00	100.00	1.70

Table 5.43: Mixing proportions summary for the olive oil data for $G = 9$.

	North Apulia	Calabria	South Apulia
True	0.0437	0.0979	0.3585
60% known	0.0436	0.1039	0.3596
70% known	0.0435	0.0958	0.3623
80% known	0.0434	0.0994	0.3567
90% known	0.0437	0.0979	0.3585
	Sicily	Inland Sardinia	Coastal Sardinia
True	0.0646	0.1119	0.0594
60% known	0.0576	0.1136	0.0577
70% known	0.0631	0.1119	0.0594
80% known	0.0632	0.1136	0.0577
90% known	0.0646	0.1119	0.0594
	East Liguria	West Liguria	Umbria
True	0.0874	0.0874	0.0892
60% known	0.0892	0.0874	0.0874
70% known	0.0910	0.0873	0.0857
80% known	0.0892	0.0874	0.0874
90% known	0.0874	0.0874	0.0892

In all four levels of known data, for both the three group and nine group structures there were one or more groups which had low estimated degrees of freedom. This indicates that, though some of the groups which had high degrees of freedom could be reasonably modelled with Gaussian distributions, there are some groups for which the t -distribution would be much more appropriate. For this reason, MMtFAs model is more appropriate for modelling this data set than a mixture of multivariate Gaussian distributions.

Also, the UUU covariance structure was chosen over the UCU covariance structure at all four levels of known data when considering both the three and nine group structures. This indicates that there is substantial variability in noise matrices between groups for this data set and that a common noise matrix does not reasonably model the olive oil data.

Chapter 6

Conclusions

6.1 Discussion

For both the Italian wines and Italian olive oil data sets, the MMtFA method outperformed linear discriminant analysis for virtually all analyzed proportions of known data. In one case the MMtFA method misclassified the same number of observations as LDA. In general, it also appeared that for both data sets the MMtFA method produced better estimated classifications at higher levels of known data. However, LDA's ability to correctly classify unknown observations did not appear to depend as much on the proportion of data that was considered known, and there was no increasing trends in the Rand or adjusted Rand indices, and no decreasing trend in error rate as the proportion of known data increased. From these data the MMtFA method was the better technique out of the two methods tested when classifying unknown observations from multiple groups.

Though the UUU structure was consistently chosen as the better of the two covariance structures considered in these analyses (UUU and UCU), this does not indicate that it will be chosen as best for all data sets. Whether the UCU structure is chosen as best out of the possible options depends on the actual structure of the data. For this reason both covariance structures are viable options for other data sets and should both be considered.

The MMtFA model allows for modelling of both t -distributed data as well as approximately Gaussian distributed data because it can both model very low and very high degrees of freedom. This results in a distinct advantage over mixtures of Gaussian distributions and allows this technique

to model a wide range of data sets.

6.2 Future Work

There are a lot of opportunities for further work in this area. Currently, only one constraint has been placed on the components of the covariance structure. Combinations of various constraints could lead to a family of mixtures of multivariate t -factor analyzers, whose members are displayed in Table 6.1. The constrained loading matrix implies that the loading matrix would be equal across all groups, such that $\mathbf{\Lambda}_g = \mathbf{\Lambda}$. The constrained error variance implies that the noise matrix would be equal across all groups, such that $\mathbf{\Psi}_g = \mathbf{\Psi}$, and the isotropic constraint indicates that the noise matrix would be isotropic such that $\mathbf{\Psi}_g = \psi_g \mathbf{I}_p$. These constraints are the same as those applied to the covariance structure of PGMMs.

This new family of mixture models would most likely provide better and more efficient methods for modeling the above data sets, as well as others with distinct group structures. Hence, exploring this area further would lead to new and interesting results, and the ideas could be applied to a wide variety of research areas.

Table 6.1: Possible covariance structures derived from the mixtures of factor analyzers model along with the number of covariance parameters.

Model ID	Loading Matrix	Error Variance	Isotropic	Covariance Structure
CCC	Constrained	Constrained	Constrained	$\Sigma_g = \Lambda\Lambda' + \psi\mathbf{I}_p$
CCU	Constrained	Constrained	Unconstrained	$\Sigma_g = \Lambda\Lambda' + \Psi$
CUC	Constrained	Unconstrained	Constrained	$\Sigma_g = \Lambda\Lambda' + \psi_g\mathbf{I}_p$
CUU	Constrained	Unconstrained	Unconstrained	$\Sigma_g = \Lambda\Lambda' + \Psi_g$
UCC	Unconstrained	Constrained	Constrained	$\Sigma_g = \Lambda_g\Lambda_g' + \psi\mathbf{I}_p$
UCU	Unconstrained	Constrained	Unconstrained	$\Sigma_g = \Lambda_g\Lambda_g' + \Psi$
UUC	Unconstrained	Unconstrained	Constrained	$\Sigma_g = \Lambda_g\Lambda_g' + \psi_g\mathbf{I}_p$
UUU	Unconstrained	Unconstrained	Unconstrained	$\Sigma_g = \Lambda_g\Lambda_g' + \Psi_g$

Bibliography

- Banfield, J. D. and A. E. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49(3), 803–821.
- Bensmail, H. and G. Celeux (1996). Regularize discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association* 91(436), 1743–1748.
- Biernacki, C., G. Celeux, and G. Govaert (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 G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern Recognition* 28, 781–793.
- Dean, N., T. B. Murphy, and G. Downey (2006). Using unlabelled data to update classification rules with applications in food authenticity studies. *Journal of the Royal Statistical Society. Series C* 55(1), 1–14.
- Dean, N. and A. E. Raftery (2006). *The clustvarsel Package*. R package version 0.2-4.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B* 39(1), 1–38.
- Fisher, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7, 179–88.
- Forina, M., C. Armanino, M. Castino, and M. Ubigli (1986). Multivariate data analysis as a discriminating method of the origin of wines. *Vitis* 25, 189–201.
- Forina, M., C. Armanino, S. Lanteri, and E. Tiscornia (1983). Classification of olive oils from their fatty acid composition. In H. Martens and H. Russwurm Jr. (Eds.), *Food Research and Data Analysis*, pp. 189–214. London: Applied Science Publishers.
- Fraley, C. and A. E. Raftery (1998). How many clusters? Which clustering methods? Answers via model-based cluster analysis. *The Computer Journal* 41(8), 578–588.
- Fraley, C. and A. E. Raftery (1999). MCLUST: Software for model-based cluster analysis. *Journal of Classification* 16, 297–306.
- Fraley, C. and A. E. Raftery (2002a, October). MCLUST: Software for model-based clustering, density estimation, and discriminant analysis. Technical Report 415, University of Washington, Department of Statistics.

- Fraley, C. and A. E. Raftery (2002b). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97(458), 611–631.
- Fraley, C. and A. E. Raftery (2003). Enhanced software for model-based clustering, density estimation, and discriminant analysis: MCLUST. *Journal of Classification* 20, 263–286.
- Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association* 84(405), 165–175.
- Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer.
- Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition* (2nd ed.). Boston: Academic Press.
- Ghahramani, Z. and G. E. Hinton (1997). The EM algorithm for factor analyzers. Technical Report CRG-TR-96-1, University Of Toronto, Toronto.
- Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification* 2, 193–218.
- Hurley, C. (2004). Clustering visualizations of multivariate data. *Journal of Computational and Graphical Statistics* 13(4), 788–806.
- Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association* 90, 773–795.
- Keribin, C. (1998). Estimation consistante de l'ordre de modèles de mélange. *Comptes Rendus de l'Académie des Sciences. Série I. Mathématique* 326(2), 243–248.
- Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā. The Indian Journal of Statistics. Series A* 62(1), 49–66.
- Leroux, B. G. (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics* 20, 1350–1360.
- Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, Volume 5. California: Institute of Mathematical Statistics: Hayward.
- McLachlan, G. and S. Ganesalingam (1982). Updating a discriminant function on the basis of unclassified data. *Communications in Statistics - Simulation and Computation* 11, 753–767.
- McLachlan, G. J. and K. E. Basford (1988). *Mixture Models: Inference and applications to clustering*. New York: Marcel Dekker Inc.

- McLachlan, G. J., R. W. Bean, and L. Ben-Tovim Jones (2007). Extension of the mixture of factor analyzers model to incorporate the multivariate t -distribution. *Computational Statistics & Data Analysis* 51(11), 5327–5338.
- McLachlan, G. J. and T. Krishnan (2008). *The EM Algorithm and Extensions* (2nd ed.). New York: Wiley.
- McLachlan, G. J. and D. Peel (1998). Robust cluster analysis via mixtures of multivariate t -distributions. In *Lecture Notes in Computer Science*, Volume 1451, pp. 658–666. Berlin: Springer-Verlag.
- McLachlan, G. J. and D. Peel (2000). *Finite Mixture Models*. New York: John Wiley & Sons.
- McLachlan, G. J., D. Peel, and R. W. Bean (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis* 41(3–4), 379–388.
- McNicholas, P. D. and T. B. Murphy (2008). Parsimonious Gaussian mixture models. *Statistics and Computing* 18(3), 285–296.
- McNicholas, P. D., T. B. Murphy, A. F. McDaid, and D. Frost (2009). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis*. To appear.
- Meng, X. L. and D. B. Rubin (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 80, 267–278.
- Meng, X. L. and van Dyk (1997). The EM algorithm — an old folk song sung to a fast new tune (with discussion). *Journal of the Royal Statistical Society Series B* 59, 511–567.
- R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Raftery, A. E. and N. Dean (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association* 101(473), 168–178.
- Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66, 846–850.
- Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 31–38.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology* 15, 72–101.
- Streuli, H. (1973). Der heutige stand der kaffeechemie. In *Association Scientifique International du Cafe, 6th International Colloquium on Coffee Chemistry*, Bogotá, Columbia, pp. 61–72.

- Swayne, D., D. Cook, A. Buja, D. Lang, H. Wickham, and M. Lawrence (2006). *GGobi Manual*.
- Tipping, T. E. and C. M. Bishop (1999). Mixtures of probabilistic principal component analysers. *Neural Computation* 11(2), 443–482.
- Titterton, D. (1976). Updating a diagnostic system using unconfirmed cases. *Applied Statistics* 25, 238–247.
- Titterton, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: John Wiley & Sons.
- Toher, D., G. Downey, and T. B. Murphy (2007). A comparison of model-based and regression classification techniques applied to near infrared spectroscopic data in food authentication studies. *Chemometrics and Intelligent Laboratory Systems* 89, 102–115.
- Venables, W. N. and B. D. Ripley (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer.
- Wolfe, J. H. (1963). Object cluster analysis of social areas. Master's thesis, University of California, Berkeley.