Discriminant Analysis for Longitudinal Data

DISCRIMINANT ANALYSIS FOR LONGITUDINAL DATA

BY

KEVIN MATIRA, B.Sc.

A THESIS

SUBMITTED TO THE DEPARTMENT OF MATHEMATICS & STATISTICS

AND THE SCHOOL OF GRADUATE STUDIES

OF MCMASTER UNIVERSITY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

Master of Science (2017) McMaster University

(Statistics) Hamilton, Ontario, Canada

TITLE:	Discriminant Analysis for Longitudinal Data
AUTHOR:	Kevin Matira B.Sc., (Mathematics & Statistics) McMaster University, Hamilton, Ontario, Canada
SUPERVISOR:	Dr. Paul D. McNicholas
NUMBER OF PAGES:	ix, 43

To my mother and father for their endless support throughout my university career.

Abstract

Various approaches for discriminant analysis of longitudinal data are investigated, with some focus on model-based approaches. The latter are typically based on the modified Cholesky decomposition of the covariance matrix in a Gaussian mixture; however, non-Gaussian mixtures are also considered. Where applicable, the Bayesian information criterion is used to select the number of components per class. The various approaches are demonstrated on real and simulated data.

Acknowledgements

I would like to thank my supervisor, Dr. Paul D. McNicholas, for his constant assistance and patience with me throughout my undergraduate and graduate studies here at McMaster University. He was always responsive whenever I ran into a problem or had a question about my research. This accomplishment would not have been possible without him.

This work is supported by an NSERC Discovery Grant (McNicholas) and the Canada Research Chairs program (McNicholas).

Contents

List of Figures

Chapter 1

Introduction

There are different types of learning in cluster analysis that include: unsupervised, semi-supervised, and supervised. The amount of supervision refers to how many labelled observations there are and how many are used. Unsupervised learning is when no labels are given to classify the data, which is essentially referred to as clustering. The other two types of learning have some labelled observations that are used to infer labels for the unlabelled observations. Consider n observations. Semi-supervised learning has k labeled and $(n-k)$ unlabeled points, and the goal is to correctly label the $(n-k)$ unlabeled points using all n points. Supervised learning refers to using the k labelled observations to build a rule that is used to label the remaining $(n-k)$ observations; this is also called discriminant analysis and the rule is called a discriminant rule. The key distinction is semi-supervised learning uses all n points whereas supervised learning uses only the k labelled points to infer labels for the unlabelled points. Often, the k labelled points are referred to as the training set and the $n-k$ unlabeled points are referred to as the test set. An analogy is to think of having two clusters on either side of a line of best fit, see Figure 1.1. Note that

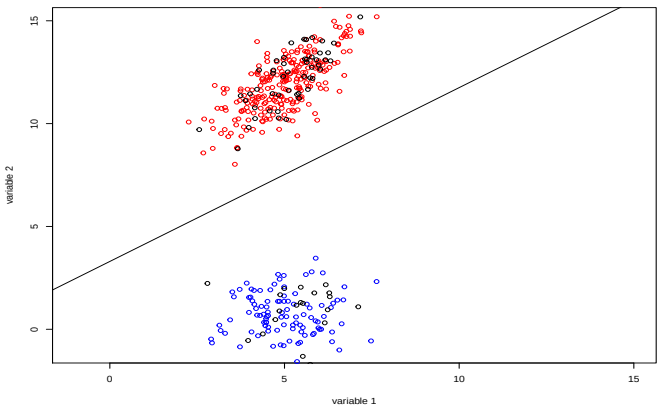


Figure 1.1: Example of discriminant analysis with cluster one in red and cluster two in blue where the discriminant rule is the line of best fit.

a line of best fit is a straight line that accurately represents the data on a scatter plot, i.e., a line is drawn through the center of a group of data points. The line of best fit may pass through some of the points, all of the points, or none of the points. Consider cluster one above the line of best fit and cluster two below the line of best fit. The line of best fit is built using the training set observations, i.e., the red and blue points in Figure 1.1. To classify the unlabelled points (black points), the line of best fit plays an important role. In Figure 1.1, a black point above the line of best fit will be classified as cluster one and an observation below the line of best fit will be classified as cluster two. It is important to note that the line of best fit does not change because it has already been trained by the labelled points.

The focus herein will be on supervised learning for longitudinal data. Longitudinal data, sometimes referred to as panel data, tracks the same observation at different points in time. For example, to assess the effectiveness of diets on rats, one may observe their body weight every week (Crowder and Hand, 1990). Another example is observing gene expressions over time—this is referred to as gene expression time

course data (Arbeitman *et al.*, 2002; Chu *et al.*, 1998). There are also studies tracking the weight loss of individuals in different weight loss programs. One study in Italy tracks daily household power consumption. These two studies, along with simulated data will be presented.

The concept of finite mixture models will be introduced, along with a special covariance decomposition to account for longitudinal data (McNicholas and Murphy, 2010a). The decomposition for the covariance matrix will be a modified Cholesky decomposition. A family of mixture models arises from this covariance decomposition. Discriminant analysis will be performed using a Gaussian mixture model with a modified Cholesky decomposition of the covariance matrix and a mixture of multivariate t-distributions McNicholas and Subedi (2012). After estimating the parameters and selecting the model, the classification performance will be assessed.

Chapter 2

Methodology

2.1 Modified Cholesky Decomposition

Let $\mathbf{A}$ represent any real positive definite matrix. The Cholesky decomposition (Benoît, 1924) of $\mathbf{A}$ is

$$\mathbf{A} = \mathbf{L}\mathbf{L}',$$

where $\mathbf{L}$ is a unique lower triangular matrix. This decomposition is commonly used in numerical analysis applications due to fast computation times and simplicity because the solution simplifies to a system of linear equations. Let $\mathbf{\Sigma}$ represent the covariance matrix of a random variable. One can apply a modified Cholesky decomposition to obtain the following:

$$\mathbf{T}\mathbf{\Sigma}\mathbf{T}' = \mathbf{D} \Leftrightarrow (\mathbf{T}\mathbf{\Sigma}\mathbf{T}')^{-1} = \mathbf{D}^{-1} \Leftrightarrow \mathbf{T}'^{-1}\mathbf{\Sigma}^{-1}\mathbf{T}^{-1} = \mathbf{D}^{-1},$$

where $\mathbf{T}$ is a unique unit lower triangular matrix and $\mathbf{D}$ is a unique diagonal matrix with strictly positive diagonal entries. Note that a unit lower triangular matrix is a

lower triangular matrix with ones along the diagonal. Finally, taking the inverse of both sides yields the covariance matrix:

$$\boldsymbol{\Sigma}^{-1} = \mathbf{T}'\mathbf{D}^{-1}\mathbf{T} \Leftrightarrow \boldsymbol{\Sigma} = (\mathbf{T}'\mathbf{D}^{-1}\mathbf{T})^{-1}, \tag{2.1}$$

due to matrix properties and the steps given above. The values of $\mathbf{T}$ and $\mathbf{D}$ can be interpreted as generalized autoregressive parameters and innovation variances, respectively (Pourahmadi, 1999). The linear least-squares predictor of X_t based on $X_{t-1}, \ldots, X_1$, is given by

$$\hat{X}_t = \mu_t + \sum_{s=1}^{t-1}(-\varphi_{ts})(X_s - \mu_s) + \sqrt{d_t}\varepsilon_t,$$

where φ_{ts} is the sub-diagonal element of $\mathbf{T}$ in position (t, s), d_t is the tth diagonal element of $\mathbf{D}$, and $\varepsilon_t \sim N(0, 1)$. In the past, the modified Cholesky decomposition has been used for joint modelling of both the mean and covariance in longitudinal studies (Pan and Mackenzie, 2003). Similarly, an approach was developed for simultaneously modelling several covariance matrices via the modified Cholesky decomposition (Pourahmadi *et al.*, 2007).

2.2 Finite Mixture Models

A finite mixture model is a convex linear combination of a finite number of probability distributions. Finite mixture models are used in cluster analysis, and commonly, a cluster is taken to equal one component in the mixture (see McNicholas, 2016a,b). Let $\mathbf{X}$ represent a random variable and let p represent the data dimensionality. The

probability density function of a mixture model is:

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

where π_g are the mixing proportions, with $\pi_g > 0$ and $\sum_{g=1}^{G} \pi_g = 1$, and $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ is the probability density function of the gth component with parameters $\boldsymbol{\vartheta} = (\pi_1, \ldots, \pi_G, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)$. It should be noted that $f(\mathbf{x} \mid \boldsymbol{\vartheta})$ is often referred to as a G-component finite mixture density. Usually, for convenience the component densities have the same distribution, e.g., $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ are all Gaussian densities.

2.3 Mixtures of Multivariate Gaussian Distributions

The Gaussian mixture model is a very common choice in literature because of ease of manipulation. The density of a Gaussian mixture model can be written as:

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

where

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

is the probability density function of a multivariate Gaussian distribution where $\boldsymbol{\mu}_g$ is the mean and $\boldsymbol{\Sigma}_g$ is the covariance matrix. Note that, in McNicholas and Murphy

(2010a), the component precision matrices utilize the decomposition in (2.1), i.e.,

$$\boldsymbol{\Sigma}_g = (\mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g)^{-1},$$

where $\mathbf{D}_g$ and $\mathbf{T}_g$ are the diagonal matrix and unit lower triangular matrix, respectively, that follow from the modified Cholesky decomposition.

A total of eight Gaussian mixture models arise depending on certain constraints, see Table 2.1. One can allow $\mathbf{T}_g$ and/or $\mathbf{D}_g$ to be the same across all components along with the isotropic constraint $\mathbf{D}_g = \delta_g \mathbf{I}_p$. Following McNicholas (2016a), this family of eight will be referred to as the Cholesky-decomposed Gaussian mixture models (CDGMMs). The CDGMMs fit longitudinal data very naturally. For example, constraining $\mathbf{T}_g = \mathbf{T}$ yields that the autoregressive relationship between time points is the same across all components, i.e., the correlation structure of the longitudinally recorded data values is the same for all classes. The constraint $\mathbf{D}_g = \mathbf{D}$ yields that the variability at each time point is the same for each of the components. Lastly, the isotropic constraint $\mathbf{D}_g = \delta_g \mathbf{I}_g$ yields that the variability is the same at each time point within the component, i.e., the noise is the same at all time points.

Table 2.1: The different constraints on the covariance matrix along with the number of free covariance parameters for each member of the CDGMM family.

Model	$\mathbf{T}_g$	$\mathbf{D}_g$	$\mathbf{D}_g$	Free Covariance Parameters
EEA	Equal	Equal	Anisotropic	$p(p-1)/2+p$
VVA	Variable	Variable	Anisotropic	$G[p(p-1)/2]+Gp$
VEA	Variable	Equal	Anisotropic	$G[p(p-1)/2]+p$
EVA	Equal	Variable	Anisotropic	$p(p-1)/2+Gp$
VVI	Variable	Variable	Isotropic	$G[p(p-1)/2]+G$
VEI	Variable	Equal	Isotropic	$G[p(p-1)/2]+1$
EVI	Equal	Variable	Isotropic	$p(p-1)/2+G$
EEI	Equal	Equal	Isotropic	$p(p-1)/2+1$

2.4 Mixtures of Multivariate *t*-distributions

The computational convenience of mixtures of Gaussian distributions explained the attention received both in literature and in applications. In practice, one may require a less heavy tailed distribution. McLachan and Peel (1998) first look at the Gaussian scale mixture model

$$(1-\epsilon)\phi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \epsilon\phi(\mathbf{x}_i \mid \boldsymbol{\mu}, c\boldsymbol{\Sigma}), \tag{2.2}$$

where c is large and ϵ is small. Alternatively, (2.2) can be written as

$$\int \phi(\mathbf{x}_i \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}/w_{ig}) dH(w_{ig}),$$

where W_{ig} represents a latent variable such that

$$\mathbf{X}_i \mid w_{ig}, z_{ig} = 1 \sim N(\boldsymbol{\mu}_g, (\mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g)^{-1} / w_{ig})$$

and $W_{ig} \mid z_{ig} = 1$ follows a gamma distribution with parameters $(\nu_g/2, \nu_g/2)$, where ν_g denotes the degrees of freedom of the gth component (McLachlan and Peel, 1998). The gth multivariate Student's t component density can then be defined as

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

where $\boldsymbol{\mu}_g$ is the mean, $\boldsymbol{\Sigma}_g = (\mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g)^{-1}$ is the scale matrix, ν_g is the degrees of freedom of component g, Γ is the gamma function, and $\delta(\mathbf{x}, \boldsymbol{\mu}_g \mid \boldsymbol{\Sigma}_g)$ is the Mahalanobis distance, i.e.,

$$\delta(\mathbf{x}, \boldsymbol{\mu}_g \mid \boldsymbol{\Sigma}_g) = (\mathbf{x} - \boldsymbol{\mu}_g)' \boldsymbol{\Sigma}_g (\mathbf{x} - \boldsymbol{\mu}_g).$$

Compared to the Gaussian distribution, the multivariate t-distribution simply introduces a new parameter known as the degrees of freedom. Note that a t-distribution is equivalent to a Gaussian distribution for a high degrees of freedom (Student, 1908). Following (McNicholas, 2016a), this family will be referred to as the CDtMM family.

2.5 Likelihood

Let z_{ig} denote an indicator variable such that $z_{ig} = 1$ if observation i belongs to component g and $z_{ig} = 0$ otherwise. Let $\mathbf{z}_i = (z_{i1}, z_{i2}, \ldots, z_{iG})$ represent the component membership of observation i. The mixture discriminant analysis likelihood for the

labelled observations $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ and their associated labels $(\mathbf{z}_1, \ldots, \mathbf{z}_k)$ is:

$$\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}_1, \ldots, \mathbf{x}_k) = \prod_{i=1}^{k} \prod_{g=1}^{G} [\pi_g f_g(\mathbf{x}_i \mid \boldsymbol{\theta}_g)]^{z_{ig}}, \tag{2.3}$$

where π_g are the mixing proportions, $f_g(\mathbf{x} \mid \boldsymbol{\theta}_g)$ is the pdf of the gth component, and $\boldsymbol{\theta}_g$ is the vector of parameters. Taking the logarithm of both sides of (2.3) yields the log-likelihood

$$l(\boldsymbol{\vartheta} \mid \mathbf{x}_1, \ldots, \mathbf{x}_k) = \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} [\log \pi_g + \log f_g(\mathbf{x}_i \mid \boldsymbol{\theta}_g)].$$

Setting the components to be Gaussian yields the likelihood

$$\mathcal{L}(\boldsymbol{\vartheta} \mid \mathbf{x}_1, \ldots, \mathbf{x}_k) = \prod_{i=1}^{k} \prod_{g=1}^{G} [\pi_g \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, (\mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g)^{-1})]^{z_{ig}},$$

where the log-likelihood is

$$l(\boldsymbol{\vartheta} \mid \mathbf{x}_1, \ldots, \mathbf{x}_k) = \sum_{i=1}^{k} \sum_{g=1}^{G} z_{ig} [\log \pi_g + \log \phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g, (\mathbf{T}_g' \mathbf{D}_g^{-1} \mathbf{T}_g)^{-1})].$$

Note that these likelihood functions assumes that each known class is modelled by one component. Using the notation of McNicholas (2016a), this constraint can be relaxed by replacing G with $\mathcal{G}$, where $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2 + \cdots + \mathcal{G}_G$. Essentially, model-based clustering is being performed on each known class to obtain the number of components per class $(\mathcal{G}_1, \ldots, \mathcal{G}_G)$ via some criterion. That is, each known class is being modelled by a mixture model itself. Then, a mixture model is fit with $\mathcal{G}$ components.

2.6 Linear Combination for Group Means

A nice feature about using the CDGMM and CDtMM families is that the means $\boldsymbol{\mu}_g$ can be modelled using a linear combination (McNicholas and Subedi, 2012), i.e.,

$$\boldsymbol{\mu}_g = \mathbf{Q}\boldsymbol{\beta}_g = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ t_1 & t_2 & t_3 & \dots & t_p \end{bmatrix}' \begin{bmatrix} a_g \\ b_g \end{bmatrix},$$

where the slope is b_g and the intercept is a_g. Given this parameterization, the likelihood can now be written as:

$$\mathcal{L}(\boldsymbol{\vartheta}) = \prod_{i=1}^{k}\prod_{g=1}^{G}[\pi_g \phi(\mathbf{x}_i \mid \mathbf{Q}\boldsymbol{\beta}_g, (\mathbf{T}_g'\mathbf{D}_g^{-1}\mathbf{T}_g)^{-1})]^{z_{ig}}.$$

This method implies that the mean will be modelled using a line of best fit. Similarly, one can model the mean using a quadratic, cubic, or other polynomial combination. In general, allowing for a linear mean is often undesirable because the data will not necessarily follow a linear trend. Examples of non-linear longitudinal data include the study of physical growth of boys and girls (Tuddenham and Snyder, 1954) or studying gene expression profiles (Arbeitman *et al.*, 2002; Chu *et al.*, 1998). Linear means will be included in one simulation and one real data set for completeness.

2.7 Parameter Estimation

The expectation-maximization (EM) algorithm (Dempster *et al.*, 1977) is commonly used in model parameter estimation. There are two steps to the algorithm. The first step is to compute the expected value of the complete data log-likelihood and is

known as the expectation step (E-step). To clarify, the complete-data in this context consist of the training set observations that belong to group $\mathcal{G}_g$. The second step is called the maximization step (M-step) and yields parameter updates that give a maximum of the log-likelihood from the E-step. This process iterates between the two steps until convergence. The objective is to obtain labels for the training set that are in class $\mathcal{G}_g$. Once labels are obtained for $\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_G$, a $\mathcal{G}$-component mixture model with these labels is fitted. Upon obtaining maximum likelihood estimates from the log-likelihood function, the resulting parameter estimates are recorded. Note that for $\mathcal{G} = 1$, it is essentially a one-component mixture.

2.8 Convergence Criterion

There are many different stopping criterion for the EM algorithm. One approach is to halt an EM algorithm based on the lack of progress in the log-likelihood. More explicitly, one can stop the algorithm when

$$l^{(k+1)} - l^{(k)} < \epsilon, \tag{2.4}$$

where ϵ is a small value and $l^{(k)}$ is the log-likelihood at the kth iteration. Clearly, this may be a good stopping rule if the log-likelihood increases and comes to a plateau at the maximum likelihood estimate. For an example of a plateau, one can refer to Figure 2.1. The criterion in (2.4) is not always effective (McNicholas *et al.*, 2010). For example, consider log-likelihood values that look like a staircase with multiple steps or have multiple local maximums.

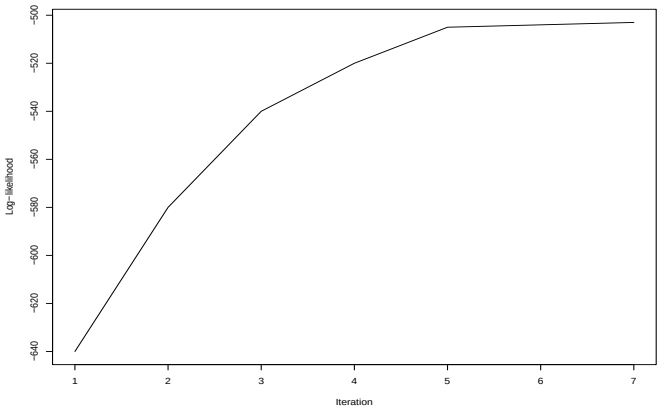


Figure 2.1: Plot of log-likelihood against EM algorithm iteration number for a real data set.

Following McNicholas *et al.* (2010), the alternative convergence criteria will be based on Aitken's acceleration (Aitken, 1926). More explicitly, Aitken's acceleration at iteration k can be written as:

$$a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}}.$$

The asymptotic estimate of the log-likelihood at the $k+1$ iteration is

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

The EM algorithm can be treated as converged when $|l_\infty^{(k+1)} - l_\infty^{(k)}| < \epsilon$ (Böhning *et al.*, 1994), or when $l_\infty^{(k)} - l^{(k)} < \epsilon$ (Lindsay, 1995), or when $l_\infty^{(k+1)} - l^{(k)} < \epsilon$, where the difference is positive (McNicholas *et al.*, 2010). The latter will be used to determine convergence because it is at least a strict as (2.4) (McNicholas *et al.*, 2010).

2.9 Model Selection

After all members of a family are fitted for a range of values of groups, the Bayesian information criterion (BIC; Schwartz, 1978) is used to select the best model and the number of components per class. The BIC can be written as:

$$\text{BIC} = 2l(\mathbf{x}, \hat{\boldsymbol{\vartheta}}) - \rho \log n,$$

where $\hat{\boldsymbol{\vartheta}}$ is the MLE of $\boldsymbol{\vartheta}$, ρ is the number of free parameters, and n is the number of observations. The BIC is used because under certain conditions, it is optimal for choosing the number of components in a mixture model (Leroux *et al.*, 1992). For the CDGMM and CDtMM families, the BIC is also used to select the decomposition. Keep in mind that while many model-based applications use the BIC, the best BIC does not necessarily yield the best classification performance (Bouchard and Celeux, 2004).

2.10 Classification Performance Assessment

One may be interested in seeing how well the model performs on the test set. This can be done by comparing the predicted test set labels to the true labels. Consider Table 2.2 below (Steinley, 2004).

Table 2.2: Table of pairs for two partitions.

	Same group	Different groups
Same group	A	B
Different groups	C	D

The Rand Index (RI) can be calculated using Table 2.2 (Rand, 1971). It is the ratio of pairwise agreements to the total number of pairs, i.e.,

$$\text{RI} = \frac{A + D}{A + B + C + D} = \frac{A + D}{N},$$

where N denotes the total number of pairs. Perfect class agreement results in RI = 1. A problem that arises is that the RI value may be higher than the true value due to agreement by chance. To compensate for this error, the adjusted Rand index (ARI; Hubert and Arabie, 1985) is introduced. For simplicity, using the notation in Table 2.2, the ARI can be written as:

$$\text{ARI} = \frac{N(A + D) - [(A + B)(A + C) + (C + D)(B + D)]}{N^2 - [(A + B)(A + C) + (C + D)(B + D)]}.$$

The ARI has an expected value of 0 under random pairwise agreement and is equal to 1 for perfect classification. One can see how this naturally arises by looking at the general form, i.e.,

$$\text{corrected index} = \frac{\text{index} - \text{expected index}}{\text{maximum index} - \text{expected index}}.$$

Negative ARI values can also occur and this implies that the classification performance is worse than would be expected by randomly classifying the observations. Classification performance is determined by the ARI of the test set. The ARI of the training set is omitted because the training set is labelled and used to build a discriminant rule.

2.11 Straightforward Discriminant Analysis

One can view straightforward discriminant analysis as follows: a data set is split into a training/test sets with 70%—80% belonging to the training set and the remaining (20%—30%) in the test set. The training/test split is usually stratified to ensure that all classes are accounted for. This is particularly useful when there are very few observations in one class. If there are G known classes in the training set, each known class is fit to exactly one component, i.e., one component per class. This implies that g ranges from 1 to G and the $\mathbf{z}_i$ for $i = 1, \ldots, k$ are known in the likelihood function. The maximum likelihood estimates are obtained via the log-likelihood. After obtaining the parameter estimates, one can compute the probability that an observation in the test set is part of one of the G classes. The ARI is used to assess the classification performance of the test set. It will be seen that straightforward discriminant analysis is often desirable. It is worth noting that there are more involved versions of straightforward discriminant analysis, e.g., with π_g treated as a prior probability (McIver and Friedl, 2002); however, these more involved versions will not be considered herein.

2.12 Mixture Discriminant Analysis

In mixture discriminant analysis, a data set is also split into a training test split that is usually stratified. Following a similar approach as in straightforward discriminant analysis, 70%—80% belong to the training set and the remaining 20%—30% are in the test set. Suppose there are G known classes in the training set. Each known class is fit to a mixture model and the labels are recorded via the EM algorithm. The number of

components for each class is obtained through the BIC and recorded $(\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_G)$. The total number of components, $\mathcal{G}$, is computed by $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2 + \cdots + \mathcal{G}_G$. A mixture model is then fit with $\mathcal{G}$ components along with the recorded labels to obtain parameter estimates. Parameter estimates are obtained via the maximum likelihood estimates of the log-likelihood function. The parameter estimates are used to compute the probability that an observation in the test set belongs to one of the $\mathcal{G}$ classes. Classification performance is determined by looking at the ARI value of the test set. Clearly, straightforward discriminant analysis is simply a special case of mixture discriminant analysis ($\mathcal{G}_g = 1$) and is often a desirable result.

2.13 Discriminant Rule

Let z_{jg} represent the probability that the jth observation in the test set belongs to the gth component. Using Bayes' theorem, this probability can be written as:

$$\hat{z}_{jg} = \frac{\hat{\pi}_g f_g(\mathbf{x}_j \mid \hat{\boldsymbol{\theta}}_g)}{\sum_{h=1}^{\mathcal{G}} \hat{\pi}_h f_g(\mathbf{x}_j \mid \hat{\boldsymbol{\theta}}_h)}, \tag{2.5}$$

where $f_g(\mathbf{x}_j \mid \hat{\boldsymbol{\theta}}_g)$ is the density of gth component. Note that all the predicted classifications are soft because they are all probabilities. For example, consider the scenario where $\mathcal{G} = 3$. The corresponding $\mathbf{z}_j$ for $j = 7$ could be $\mathbf{z}_7 = (0.3, 0.2, 0.5)$. This means that observation seven has a 30% chance of belonging to component one, a 20% chance of belonging to component two, and a 50% chance of belonging to component three. These soft classifications are considered to be a nice feature with mixture models because they allow one to evaluate borderline observations. This is particularly helpful when comparing borderline observations across different

supervised learning methods. For example, $\mathbf{z}_7 = (0.32, 0.31, 0.37)$ has very different connotations to $\mathbf{z}_7 = (0.05, 0.15, 0.8)$. In the first scenario, it is unclear whether or not observation seven truly belongs in class three, whereas it is more evident that observation seven belongs in class three in the second scenario. In general, one may wish to harden these probabilities for comparison in practical applications. A common method is using the maximum *a posteriori* (MAP) classifications, i.e., $\text{MAP}\{\hat{z}_{jg}\}$, where

$$\text{MAP}\{\hat{z}_{jg}\} = \begin{cases} 1 & \text{if } g = \arg\max_h\{\hat{z}_{jh}\}, \\ 0 & \text{otherwise.} \end{cases}$$

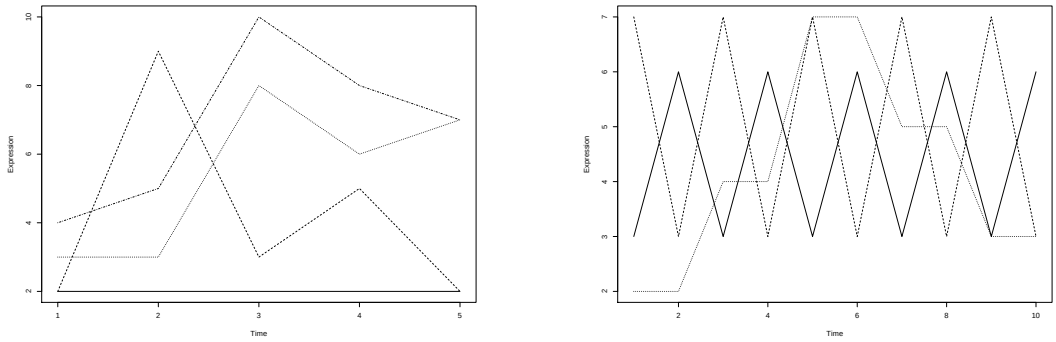
Chapter 3

Simulation Study

3.1 Introduction

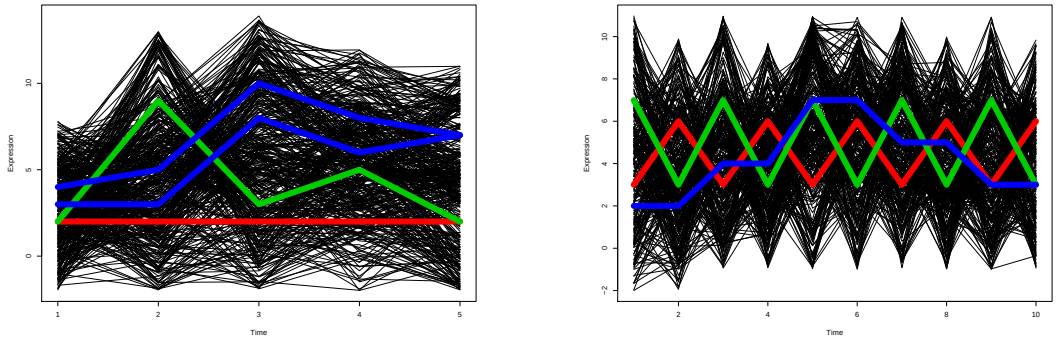
One method to simulate data is to make a representative time course for one or more classes (McNicholas, 2016a). In Figure 3.1, both graphs clearly have distinct time courses. The graph on the left has five time points whereas the graph on the right has 10 time points. Following Section 8.4 in McNicholas (2016a), to simulate more points, random values between $(-u, u)$ are added to the expression (y-value) at each time point. Increasing the value of u makes the clusters less distinguishable from one another whereas a small value for u keeps them fairly distinct. Repeating this process 99 times for each component and setting $u = 4.0$ yields the two graphs in Figure 3.2. This implies that there are now a total of 100 observations belonging to each time course. Treating this as a real problem, a training test split is applied to the simulated data. A 75/25 split is used, where 75% of the data will be part of the training set and 25% will be part of the test set. It should be noted that the results did not change whether stratification was used or not because each class is

represented by at least 100 observations.



(a) (b)

Figure 3.1: Four representative time courses on the left and three representative time courses on the right.



(a) (b)

Figure 3.2: Simulated longitudinal data coloured by classes with $u = 4.0$ and $n = 100$ for each representative time course.

3.2 Mixture Discriminant Analysis

3.2.1 First Simulation

Each known class in the training set in Figure 3.2a is fit using the CDGMM family. The BIC selected $\mathcal{G}_1 = \mathcal{G}_2 = 1$ and $\mathcal{G}_3 = 2$, see Figure 3.3. This results in the total number of components being $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2 + \mathcal{G}_3 = 4$ and the recorded labels ranging from one to four. Note that labels three and four correspond to belonging to class three. A four-component CDGMM with the corresponding labels are fit to get parameter estimates for $\boldsymbol{\mu}_g$ and $\boldsymbol{\Sigma}_g$. These parameter estimates are used in (2.5) to compute the soft classifications for the test set. The MAP classifications for 10 different test sets are given in Table 3.1. This corresponds to an average ARI value of 0.9037 over the course of 10 runs. The overall classification performance is very good, i.e., only 42 observations out of 1000 were misclassified.

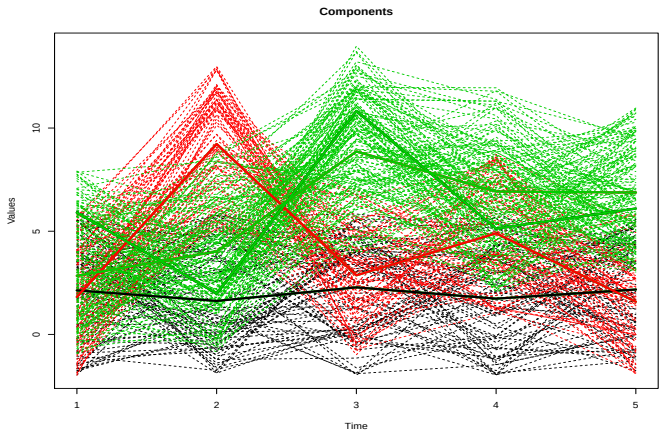


Figure 3.3: Plot of one training set using a CDGMM for the simulated data set with four distinct time courses and $n = 400$.

Table 3.1: Cross-tabulation of the MAP classifications associated with the 10 test sets using the CDGMM (A-C) against true classes for the first simulation.

	A	B	C
1	232	16	2
2	16	234	0
3	4	4	492

Similarly, each known class in the training set is also fit to the CDtMM family. The BIC selected $\mathcal{G}_1 = \mathcal{G}_2 = 1$ and $\mathcal{G}_3 = 2$. This can be seen in Figure 3.4. This implies that the total number of components is $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2 + \mathcal{G}_3 = 4$ and the recorded labels ranging from one to four. Note that labels three and four correspond to belonging to class three. A four-component CDtMM with the aforementioned labels are fit to obtain parameter estimates for $\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g$, and ν_g. To compute soft classifications for the test set, these parameter estimates will used in the discriminant rule in (2.5). The MAP classifications of 10 distinct test sets are in Table 3.2. The average ARI value is approximately 0.9001. The good classification results may be attributed to the high degrees of freedom for each component ($\nu_g \geq 100$). This implies that the fit is essentially a CDGMM.

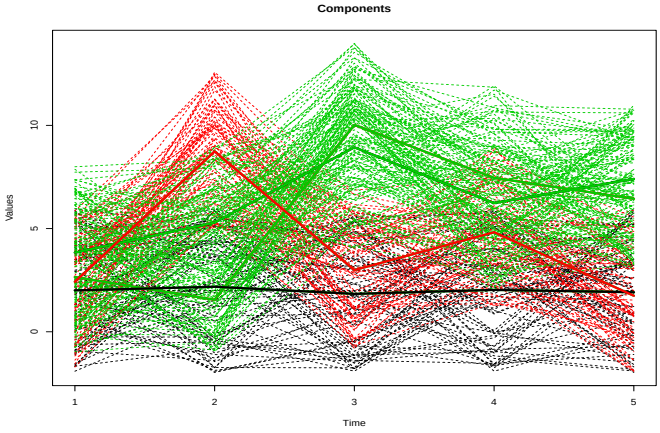


Figure 3.4: Plot of one training set using a CDtMM for the simulated data set with four distinct time courses and $n = 400$.

Table 3.2: Cross-tabulation of the MAP classifications associated with the 10 test sets using the CDtMM (A-C) against true classes for the first simulation.

	A	B	C
1	237	13	0
2	16	233	1
3	5	7	488

3.2.2 Second Simulation

The CDGMM family is fit for each known class in the training set in Figure 3.2b. This data set appears to be a more difficult problem. The BIC selected $\mathcal{G}_1 = \mathcal{G}_2 = \mathcal{G}_3 = 1$, see Figure 3.5. This yields that the total number of components is $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2 + \mathcal{G}_3 = 3$ and the recorded labels ranging from one to three. To obtain parameter estimates, a three-component CDGMM is fit with the corresponding labels. Parameter estimates are then used to compute soft classifications for 10 distinct test

sets and the MAP classifications are given in Table 3.3. The average ARI value over 10 runs is 0.8443. This is a relatively high ARI value with minimal classifications, i.e., only 40 observations out of 750 were misclassified.

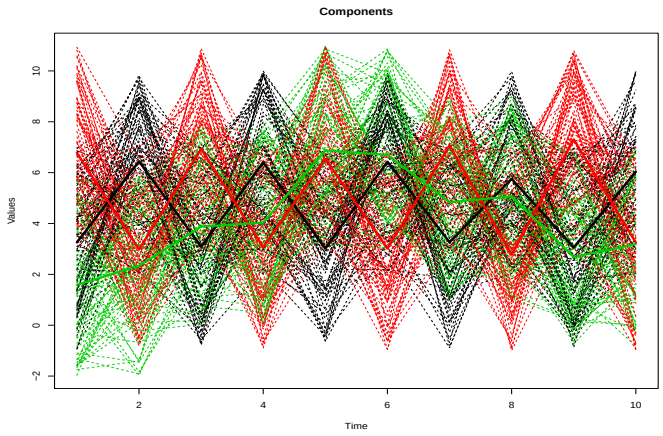


Figure 3.5: Plot of one training set using a CDGMM for the simulated data set with three distinct time courses and $n = 300$.

Table 3.3: Cross-tabulation of the MAP classifications associated with the 10 test sets using the CDGMM (A-C) against true classes for the second simulation.

	A	B	C
1	246	3	11
2	2	233	6
3	15	3	231

Allowing for linear means in the CDGMM family returns lower ARI values and poor classification results. The ARI values were approximately 0.1382 over 10 different runs for the CDGMM family. The hardened classifications of 10 different test sets are given in Table 3.4. This implies that the mean is better modelled by a weighted average than a line of best fit. One can see how this is true by looking at Figure 3.6.

It does not look natural to model the means by lines of best fit given the different patterns.

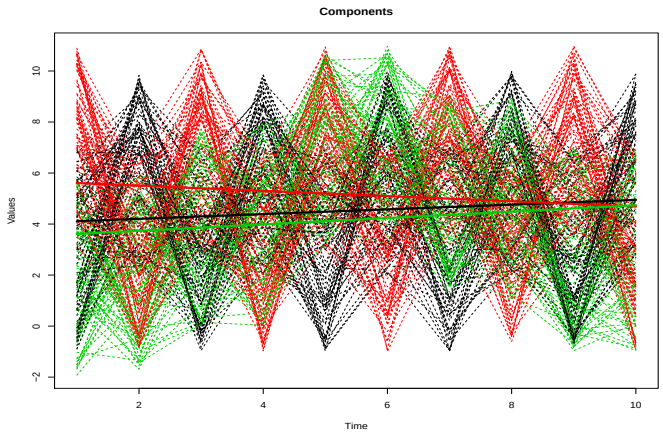


Figure 3.6: Plot of one training set using a CDGMM with linear means for the simulated data set with three distinct time courses and $n = 300$.

Table 3.4: Cross-tabulation of the MAP classifications associated with the 10 test sets using the CDGMM (A-C) with linear means against true classes for the second simulation.

	A	B	C
1	57	92	101
2	43	176	29
3	57	37	158

Next, the CDtMM family is also fit for each class in the training set in Figure 3.2b. The BIC selected $\mathcal{G}_1 = \mathcal{G}_2 = \mathcal{G}_3 = 1$. This implies that the total number of components is $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2 + \mathcal{G}_3 = 3$ and the recorded labels ranging from one to three. Parameter estimates are obtained by fitting a three-component CDtMM with the aforementioned labels. After obtaining parameter estimates, the soft classifications for the test set can

be calculated. The classification results for 10 different test sets is given in Table 3.5. The average ARI over the 10 runs is approximately 0.8327. It is important to note that the degrees of freedom was very high for all components, i.e., $\boldsymbol{\nu} = (\nu_1, \nu_2, \nu_3) = (200, 200, 200)$, which is the maximum in the software. This indicates that this model is essentially a CDGMM.

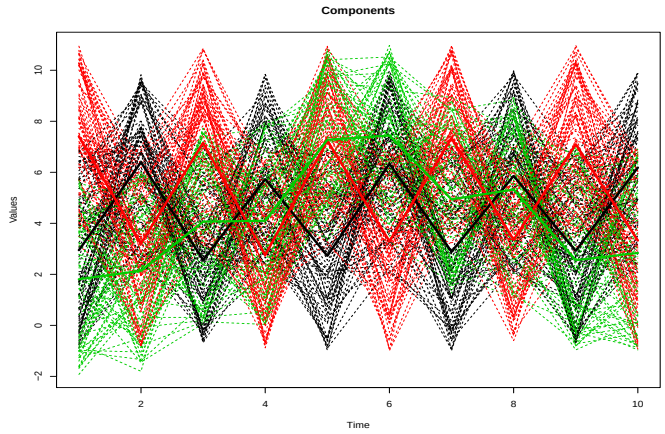


Figure 3.7: Plot of one training set using a CDtMM for the simulated data set with three distinct time courses and $n = 300$.

Table 3.5: Cross-tabulation of the MAP classifications associated with the 10 test sets using the CDtMM (A-C) against true classes for the second simulation.

	A	B	C
1	233	0	17
2	0	244	6
3	9	12	229

Chapter 4

Real Data Analyses

4.1 Introduction

In real data analyses, one may want to perform discriminant analysis on gene expression time course data to find groups of genes with similar expression patterns. Expression patterns are often referred to as expression profiles. Co-expressed genes are genes that have similar expression profiles. Practical applications include being interested in finding groups of genes that have similar activation patterns over time. For example, a study was conducted to look at the behaviour of yeast genes during sporulation (Chu *et al.*, 1998). A problem in this study was that only 40 genes of over 5000 genes have known labels. Labels are biological functions in this case.

4.2 Weight Loss Data Set

The first real data set used for illustration is the weight loss data set (Dominici, 2005). There are a total of 100 observations and three different weight loss programs. More

explicitly, there are 34 individuals in program one, 28 participants in program two, and the remaining 38 individuals are in program three. There are five time points evenly spread over 12 months. A plot of the data is given below in Figure 4.1. The three classes appear to have distinct paths. Again, the training set will consist of 75% of the data and the remaining 25% will be in the test set. Simply put, 75 observations will be in the training set and 25 observations will be a part of the test set. The data were stratified to ensure a suitable number of points from each class were allocated to the test set.

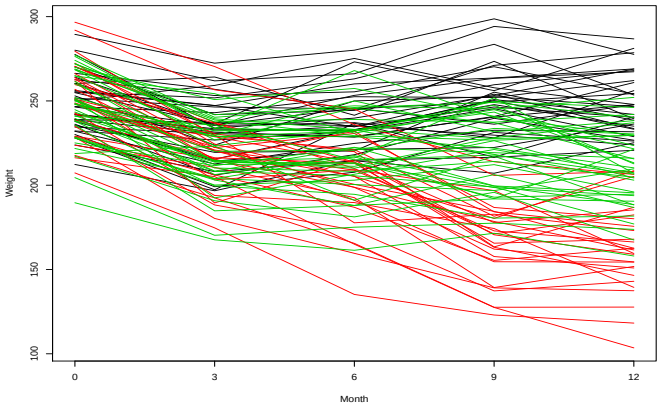


Figure 4.1: Plot of weight loss data set coloured by the three programs.

4.2.1 Mixture Discriminant Analysis

The CDGMM family is fit to each known class in the training set for the weight loss data set. The BIC selected $\mathcal{G}_1 = \mathcal{G}_2 = \mathcal{G}_3 = 1$. This yields that the total number of components is $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2 + \mathcal{G}_3 = 3$ and recorded labels ranging from one to three. Parameter estimates are obtained by fitting a three-component CDGMM with the aforementioned labels. The selected CDGMM was an EEI model which has equal autoregressive structure and equal, isotropic noise across groups. Parameter

estimates are then used to compute soft classifications for each test set. The MAP classifications for 10 different test sets are given in Table 4.1. The average ARI value for the 10 test sets is approximately 0.8982. One may argue that these are very good results. It is interesting to note how the only misclassifications occurred when a test set observation was predicted to be in program three but was in program one or vice versa. This makes sense when looking at Figure 4.1 because programs one and three are more difficult to differentiate between than programs two and three or programs one and two.

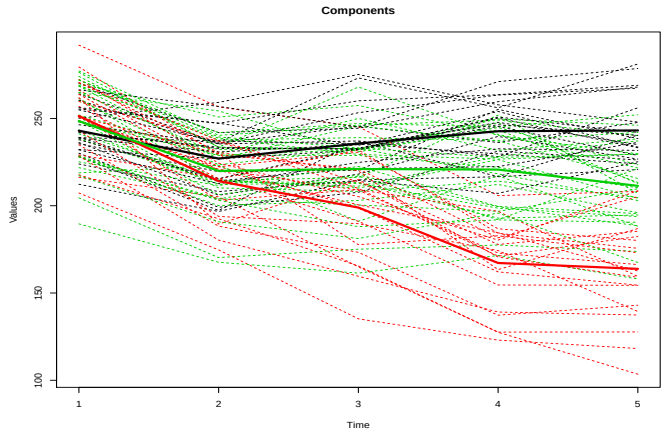


Figure 4.2: Plot of one training set using a CDGMM for the weight loss data.

Table 4.1: Cross-tabulation of the MAP classifications associated with the 10 test sets using the CDGMM (A-C) against true classes for the weight loss data.

	A	B	C
1	78	0	3
2	0	74	0
3	5	0	90

Allowing for linear means in the CDGMM yielded an EEI model. The EEI model

has equal autoregressive structure and equal, isotropic noise across groups. This yielded an ARI of approximately 0.8329 over the course of ten runs. The hardened classifications of 10 different test sets are given below in Table 4.2. While these results are still good, the mean is better modelled by a weighted average than a linear combination. This makes sense when comparing the patterns in Figures 4.3 and 4.2 because the paths are not strictly linear.

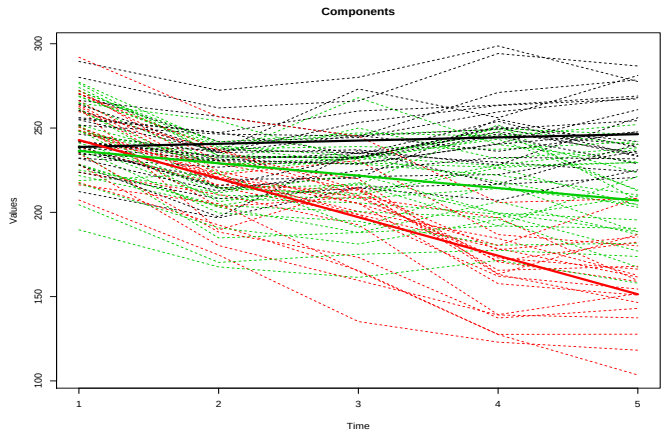


Figure 4.3: Plot of one training set using a CDGMM with linear means for the weight loss data.

Table 4.2: Cross-tabulation of the MAP classifications associated with the 10 test sets using the CDGMM with linear means (A-C) against true classes for the weight loss data.

	A	B	C
1	74	0	8
2	0	67	0
3	4	1	90

Next, each class in the training set is fit to a CDtMM for the weight loss data. The BIC selected $\mathcal{G}_1 = \mathcal{G}_2 = \mathcal{G}_3 = 1$. This implies that the total number of components

is $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2 + \mathcal{G}_3 = 3$ and the recorded labels ranging from one to three. Then, a three-component CDtMM is fit along with the corresponding labels. This is done to obtain parameter estimates that are used to compute soft classifications for the test set. The MAP classifications over 10 distinct test sets is given in Table 4.3. The average ARI value of these runs is approximately 0.8129. The degrees of freedom were all well over 50 for each component and this indicates that this model was essentially a CDGMM. This makes sense as both Tables 4.1 and 4.3 have very similar results.

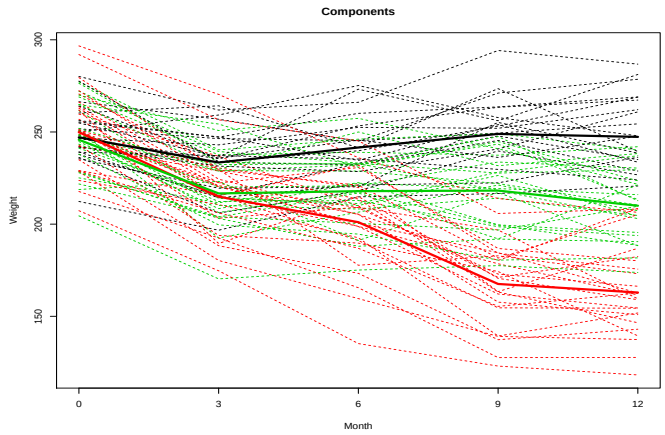


Figure 4.4: Plot of one training set using a CDtMM for the weight loss data.

Table 4.3: Cross-tabulation of the MAP classifications associated with the 10 test sets using the CDtMM (A-C) against true classes for the weight loss data.

	A	B	C
1	76	0	4
2	0	70	0
3	8	0	82

4.3 Italy Power Demand

The second data set that will be looked at is the Italy power demand data set (Keogh *et al.*, 2006). There are a total of 1096 observations and two classes. Class one corresponds to winter months (October to March) and class two relates to summer months (April to September). There are 547 days (observations) in the winter months and 549 days (observations) in the summer months. There are 24 time points over the course of one year. The classification task is to distinguish days from winter and summer months. A subset of the data is given in Figure 4.5. The two classes have fairly different paths. The training set will contain 70% of the data and the remaining 30% will remain as the test set. This implies that 767 observations will belong to the training set and the remaining 329 observations will be in the test set. It should be noted that the results did not change whether stratification was used or not because each class is represented by over 500 observations.

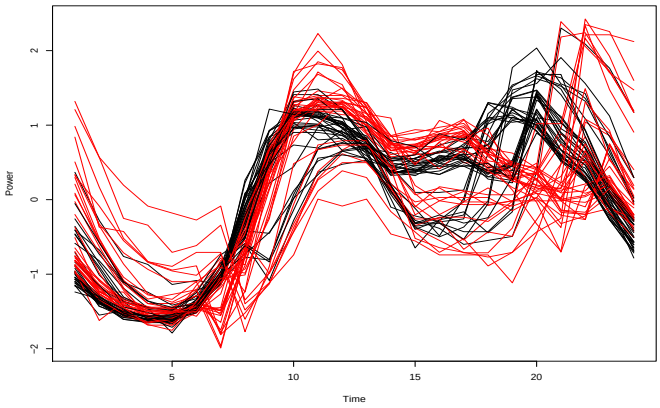


Figure 4.5: Plot of a subset of the Italy power demand data set coloured by winter versus summer months.

4.3.1 Mixture Discriminant Analysis

The classes in the training set for the Italy power demand data are fit using a CDGMM. The BIC selected $\mathcal{G}_1 = \mathcal{G}_2 = 1$. This implies that the total number of components is $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2 = 2$ and the recorded labels ranging from one to two. A two-component CDGMM is fit with the aforementioned labels to obtain parameter estimates. The resulting CDGMM was an EEI model which has equal autoregressive structure and equal, isotropic noise across groups. The parameter estimates are used to compute soft classifications for each test set. The MAP classifications of 10 different test sets are given below in Table 4.4. The average ARI value for all the test sets is approximately 0.8887, which indicates good classification performance. Overall, the CDGMM provides excellent results with minimal classifications, i.e., only 94 observations out of 3290 were misclassified.

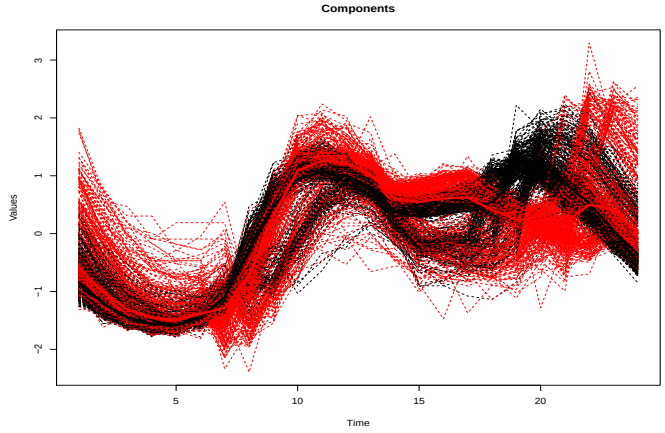


Figure 4.6: Plot of one training set using a CDGMM for the Italy power demand data.

Table 4.4: Cross-tabulation of the MAP classifications associated with the 10 test sets using the CDGMM (A-B) against true classes for the Italy power demand data set.

	A	B
1	1626	58
2	36	1570

Next, a CDtMM is fit to the classes in the training set for the Italy power demand data. The BIC yielded $\mathcal{G}_1 = \mathcal{G}_2 = 1$. This implies that the total number of components is $\mathcal{G} = \mathcal{G}_1 + \mathcal{G}_2 = 2$ and the recorded labels ranging from one to two. A two-component CDtMM is fit along with the aforementioned labels. The selected CDtMM was an EEI model which has equal autoregressive structure and equal, isotropic noise across groups. Once the parameter estimates are obtained, the soft classifications for the test set can be calculated. For 10 different test sets, the MAP classifications are given in Table 4.5. The average ARI value over the 10 runs is approximately 0.8831. The average degrees of freedom for component one (ν_1) was 3.3149 and was 2.3755 for component two (ν_2).

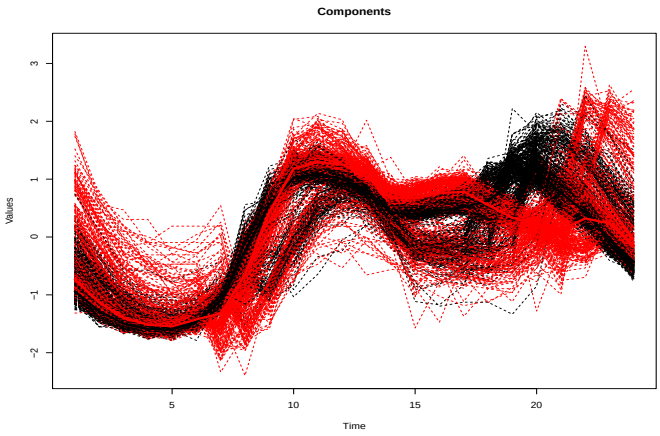


Figure 4.7: Plot of one training set using a CDtMM for the Italy power demand data.

Table 4.5: Cross-tabulation of the MAP classifications associated with the 10 test sets using the CDtMM (A-B) against true classes for the Italy power demand data set.

	A	B
1	1548	71
2	28	1643

Chapter 5

Discussion

Discriminant analysis yielded good results for both simulated data sets. In most cases, each known class was modelled by exactly one component and the average ARI values were fairly high for both simulated data sets. As expected, allowing for a linear mean in the CDGMM yielded worse results.

Applying discriminant analysis to the weight loss data set proved to be very successful. Each weight loss program was modelled by strictly one component. The classification table gave good results for both mixture models with near perfect classification. It is interesting to note how allowing for a linear combination for each group mean in the CDGMM performed slightly worse. One may argue that this is because the data set follows a slight linear path downwards.

Similarly, discriminant analysis was applied to the Italy power demand data set. For the Italy power demand data set, each class was modelled by exactly one component as well. The classification table yielded high overall ARI values for both mixture models and separated the two classes very well despite the two classes having slightly similar paths. Unlike the weight loss data set, the Italy power demand data set does

not follow a linear trend. Thus, modelling the mean by a line of best fit would perform much worse.

In future work, it would also be useful to those in the health care industry when looking at healthy versus ill individuals over time. It would be interesting to look at a cost function, e.g., incorrectly labelling a sick individual as healthy is much worse than incorrectly labelling a healthy individual as sick. A healthy individual labelled as sick is getting care that is not needed whereas the sick individual labelled as healthy is not getting treatment that is needed. Moving forward, one may also be interested in applying other distributions, e.g., mixtures of power exponential distributions (Dang *et al.*, 2015), where mixtures of power exponential distributions are alternative heavy-tailed or less heavy-tailed distributions. One would have to be careful with the covariance decomposition (modified Cholesky).

Bibliography

Aitken, A. (1926). A series formula for the roots of algebraic and transcendental equations. *Proceedings of the Royal Society of Edinburgh*, **45**(1), 14–22.

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

Arbeitman, M. N., Furlong, E. E., Imam, F., Johnson, E., Null, B. H., Baker, B. S., Krasnow, M. A., Scott, M. P., Davis, R. W., and White, K. P. (2002). Gene expression during the life cycle of drosophila melanogaster. *Science*, **297**(5590), 2270–2275.

Benoît (1924). Note sur une méthode de résolution des équations normales provenant de l'application de la méthode des moindres carrés à un système d'équations linéaires en nombre inférieur ï¿œ celui des inconnues (Procédé du Commandant Cholesky). *Bulletin Géodésique*, **2**, 67–77.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P., and Lindsay, B. G. (1994).

The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics*, **46**(2), 373–388.

Bouchard, G. and Celeux, G. (2004). *Model selection in supervised classification*. Ph.D. thesis, INRIA.

Browne, R. P., ElSherbiny, A., and McNicholas, P. D. (2015). *mixture: Mixture Models for Clustering and Classification*. R package version 1.4.

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

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science*, **282**(5389), 699–705.

Crowder, M. J. and Hand, D. J. (1990). *Analysis of Repeated Measures*. London: Chapman and Hall/CRC Press.

Dahl, D. B. (2016). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-2.

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

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

Dominici, F. (2005). Longitudinal Data Analysis. `http://www.biostat.jhsph.edu/~fdominic/teaching/LDA/lda.html`.

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. R package version 2.0-10.

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

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag, Heidelberg.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2017). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-6.

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

Keogh, E., Xi, X., Wei, L., and Ratanamahatana, C. A. (2006). The UCR time series classification/clustering homepage. `http://www.cs.ucr.edu/~eamonn/time\_series\_data`.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, **61**(1), 49–66.

Leroux, B. G. *et al.* (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, **20**(3), 1350–1360.

Lindsay, B. G. (1995). Mixture models: Theory, geometry and applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 5*, Hayward, CA: Institute of Mathematical Statistics.

McIver, D. and Friedl, M. (2002). Using prior probabilities in decision-tree classification of remotely sensed data. *Remote sensing of Environment*, **81**(2), 253–261.

McLachlan, G. J. and Peel, D. (1998). Robust cluster analysis via mixtures of multivariate t-distributions. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pages 658–666. Springer.

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

McNicholas, P. D. (2016b). Model-based clustering. *Journal of Classification*, **33**(3), 331–373.

McNicholas, P. D. and Murphy, T. B. (2010a). Model-based clustering of longitudinal data. *Canadian Journal of Statistics*, **38**(1), 153–168.

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

McNicholas, P. D. and Subedi, S. (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference*, **142**(5), 1114–1127.

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

McNicholas, P. D., Jampani, K. R., and Subedi, S. (2015). *longclust: Model-Based Clustering and Classification for Longitudinal Data*. R package version 1.2.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2017). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.6-8.

Pan, J. and Mackenzie, G. (2003). On modelling mean-covariance structures in longitudinal studies. *Biometrika*, **90**(1), 239–244.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, **86**(3), 677–690.

Pourahmadi, M. (2000). Maximum likelihood estimation of generalised linear models for multivariate normal covariance matrix. *Biometrika*, **87**(2), 425–435.

Pourahmadi, M., Daniels, M. J., and Park, T. (2007). Simultaneous modelling of the cholesky decomposition of several covariance matrices. *Journal of Multivariate Analysis*, **98**(3), 568–587.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**(336), 846–850.

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

Steinley, D. (2004). Properties of the Hubert-Arable adjusted Rand index. *Psychological Methods*, **9**(3), 386.

Student (1908). The probable error of a mean. *Biometrika*, pages 1–25.

Tuddenham, R. D. and Snyder, M. M. (1954). Physical growth of california boys and girls from birth to eighteen years. *Publications in Child Development.*, **1**(2), 183.