UNIVERSITY OF
LIVERPOOL

Discriminant analysis of multivariate longitudinal data: Statistical methods and clinical applications

Thesis submitted in accordance with the requirements of
the University of Liverpool for the degree of Doctor in Philosophy
by

Riham El Saeiti

May, 2019

Abstract

Thesis title: Discriminant analysis of multivariate longitudinal data: Statistical methods and clinical applications
Author: Riham El Seaiti

There is an increasing interest in longitudinal discriminant analysis (LoDA) approaches that use patients' longitudinal data to predict their future clinical status. LoDA utilises (multivariate) generalised linear mixed models (GLMM) where changes over time of markers with predictive ability can be jointly modelled. This thesis aimed to address the following questions: (i) What is the benefit of using LoDA rather than classical quadratic discriminant analysis (QDA) for clinical classifications? (ii) What is the best way to utilise longitudinal data for clinical classification? (iii) How does the misspecification of the random-effects distribution impact on classification accuracy? Additionally, I investigated whether the number of patients and repeated measurements affect the ability of the predictive tool to classify patients correctly.

Three approaches of LoDA, using different types of markers (continuous, binary, and discrete) were compared, first using data from the Mayo Primary Biliary Cirrhosis study, and then using a simulation study to investigate different uses of a patient's longitudinal data. The impact of random-effects misspecification on classification accuracy was assessed by examining several scenarios where data was generated using four different 'true' random-effects distributions.

I found that methods that take the relationship between repeated measurements from the same subject into account provided more accurate classification than methods that treated each time point as a single variable and also used the data more efficiently. Using the marginal distribution of a patient's longitudinal data often provided the best results if the average profiles in each clinical group were not the same. However, when differences in variability between groups were apparent, then the distribution of a patient's random effects gave the best classification. Consequently, I recommend that researchers should consider the marginal and random effects approaches as first options when performing LoDA.

Misspecification of the random-effects distribution has a minimal impact on classification accuracy when the departure from normality is *small*. However, when the departure from normality is *large*, assuming a more flexible random effects distribution can provide greater classification accuracy.

Acknowledgements

First and foremost I thank Allah for the guidance and enlightenment provided to me which without this project would have been not completed.

I would like to express my sincere gratitude to my supervisors Prof. Marta García-Fiñana, Dr. Gabriela Czanner and Dr. David M. Hughes for the continuous support of my PhD study and related research, for their patience, motivation, and immense knowledge. Their guidance helped me during my time doing research and during the writing up of this thesis. I have been fortunate to have a supervisory team who cared so much about my work, and who responded to my questions and queries so promptly. I could not have imagined having a better supervisors and mentors for my PhD study.

I would like to give special thanks to Prof. Marta García-Fiñana for recognising how difficult of being far from home at times when I was experiencing tragedies in my home country.

I am grateful to Prof. Simon P Harding from St. Paul's Eye Unit at Royal Liverpool Hospital for providing the ophthalmic clinical data. I would like to thanks Dr. Ian Smith for his help with using Condor Service. I also express my sincere appreciation to all my friends and staff of the Department of Biostatistics, especially the members of Multivariate Modelling Group.

I gratefully acknowledge the funding received towards my PhD from the Ministry of Higher Education and Scientific Research, Libya and the University of Benghazi for giving me this opportunity.

Last but not least, I would like to thank my parents (*Awad* and *Fathia*) and my brothers (*Serag*, *Anas*, *Islam* and *Sanad*) and sister (*Aml*) for supporting and encouraging me to spiritually throughout my PhD. Also, I acknowledge with love my family, my husband *Taher* who has been a great source of patience, understanding during my PhD and my children *Khadija*, *Aisha* and *Ali* who make my every day worthwhile.

Contents

Abstract i

Acknowledgements ii

Contents vii

List of Figures xiii

List of Tables xvi

List of Abbreviations xviii

Publications xix

1 Introduction 1

1.1 Discriminant analysis and classification in multivariate longitudinal data 1

1.1.1 Longitudinal data 2

1.1.2 Discrimination and classification 3

1.2 Aims, Objectives and Motivation 4

1.3 Introduction to the datasets 6

1.3.1 Data description: Ophthalmic data 6

1.3.2 Data description: PBC data 9

1.4 The structure of this thesis 10

2 Review of current methodologies 11

2.1 Introduction 11
2.2 Classical discriminant analysis 12
2.2.1 Linear discriminant analysis (LDA) 14
2.2.2 Quadratic discriminant analysis (QDA) 16
2.2.3 Test for homogeneity of variance-covariance matrices 17
2.2.4 Classification rule 20
2.3 Longitudinal discriminant analysis (LoDA) approaches 20
2.3.1 Modified discriminant analysis 25
2.3.2 Mixture multivariate generalised linear mixed model (MMGLMM) 26
2.3.3 Bayesian method and Markov chain Monte Carlo (MCMC) method 28
2.3.4 Estimation 29
2.3.5 Marginal approach 33
2.3.6 Conditional approach 35
2.3.7 Random-effects approach 35
2.3.8 Classification rules 36
2.4 Assessment of classification accuracy 36
2.4.1 Validation Method 36
2.4.2 Receiver operating characteristic curve (ROC) 38
2.4.3 Area under curve AUC 40
2.5 Summary 42

3 Analysis of clinical data: Ophthalmic application 44

3.1 Introduction 44
3.2 Description of data types 45
3.3 Classical linear and quadratic discriminant analysis 48
3.4 Comparison of classical and modified discriminant analysis 61
3.4.1 Modified discriminant analysis 62
3.5 Simulation Study 70

3.6 Discussion 77

4 Comparison of prediction approaches in longitudinal discriminant analysis 79

4.1 Introduction 79

4.2 Primary Biliary Cirrhosis Data 80

4.3 Simulation Study 87

4.3.1 Scenario 1 92

4.3.2 Scenario 2 97

4.4 Discussion 103

5 Impact of misspecified random effects distribution 105

5.1 Introduction 105

5.2 PBC application 108

5.3 Simulation 113

5.3.1 Simulation Setup 114

5.3.2 Effect of the misspecification of the random effects on classification accuracy 133

5.4 Summary 151

6 Conclusions and Further Work 152

6.1 Introduction 152

6.2 Summary of the main findings 154

6.2.1 Classical discriminant analysis versus modified discriminant approach 154

6.2.2 Longitudinal discriminant analysis (LoDA) approaches 155

6.2.3 Impact of the misspecification of the random-effects distribution on the classification accuracy 157

6.3 Recommendations for practice 158

6.4 Future work 160

6.5 Conclusions 161

Appendices - R code **173**

A Modified discriminant analysis step two **174**

B Creating Simulated Datasets **177**

C Performing Longitudinal Discriminant Analysis **180**

D Calculating classification accuracy measurements **182**

List of Figures

1.1 Types of longitudinal data where (-) indicates missing values. 3

1.2 Profile of visual acuity and contrast sensitivity of success (left panel) and failure (right panel) groups over a year. Thick lines show the overall mean over time and across patients. 7

2.1 Finding best cut-off from the ROC curve 39

2.2 Comparison of three ROC curves with different areas. 41

3.1 Flow chart illustrating the comparisons between an unbalanced dataset (D1) and a balanced dataset (D2) and the time approximation. 48

3.2 Flow chart illustrating an unbalanced dataset (D3) and a balanced and imputed dataset (D4). 49

3.3 ROC curves for the linear discriminant analysis (left panel), and for the quadratic discriminant analysis(right panel) using contrast sensitivity. . 57

3.4 ROC curves for the linear discriminant analysis (left panel), and for the quadratic discriminant analysis(right panel) using visual acuity. 58

3.5 ROC curves for the multivariate models. ROC curves for the linear discriminant analysis (left panel) and for the quadratic discriminant analysis (right panel). 59

3.6 Comparison of AUC for linear and quadratic functions that used CS and VA (separately and together), measured using different follow-ups, to predict treatment success or treatment failure. 60

3.7 Receiver operating characteristic (ROC) curves of the classical discriminant model using the balanced data (D2, left plot) and of the modified discriminant model using the unbalanced data (D1, right plot) at four time points. 66

3.8 Area under ROC curve (AUC) for the modified and the classical discriminant analysis approaches to predict failure of treatment over time. 67

3.9 Receiver operating characteristic (ROC) curves of the modified discriminant analysis for the unbalanced dataset (D3, left plot) and of the classical discriminant analysis for the balanced, imputed dataset (D4, right plot) at four time points. 69

3.10 Area under ROC curve (AUC) for the modified and the classical approaches that applied to the unbalanced dataset (D3) and balanced, imputed dataset (D4), respectively to predict failure of treatment over time. 69

3.11 ROC curves of the modified discriminant analysis (right plot) and the classical discriminant analysis (left plot) at four time points (10% of data missing). 73

3.12 ROC curves of the modified discriminant analysis (right plot) and the classical discriminant analysis (left plot) at four time points (20% of data are missing). 73

3.13 ROC curves of the modified discriminant analysis (right plot) and the classical discriminant analysis (left plot) at four time points (40% of data missing). 75

3.14 Area under ROC curve (AUC) for the modified (dash lines) and the classical (solid lines) approaches with 10%, 20% and 40% of data missing to predict failure of treatment over time. 76

4.1 Trace plot of the MCMC chain of the model deviance $\mathbb{D}$. Right Figure refers to Group 1 and left Figure refers to Group 0. 82

4.2 Trace plot of the MCMC chain of Group 0 of the model means μ. . . . 83

4.3 Trace plot of the MCMC chain of Group 1 of the model means μ. 84

4.4 Trace plot of the MCMC chain of Group 0 of the model standard deviations derived from the covariance matrices $\mathbb{D}$. 85

4.5 Trace plot of the MCMC chain of Group 1 of the model standard deviations derived from the covariance matrices $\mathbb{D}$. 86

4.6 Longitudinal profiles of albumin (mg/dl), log(bilirubin) (log(mg/dl)), platelet counts and blood vessel malform (spiders) for patients who were known to be alive at 5 years (Group 0, solid lines) and who died between 2.5 and 5 years (Group 1, dashed lines). The thick lines show fitted mean of patients over time, calculated using loess. 87

4.7 Receiver Operating Characteristic curves of the dynamic LoDA using the random effects (solid), marginal (dotted) and conditional (dot-dashed) prediction approaches. PBC data are collected during the first 2.5 years data are used for the modelling and prediction. 89

4.8 Receiver Operating Characteristic curves of the dynamic LoDA using the random effects (solid), marginal (dotted) and conditional (dot-dashed) prediction methods for the whole PBC data with average of 7.03 visits per patient. 90

4.9 Simulation study Scenario 1: Longitudinal profiles of albumin, bilirubin, platelet count and blood vessel malform (spiders) for patients who were known to be alive at 5 years (Group 0, solid lines) and who died between 2.5 and 5 years (Group 1, dashed lines). The thick lines show fitted mean of patients over time, estimated using loess. 93

4.10 Receiver Operating Characteristic curves of the dynamic LoDA using the random effects (solid), marginal (dotted) and conditional (dot-dashed) prediction methods for Scenario 1. 96

4.11 Histograms showing the sensitivity, specificity, PCC and AUC of each of the three approaches for each of the 100 simulated datasets under Scenario 1. 97

4.12 Receiver Operating Characteristic curves of the dynamic LoDA using the random effects (solid), marginal (dotted) and conditional (dot-dashed) prediction methods for Scenario 2. 102

4.13 Histograms showing the sensitivity, specificity, PCC and AUC of each of the three approaches for each of the 100 simulated datasets under Scenario 2. 103

5.1 Profiles of three markers log(bilirubin), platelet counts and blood vessel malformation (spiders) for patients who were known to be alive at 5 years (Group 0, left panel) and who died after 2.5 years (Group 1, right panel). The thin lines show the profiles of individuals in the PBC data, and the thick lines show the overall mean, as estimated by loess. 110

5.2 Receiver Operating Characteristic (ROC) curves for models with $K = 1, 2, 3, 4$ mixture components in the PBC data. 113

5.3 Density functions of five random effects distributions for the single normal assumption. Red and blue lines show the density functions for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1) respectively. 119

5.4 Profiles from randomly selected simulated dataset with 2 components normal distribution with small departure from normality. Profiles of three markers log(bilirubin), platelet counts and blood vessel malformation(spiders) for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1). The black line shows the overall mean, the pink and blue lines show the mean of two mixture components, estimated using loess. 120

5.5 Density functions of the random effects for the 2-components assumption with small departure from normality for two groups. Red and blue lines show the density function for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1) respectively. . 121

5.6 Profiles from randomly selected simulated data with 3 mixture components normal distribution with small departure from normality. Profiles of three markers log(bilirubin), platelet counts and blood vessel malformation(spiders) for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1). The black line shows the overall mean, while the pink, green and blue lines show the mean of three mixture components, estimated using loess. 122

5.7 Probability density functions of five random-effects elements which follow a 3 mixture normal distribution with small departure from normality for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1). 123

5.8 Simulated data from 2 mixture components normal distribution with a large departure from normality. Profiles of three markers log(bilirubin), platelet counts and blood vessel malformation(spiders) for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1). The black line shows the overall mean, while the pink and blue lines show the mean of two mixture components, estimated using loess. 127

5.9 Density function for the 2-components normal assumption with large departure from normality. Red and blue lines show the density function for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1) respectively. 128

5.10 Density function for the 3-components normal assumption with a large departure from normality. Red and blue lines show the density function for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1) respectively. 129

5.11 Simulated data from 3 mixture components normal distribution with a large departure from normality. Profiles of three markers log(bilirubin), platelet counts and blood vessel malformation(spiders) for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1). The black line shows the overall mean, while the pink, green and blue lines show the mean of three mixture components, estimated using loess. 130

5.12 T-distribution density function for two groups with 3 (solid lines) and 5 (dashed lines) degrees of freedom. Red and blue lines show the density function for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1) respectively. 132

5.13 Receiver Operating Characteristic curves for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a single normal distribution. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets. 135

5.14 Receiver Operating Characteristic curves for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a 2-component mixture of normals with small departure from normality. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets. 137

5.15 Receiver Operating Characteristic curves for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a 2-component mixture of normals with large departure from normality. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets. 138

5.16 Receiver Operating Characteristic curves for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a 3-component mixture of normals with small departure from normality. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets. 140

5.17 Receiver Operating Characteristic curves for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a 3-component mixture of normals with large departure from normality. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets. 141

5.18 Receiver Operating Characteristic curves for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a T-distribution with 3 degrees of freedom. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets. 143

5.19 Receiver Operating Characteristic curves for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a T-distribution with 5 degrees of freedom. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets. 144

5.20 Receiver Operating Characteristic curves of the random effects approach for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a single normal. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets. 148

5.21 Receiver Operating Characteristic curves of the random effects approach for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a 2-components multivariate normal with a high degree of departure from normality. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets. 149

5.22 Receiver Operating Characteristic curves of the random effects approach for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a t-distribution with 3 degrees of freedom. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets. 150

List of Tables

2.1 Classification table: Predicted versus true outcome. 38

3.1 Statistics results of the Box's test that uses the univariate markers separately (CS, VA) and the multivariate markers (CS and VA together). Each model included age at baseline. 52

3.2 Results of the Levene's test that uses the univariate markers separately (CS, VA) and the multivariate markers (CS and VA together). Each model included age at baseline. 53

3.3 The test of equality of variances results of the Bartlett's test that uses the univariate markers separately (CS, VA) and the multivariate markers (CS and VA together). Each model included the age at baseline. 54

3.4 Results of the linear and quadratic discriminant analyses that uses a univariate marker (CS and VA, separately) and multivariate markers CS and VA together. 55

3.5 Fixed effects parameters of the MLMM based on contrast sensitivity (CS) and visual acuity (VA) for the prognostic groups (i.e., success and failure of treatment). 64

3.6 Accuracy measures for the modified QDA that used the unbalanced dataset (D1) and for the classical QDA that used the balanced data (D2). Note that both dataset use the same patients and same measurements. 65

3.7 Accuracy values of the modified QDA model that used the unbalanced dataset (D3) and of the classical QDA that used the balanced and imputed data (D4). 68

3.8 Parameter estimates for each simulation scenario. 71

3.9 Accuracy values of the modified QDA model that used the unbalanced dataset and of the classical QDA that used the balanced and imputed data. (Simulation study when 10% of data are missing). 72

3.10 Accuracy values of the modified QDA model that used the unbalanced dataset and of the classical QDA that used the balanced and imputed data. (Simulation study when 20% of data are missing). 74

3.11 Accuracy values of the modified QDA model that used the unbalanced dataset and of the classical QDA that used the balanced and imputed data. (Simulation study when 40% of data are missing). 75

4.1 Prediction accuracy using leave-one-out cross validation for the random-effects, marginal and conditional approaches. PBC data collected during the first 2.5 years were used for the modelling and prediction. The average number of visits per patient was cohort 3.53, in this 2.5 years period. 88

4.2 Prediction accuracy using leave-one-out cross validation based on the random-effects, marginal and conditional approaches. The average number of visits per patient was 7.03 visits in the full PBC dataset. 88

4.3 Parameter estimates for the PBC data and the modifications used for each simulation scenario. Blank entries occur when the parameter was not used in Scenario 2. 91

4.4 Simulation study Scenario 1: Posterior Means, highly probable density (HPD) intervals, bias, standard deviation (SD), mean square error (MSE) and coverage for the fixed and random effects. These measurements were the average of 100 simulations. 94

4.5 Prediction accuracy for the simulated data from Scenario 1 based on leave-one-out cross validation for the random-effects, marginal and conditional approaches. 95

4.6 Simulation study Scenario 2: Posterior Means, highly probable density (HPD) intervals, bias, standard deviation (SD), mean square error (MSE) and coverage for the fixed and random effects. These measurements were the average of 100 simulations. 100

4.7 Scenario 2 prediction accuracy based on leave-one-out cross validation for the random-effects, marginal and conditional approaches. 101

5.1 Penalized Expected Deviance for models with different number of mixture components (K = 1, 2, 3, 4) in the random effects. 111

5.2 Prediction accuracy from leave-one-out cross-validation of the marginal approach with K mixture components in the random effects distribution ($K = 1, 2, 3, 4$) using the PBC data. 112

5.3 Model parameters for the random effects under the assumption that the random effects jointly follow a single component multivariate normal distribution. 116

5.4 Model parameters for the random effects under the assumption that the random effects jointly follow a 2-component multivariate normal distribution with a small departure from normality. 117

5.5 Model parameters for the random effects under the assumption that the random effects jointly follow a 3-component multivariate normal distribution with a small departure from normality. 118

5.6 Model parameters for the random effects under the assumption that the random effects jointly follow a 2-component multivariate normal distribution with a large departure from normality. 125

5.7 Model parameters for the random effects under the assumption that the random effects jointly follow a 3-component multivariate normal distribution with a large departure from normality. 126

5.8 Results of the simulation study under the assumption of a single normal distribution for the random effects. Prediction accuracy of the marginal approach from leave-one-out cross validation for N = 250 patients, 70% training and 30% testing for N = 2,500. 134

5.9 Prediction accuracy for the marginal approach under the assumption that the random effects follow a 2 component normal mixture distribution. 136

5.10 Prediction accuracy for the marginal approach under the assumption that the random effects follow a 3 component normal mixture distribution. 139

5.11 Prediction accuracy of the marginal approach under that assumption that the random effects follow a T-distribution with 3 and 5 degrees of freedom. 142

5.12 Prediction accuracy of the random-effect approach under the assumption that the random effects distribution follow a single normal distribution for N = 250 and 2,500 patients. 146

5.13 Prediction accuracy of the random-effect approach under the assumption that the random effects distribution follow a 2-component normal distribution with large departure from normality for N = 250 and 2,500 patients. 147

5.14 Prediction accuracy of the random-effect approach under the assumption that the random effects distribution follow a T-distribution with 3 degrees of freedom for N = 250 and 2,500 patients. 147

List of Abbreviations

AMD	Age-related Macular Degeneration
AUC	Area Under the Curve
CS	Contrast Sensitivity
DA	Discriminant Analysis
DDA	Descriptive Discriminant Analysis
EM	Expectation Maximization
LDA	Linear Discriminant Analysis
LoDA	Longitudinal Discriminant Analysis
LOOCV	Leave-one-out Cross-Validation
LMM	Linear Mixed-effects Model
LRT	Likelihood Ratio Test
MANOVA	Multivariate analysis of variance
MCMC	Markov chain Monte Carlo
MGLMM	Multivariate Generalised Linear Mixed Model
ML	Maximum likelihood approach
MLMM	Multivariate Linear Mixed Model
MMGLMM	Mixture Multivariate Generalised Linear Mixed Model

nAMD	Nonvascular Age-related Macular Degeneration
NPV	Negative Predictive Value
QDA	Quadratic Discriminant Analysis
PBC	Primary Biliary Cirrhosis
PCC	Probability of Correct Classification
PDA	Predictive Discriminant Analysis
PPV	Positive Predictive Value
ROC	Receiver Operating Characteristic
VA	Visual Acuity

Publications of Work in this Thesis

Chapter 4 has been published in:

Hughes, D. M., El Saeiti, R. and Garcia-Fiñana, M. (2018), 'A comparison of group prediction approaches in longitudinal discriminant analysis', *Biometrical Journal* 60(2), 307-322.

Chapter 5 has been submitted in:

El Saeiti, R., Garcia-Fiñana, M. and Hughes, D. M. (2019), The effect of random-effects misspecification on classification accuracy. submitted.

Chapter 1

Introduction

This thesis describes a number of multivariate longitudinal approaches and explores the advantages and disadvantages of applying these methods for clinical discrimination. Discriminant analysis (DA) is a technique for classification of data (particularly new observational units) into categories or groups to which they most probably belong. DA has been applied in different areas, for example, finance (Joy and Tollefson (1975)), psychology (Alexakos (1966)) and in medical research (Jain and Jain (1994)). DA can be applied to multivariate longitudinal problems in which data are collected for the same subjects repeatedly over a period of time. An introduction to the discriminant analysis and classification in multivariate longitudinal data is given in Section 1.1. Section 1.2 states the objectives of this thesis. The datasets used in this thesis are described in Section 1.3. Finally, the structure of this thesis is explained.

1.1 Discriminant analysis and classification in multivariate longitudinal data

Discrimination and classification are an important part of medical research. The number of clinical studies that involve the collection of multiple variables repeatedly for the same patient through time (i.e., longitudinal data) is currently increasing. Conse-

quently, research in the area of discrimination and classification has also expanded to analyse this type of data, in order to assist clinical care by providing decision support and predictions of future status.

1.1.1 Longitudinal data

Longitudinal data, where observations are measured repeatedly from patients at several occasions are often collected for clinical research. Longitudinal data is in contrast to cross-sectional data, in which observations are collected from patients at a particular time point (Hand (2017), Tang and Tu (2012), Diggle et al. (2002)). A key advantage of measurements repeatedly over time is that they can be used to explore the variations in both between- and within-individual which, is not the case with cross-sectional data (Tang and Tu (2012)).

Longitudinal data can be structured where all subjects have repeated measurements for each marker at the same n time points (known as balanced design). Alternatively, datasets could involve subjects that are observed, for each marker, at different time points and the number of measurements may also vary across patients (unbalanced design). Longitudinal data may show missing values such that markers' measurements for some subjects are incomplete, leading to so-called incomplete cases (Shah et al. (1997), Marshall and Barón (2000)).

Usually, longitudinal data includes more than one marker, all measured over time and which may be similar or of different types (e.g., continuous, binary and counts). Longitudinal data can be utilised for analysis in two ways that depend on data available and research question. The first case involves only analysing a single marker and is referred to as univariate longitudinal analysis. The second case considers multiple longitudinal markers in one study and is referred to as multivariate longitudinal analysis (Verbeke et al. (2014)).

In longitudinal data, there are different data structures (see Figure 1.1). Longitudinal data involves repeated observations of the same patients over a period of time;

these repeated observations over time could be balanced or unbalanced.

In this thesis, I define a balanced dataset as a dataset in which all patients are observed at the same time points, Figure 1.1 (a) and (b). An unbalanced dataset is one in which the time points of observation differ across patients, Figure 1.1 (c).

A complete dataset is defined as one in which all patients have been measured at all of the scheduled markers (all patients must have observations at each marker at different/same time points, Figure 1.1 (a)), whereas an incomplete dataset occurs when some patients have some incomplete profiles (i.e., missing measurements, Figure 1.1 (b)).

Patient	Time	Y1	Y2	...	Yr
1	0	X	X	...	x
1	3	X	X	...	X
1	6	X	X	...	X
⋮	⋮	⋮	⋮	⋮	⋮
N	0	X	X	...	X
N	3	X	X	...	X
N	6	X	X	...	X

(a) Balanced and complete data

Patient	Time	Y1	Y2	...	Yr
1	0	-	X	...	x
1	3	X	-	...	X
1	6	X	X	...	X
N	0	X	-	...	-
N	3	X	X	...	-
N	6	-	X	...	X

(b) Balanced and incomplete data

Patient	Time	Y1	Y2	...	Yr
1	0	X	X	...	X
1	3	X	X	...	X
1	6	X	X	...	X
⋮	⋮	⋮	⋮	⋮	⋮
N	0	X	X	...	x
N	5	X	X	...	X

(c) Unbalanced and complete data

Figure 1.1: Types of longitudinal data where (-) indicates missing values.

1.1.2 Discrimination and classification

In statistics, there is a strong association between discrimination and classification. DA aims to search for an optimal way to distinguish between two or more groups by maximising their differences. Meanwhile, classification seeks to allocate a new patient

into one of the predefined groups by minimising the misclassification rate (Krzanowski (2000)). In this thesis, I used the area under the curve (AUC) and sensitivity, specificity, probability of correct classification (PCC), positive predictive value (PPV) and negative predictive value (NPV) to assess classification accuracy.

DA has been applied to longitudinal data where each visit is treated as a separate variable. Several examples in which longitudinal data can be used within a classical discriminant analysis framework are discussed in Chapter 3. DA approaches are not restricted to complete and balanced longitudinal datasets. Recently, methods for DA have been further developed to analyse unbalanced datasets (Tomasko et al. (1999), Marshall and Barón (2000)). Longitudinal discriminant analyses (LoDA) have been recently developed to use a patient's multivariate longitudinal history to predict their group membership (Morrell et al. (2007), Komárek et al. (2010), Hughes et al. (2018b)).

Several possible rules can be used to allocate a new patient into one of G (prognostic) groups. A considerable amount of literature has been published on classification rules (Huberty and Olejnik (2006), Rizopoulos (2012), Hansen et al. (2010), Hughes et al. (2017)). For example, Huberty and Olejnik (2006) discussed in their book four possible rules (namely: maximum likelihood, typicality probability, posterior probability and prior probability) to calculate probabilities that can be used to assign subjects to a suitable group. They used a general scheme which is to allocate the individual to which group membership probability is highest. Hughes et al. (2017) addressed possible allocation rules to allocate patients based on their estimated group probabilities. Additionally, they introduced a new classification scheme based on credible intervals for group membership probabilities for improving classification in a dynamic LoDA.

1.2 Aims, Objectives and Motivation

The aims of this thesis are:

1. To investigate the benefits of using longitudinal discriminant analysis (LoDA)

rather than classical linear and quadratic discriminant analysis for clinical classifications. While LoDA takes into account correlation between repeated measurements on the same patients, classical linear/quadratic discriminant analysis deals with each time point as a separate variable, and so does not allow for correlation between repeated measures on the same individual.

2. To explore the classification accuracy of three approaches of LoDA (namely: marginal, conditional and random-effects). Each of these approaches uses a different aspect of the patients' longitudinal history to predict the patient prognostic group.

3. To assess the impact that misspecification of the random-effects distribution has on classification accuracy. Additionally, I investigate whether the sample size (i.e., number of individuals) and the number of repeated measurements per individual may affect the ability of the predictive tool to classify patients correctly.

Motivation

The motivation for my statistical work comes from two areas: ophthalmology and cirrhosis disease. Ophthalmology is an area of medicine devoted to the diagnosis and treatment of eye diseases. Cirrhosis is a disease in which the liver does not work properly as a result of long-term damage. The motivation behind the ophthalmic data is that clinicians want to predict treatment success or treatment failure early in patients who have age-related macular degeneration (AMD). Current practice is that the effectivity of the treatment is evaluated one year after the treatment initiation. The clinical question that I will address with the cirrhosis dataset is to identify those patients who will not survive or need a transplant during the five years after transplantation. These patients met the eligibility criteria for the randomised placebo-controlled trial of the drug D-penicillamine (DPCA).

1.3 Introduction to the datasets

This thesis included analyses of two longitudinal datasets. The first dataset came from a clinical ophthalmic study. The ophthalmic dataset was called LOUISE and consisted of information on 1008 patients who were treated with verteporfin photodynamic therapy at St. Paul's eye unit royal Liverpool University between 1999-2006. This dataset will be described in detail in next section.

The second dataset was a subset of the data from a Mayo Clinic trial with primary biliary cirrhosis (PBC) conducted in 1974-1984 Dickson et al. (1989) and included information on 312 patients who required liver transplantation. This dataset will be described in detail in Section 1.3.2.

1.3.1 Data description: Ophthalmic data

The ophthalmic dataset included clinical information from 1008 patients who had non-vascular age-related macular degeneration (nAMD) during the seven-years interval. AMD is a progressive, irreversible painless eye condition that generally leads to the gradual loss of central vision, which occurs in the central area of the retina (macula) in people aged 55 years and over, (Ferris III et al. (2013)), and affects more than 600,000 people in the UK to some degree every year. In 2011, the British Journal of Ophthalmology presented that by 2020 the number of people who will have AMD could increase to approximately 756,000 (Minassian et al. (2011)). The ophthalmic dataset can be made available upon request.

In this thesis, the measurements of visual acuity and contrast sensitivity were used, which were repeatedly measured over time. These measurements are essential in monitoring the function of the eye, and the idea is that they should help ophthalmologists to identify the deterioration in visual function. The first measure of sight deterioration was a loss of visual acuity (VA), which is the loss of ability to read small letters and to detect fine details in central vision, for example when driving or reading. The second

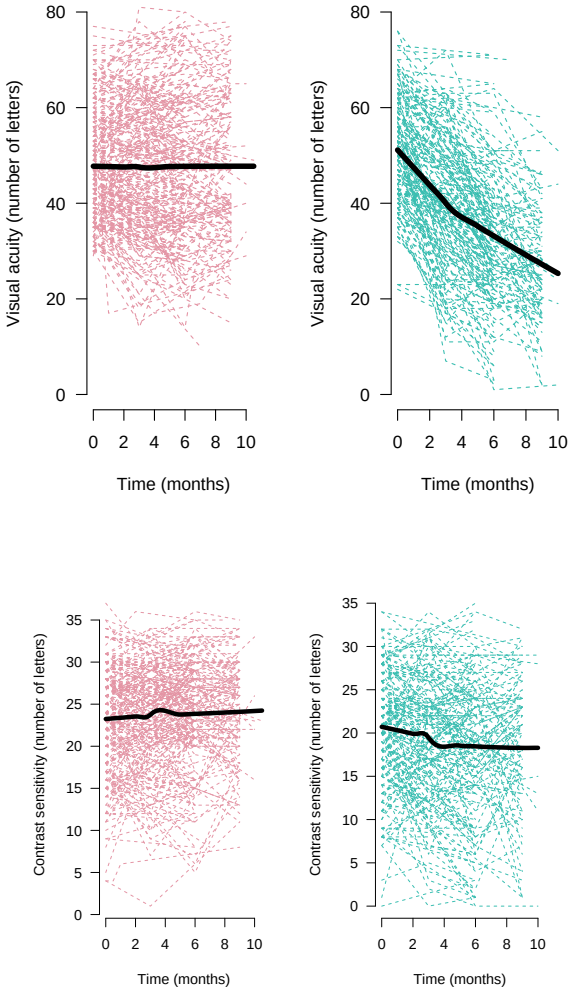

Figure 1.2: Profile of visual acuity and contrast sensitivity of success (left panel) and failure (right panel) groups over a year. Thick lines show the overall mean over time and across patients.

measurement was contrast sensitivity (CS), which is the ability to see less well-defined objects, such as faces against a background (Bellmann et al. (2003)). Visual acuity was measured as letters read at a distance of one metre using the ETDRS (Early Treatment Diabetic Retinopathy Study test) with logMAR charts, while contrast sensitivity was assessed using a Pelli–Robson chart (García-Finana et al. (2010)).

My observational ophthalmic dataset contains information on patients who were treated with verteporfin photodynamic therapy for neovascular AMD over 12 months. One of the clinical purposes was to be able to identify patients who will benefit from

the treatment. The clinical dataset was described by García-Finana et al. (2010) and included covariates, such as CS, VA, lesion type, age and gender. In cases of bilateral treatment, the first treated eye was accounted for as the study eye.

In order to be able to address the research question described above, I included patients who were followed for at least 12 months. Therefore, my sample consisted of 447 patients. I used measurements of CS and VA which were repeatedly measured over time (more detail is presented in the Chapter 3).

Data was intended to be collected at specific time points: at baseline, 3, 6, 9 and 12 months as per protocol. Treatment failure was determined as a fall in VA to below 20 letters at 12 months and/or a loss of VA $\geqslant 15$ letters from baseline (García-Finana et al. (2010)). Profiles of VA and CS of treatment success and treatment failure for patients from baseline to 12 months are shown in Figure 1.2. In the treatment success group, the profiles of VA remain stable over a year. However, the profiles of treatment failure show an overall decreasing trend in VA over 12 months. Profiles of CS of success and failure groups stay constant over time (Figure 1.2).

1.3.2 Data description: PBC data

The Dutch Multicenter Primary Biliary Cirrhosis (PBC) study was used in an application of longitudinal discriminant analysis by Komárek et al. (2010). They presented a discriminant method that relaxes the normality assumption on random effects by using a normal mixture in the random effects distribution. In this thesis, the motivation came from a similar PBC dataset, known as the Mayo Clinic trial, which collected data on patients with primary biliary cirrhosis (PBC) between 1974–1984 (Dickson et al. (1989)). Komárek et al. (2013) applied cluster analysis to the PBC dataset where three markers were used to classify patients who survived without liver transplantation the first 2.5 years (910 days). This dataset is available within the `R` package `mixAK` Komárek and Komárková (2014), in Appendix D of Fleming and Harrington (1991) and electronically at *http* : *//lib.stat.cmu.edu/datasets/pbcseq*. This dataset included information on 312 patients with a large number of clinical variables for each patient and with a median follow-up time equal to 6.3 years.

For the purpose of this research, I focussed on patients who survived without liver transplantation after five years and who died or had a liver transplant at some point between two and a half and five years. I classified the patients into two groups, 202 patients were classified as patients known to be alive at five years (Group 0) and 51 patients who died or had a liver transplant between 2.5 and 5 years (Group 1). Four markers (albumin, platelet count, bilirubin and blood vessel malformation) were used for the comparison of three prediction approaches (marginal, conditional and random-effects) of longitudinal discriminant analysis (more detail is given in the Chapter 4). In the second application with this dataset, three of these markers (platelet count, bilirubin and blood vessel malformation) were used to investigate whether the misspecification of random effects may affect the model ability to classify patients into prognostic groups accurately (Chapter 5 discusses in more detail).

1.4 The structure of this thesis

The overall structure of the thesis takes the form of six chapters, including this introductory chapter. In Chapter 2, I present a range of discriminant analysis approaches (classical discriminant analysis approaches and longitudinal discriminant analysis approaches) that can be used to analyse longitudinal data. Also, I review assessments of classification that have been used to evaluate the results.

In Chapter 3, I use the AMD dataset to compare classical quadratic discriminant analysis which deals with each time point as a single variable with the modified quadratic discriminant analysis. In the latter, the multivariate linear mixed model is used to estimate the parameter means and covariance matrices.

In Chapter 4, I describe the three approaches for LoDA (marginal, conditional and random effects) which take different strategies when using the patients' longitudinal data to predict the future status for a new patient. I use the PBC dataset and several simulation studies to investigate in which situation each of the three approaches is superior.

The three approaches of LoDA (which are compared in Chapter 4) are based on a mixed model that uses random effects to model the correlation between repeated measurements on the same subject. The random effects are typically assumed to follow a (possibly multivariate) normal distribution. In Chapter 5, I use the PBC dataset to investigate whether the misspecification of the random effects distribution affects the classification performance. Finally, the conclusions drawn from these investigations and further work are described in Chapter 6.

Chapter 2

Review of current methodologies

The previous chapter introduced the clinical motivations for this thesis. In the ophthalmic application, the aim was to predict age-related macular degeneration (AMD) treatment failure as early as possible using the patient's longitudinal history. The second aim was to predict patients' mortality or their need for a liver transplant within the next five years in patients with Primary Biliary Cirrhosis.

In this chapter, I describe the methodology related to discriminant analysis using longitudinal data which will be used in this thesis.

2.1 Introduction

Many recent statistical methodological studies have focused on the further development of longitudinal discriminant analysis (LoDA) to be used for classification purposes. In particular, since a large amount of longitudinal information can be collected for multiple markers over time for each patient clinicians may want to use this extra data to predict the future status of a patient. This type of data has multiple correlations. For each patient, the repeated measurements of a marker are correlated and multiple markers observed at a particular time point, or even different time points, may be correlated. Further, different types of longitudinal markers can be collected including continuous,

binary or count variables.

The structure of this chapter is as follows. First, I describe the classical linear and quadratic discriminant analyses (Section 2.2). In the same section, a variety of test statistics for homogeneity of variance-covariance matrices are presented. In Section 2.3, I review four approaches of longitudinal discriminant analysis: modified, marginal, conditional and random-effects approaches. These four approaches follow different ways to estimate the parameters which will be used to generate the classification model. The modified approach is based on the multivariate linear mixed effect model where maximum likelihood (ML) is used for estimating the parameters. The marginal, conditional and the random effects approaches are based on multivariate generalised linear mixed models with a mixture of normal distributions assumed for the random effects. A Bayesian approach is used to estimate the parameters. In Section 2.4, I describe the tools used to assess the accuracy of predictions made using the discriminant analysis approaches described in this chapter. I summarise Chapter 2 in Section 2.5.

2.2 Classical discriminant analysis

Huberty and Olejnik (2006) described two forms of discriminant analysis (DA) based on the research question: Descriptive discriminant analysis and predictive discriminant analysis. Descriptive discriminant analysis (DDA) is used to describe how the groups differ. Applications of DDA have been viewed as a follow-up to a multivariate analysis of variance (MANOVA) (Huberty and Wisenbaker, 1992). Predictive discriminant analysis (PDA) is usually used for classification purposes. In classical discriminant analysis, cross-sectional data is often used with data collected at a single time point per patient.

In diagnostic medical studies, discriminant analysis can be used to assess the ability of biomarkers to classify patients into different groups. When markers are measured at a single time for each patient, classical discriminant analysis offers an approach to predict patient's group membership. In this thesis, classical discriminant analysis will be

applied using longitudinal data by considering each time point as a separate variable. However, note that this does not take into account correlation between repeated measurements and also requires that patients have a complete record of clinic visits that are all arranged at the same time points, which I will refer to as a balanced dataset.

Classical discriminant analysis has been applied to some longitudinal data applications. A comprehensive review of discriminant analysis using repeated measures data can be found in Lix and Sajobi (2010). One of the recent examples of using classical discriminant analysis is presented by Coster et al. (2005) who developed a classification rule for first-stroke patients. They developed a linear discriminant model which involved information from 206 patients of 17-items of the Hamilton Depression Rating Scale (HAM-D) at 1, 3, 6, 9, 12 months. One drawback of the study is that they did a complete data analysis (i.e., they removed patients for whom at least one measurement was missing) which led to the removal of 10% of patients.

Another example is presented by Rietveld et al. (2000) who were interested in classifying zygosis diagnosis in twins. Data from 691 twins zygosity were collected at ages 6, 8, 10 years by both parents. However, nearly 53% (367) and 60% (412) of twin pairs information from mother and father, respectively were missing. Again, data from the children with missing data were not used in the analysis.

Levesque et al. (2008) were interested in classifying husband caregivers caring for their wives, into three groups of psychological distress based on variation in psychological distress. They used two measurements at baseline and changes over time to build classical linear discriminant analysis. In total 91 of 323 individuals were excluded from the study.

In summary, the applications of classical discriminant analysis on longitudinal data often leads to the removal of patients with incomplete observations. This procedure is not optimal, as not all data are used for the analyses and it is not clear how to perform the predictions for the patients with missing data.

Discriminant analysis has been extended by Cochran et al. (1948) to include covari-

ates. Then, Tomasko et al. (1999) developed an approach for classical linear discriminant analysis that can be used with incomplete datasets. The idea of their approach is that a linear mixed model is first used to estimate the group means and the covariance matrices which are then used in classical LDA equations. At the same time, Marshall and Barón (2000) combined the traditional discriminant analysis and linear/non-linear mixed models to allow for unbalanced data (i.e., patients can have a different number of observations at different time points).

Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are commonly used for clinical classification. LDA is designed for cases where the variance-covariance matrices are the same across groups. However, such an assumption is difficult to assess in practice. Meanwhile, QDA is considered as a generalisation of LDA, since it allows each class to have a class-specific covariance matrix (Marks and Dunn (1974), Flury and Schmid (1992)). Like LDA, QDA assumes that the observations for each group are sampled from the multivariate normal distribution.

2.2.1 Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) was first introduced in 1936 by Fisher (1936). His idea was to obtain a linear combination of the predictor variables $\mathbf{X}$ that best separate the groups of observations. This combination is called the discriminant function and is represented by:

$$d_{ig}(x) = \mathbf{a}_{1g}\mathbf{x}_{1i} + \mathbf{a}_{2g}\mathbf{x}_{2i} + \cdots + \mathbf{a}_{pg}\mathbf{x}_{pi} + c_g = \mathbf{a}_g'\mathbf{X}_i + c_g \tag{2.1}$$

where the vector of weights is

$$\mathbf{a}_g' = \mu_g'\Sigma^{-1}$$

and constant

$$c_g = -\frac{1}{2}\mu_g'\Sigma^{-1}\mu_g + \log \pi_g$$

where $d_{ig}(x)$ is a discriminant score for group $g, (g = 1, \ldots, G)$, μ_g is a mean vector of group g, Σ is common covariance matrix, and $\mathbf{X}_i$ is the observation vector of the i-th individual.

Linear discriminant analysis assumes that the data are normally distributed. LDA is designed for cases where the variance-covariance matrices are the same across groups. Assume that there are G different groups, and for each $\mathbf{X}$ is supposed to follow a multivariate normal distribution with group-specific mean vector μ_g (where $g = 1, \ldots, G$) and common covariance matrix Σ. Also, suppose that the aim is to classify an observation into one of G groups or classes ($G \geqslant 2$). π_g is defined as prior probability of the gth group. Prior probabilities (denoted by π_g) are often estimated using the proportion of observations in each group to the total (such that, $\pi_g = \frac{n_g}{N}$ where $N = \sum_{g=1}^{G} n_g$ and $\sum_{g=1}^{G} \pi_g = 1$). Usually the population means and covariance matrices are unknown and the maximum likelihood is used to estimate them.

The linear discriminant function (LDF) also known as linear classification function can be expressed by:

$$d_g(x) = x'\Sigma^{-1}\mu_g - \frac{1}{2}\,\mu_g'\Sigma^{-1}\mu_g + \log \pi_g. \tag{2.2}$$

The covariance matrix and mean for each group can be estimated from the data sample using the sample mean $\bar{x}$ and covariance matrix S_g. A new subject with $X = x$ is allocated to class g, if the following decision rule applies:

$$d_g > d_l \tag{2.3}$$

otherwise x is assigned to group l. The decision rule in Equation 2.3 is known as the linear classification rule. In the binary case (G = 2), two linear discriminant functions are built as follows:

$$\begin{aligned} d_1(x) &= x'\Sigma^{-1}\mu_1 - \frac{1}{2}\,\mu_1'\Sigma^{-1}\mu_1 + \log \pi_1 \\ d_2(x) &= x'\Sigma^{-1}\mu_2 - \frac{1}{2}\,\mu_2'\Sigma^{-1}\mu_2 + \log \pi_2 \end{aligned} \tag{2.4}$$

These two discriminant functions can be combined by subtracting $d_2(x)$ from $d_1(x)$ as:

$$d(x) = x'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2) + \log\frac{\pi_1}{\pi_2}. \tag{2.5}$$

such that if $d(x) > 0$, the observation x will be allocated to group 1, otherwise to group 2. The last two parts in the Equation (2.5) $(-\frac{1}{2}(\mu_1 + \mu_2)'\Sigma^{-1}(\mu_1 - \mu_2), \log\frac{\pi_1}{\pi_2})$ are constant given a dataset.

In this thesis, a patient is classified into prognostic group 1 if the group membership probability is larger than c, where c is a chosen cutoff value that selected by analysis of the receiver operating characteristic (ROC) curve.

2.2.2 Quadratic discriminant analysis (QDA)

Similar to LDA, QDA assumes the observations for each group are sampled from a multivariate normal distribution. However, in QDA the covariance matrices can be different across groups (Marks and Dunn (1974), Flury and Schmid (1992)). As indicated above, the covariance matrix in each class can be estimated from the data sample using the sample covariance matrix S_g, $g = 1, \ldots, G$.
In particular, the quadratic discriminant function is expressed as:

$$\begin{aligned} d_g(x) &= -\frac{1}{2}(x - \mu_g)'\Sigma_g^{-1}(x - \mu_g) - \frac{1}{2}\log|\Sigma_g| + \log\pi_g \\ &= -\frac{1}{2}x'\Sigma_g^{-1}x + x'\Sigma_g^{-1}\mu_g - \frac{1}{2}\mu_g'\Sigma_g^{-1}\mu_g - \frac{1}{2}\log|\Sigma_g| + \log|\pi_g| \end{aligned} \tag{2.6}$$

where Σ_g is covariance matrix for class g and π_g is the prior probability for group g.

For classification purposes, the same approach that is used in LDA to allocate patients can be used in QDA (i.e., a patient is allocated into a prognostic group 1 if the group membership probability of the group one is larger than the chosen cutoff value c). James et al. (2013) recommended that if the dataset is large, the QDA is a better choice than LDA.

In both LDA and QDA, the decision boundaries are functions of the parameters of

the densities. If the number of variables p increases then the number of parameters can be increased considerably. The number of parameters required by LDA is less than that required by QDA. While for LDA there are $(G-1)\times(p+1)$ parameters (where p is the number of parameters), QDA requires $(G-1)\times\{p\,(p+3)/2+1\}$ parameters (Hastie et al. (2009)). Regarding the covariance matrices, LDA assumes a common covariance matrix that requires $\frac{p\,(p+1)}{2}$ parameters. However, the QDA assumes that each group has a specific covariance matrix such that this requires $G\times\frac{p\,(p+1)}{2}$ parameters to be estimated.

If the assumption of a common covariance matrix is not met, a different covariance matrix for each group is estimated. This leads to QDA where the discriminating boundaries are not straight lines, but quadratic curves. Box's M test (described in the next section) is used to test the homogeneity of variance-covariance matrices (Box (1949), Geisser and Greenhouse (1958)). However, Bouveyron et al. (2007) pointed out that QDA does not guarantee an improved classification rate even when the test is significant.

2.2.3 Test for homogeneity of variance-covariance matrices

In this section, a range of methods that are used to test for homogeneity of covariance matrices in order to decide whether QDA is superior to LDA are presented.

Box's M test

Box's M statistic is usually used in discriminant analysis to decide whether LDA or QDA should be used. Box (1949) suggested a test based on the likelihood-ratio test (LRT) statistic for testing the hypothesis of equal covariance matrices. In particular, it assumes both χ^2- and F-approximations for the distribution of the LRT statistic M under the assumption of multivariate normality (see Rencher (1998)). The null hypothesis of the test for homogeneity of covariance matrices is $H_0 : \Sigma_1 = \Sigma_2 = \cdots =$

Σ_G. The test statistic can be given as follows:

$$M = \gamma \sum_{g=1}^{G} (n_g - 1)\, log|\Sigma_g^{-1}\Sigma|, \tag{2.7}$$

where $\gamma = 1 - \dfrac{2p^2 + 3p - 1}{6(p+1)(G-1)} \left(\sum \frac{1}{n_g - 1} - \frac{1}{N-G}\right)$, $N = \sum n_g$ is the total sample size and Σ is the pooled covariance matrix that is estimated as $\Sigma = \dfrac{\sum_{g=1}^{G}(n_g - 1)\,\Sigma_g}{(N-G)}$. Box's M statistic has an asymptotic χ^2 distribution with degrees of freedom $df = \frac{p\,(p+1)\,(G-1)}{2}$. Mardia et al. (1979) pointed out that the Box's approximation (χ^2) is expected to work well when $N > 20$, and if G and $p < 5$. Otherwise, the F-approximation is more accurate.

If Box's M test is statistically significant at a significant level $\boldsymbol{\alpha}$, the covariance matrices can be regarded as not homogeneous.

Bartlett's test

Bartlett's test Bartlett (1937) is designed to assess the assumption that the equality of variances across groups holds $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_G^2$ against the alternative that variances are unequal for at least two groups. Bartlett's test is based on a chi-square statistic with $(G-1)$ degrees of freedom, where G is the number of groups. The test statistic can be expressed as follows:

$$T = \frac{(N-G)\, ln(\sigma^2) - \sum_{g=1}^{G}(n_g - 1)\, ln(\sigma_g^2)}{1 + \dfrac{1}{3(G-1)} \left(\sum_{g=1}^{G} \dfrac{1}{n_g - 1} - \dfrac{1}{N-G} \right)} \tag{2.8}$$

where σ_g^2 indicates the variance of group g, $\sigma^2 = \dfrac{\sum_{g=1}^{G}(n_g - 1)\,\sigma_g^2}{(N-G)}$ is pooled variance and n is defined as in Box's M test. The T statistic is distributed as a χ^2 distribution with $(G-1)$ degrees of freedom under the null hypothesis of equality of variances.

Bartlett's test is unbiased for any sample sizes Pitman (1939). However, it is sensitive to departures from normality. If the null hypothesis H_0 is rejected, this could be due to the variances of the two groups being unequal, or the samples of two groups not

following a normal distribution or both.

Levene's test

Levene's test is another approach that is used to test for homogeneity of variance across groups (Levene (1960)). The Levene's test is based on F distribution with $(G-1)$ and $(N-G)$ degrees of freedom. The test statistic is given as:

$$F = \frac{(N-G)\sum_{g=1}^{G} n_g(Z_g - Z_{..})^2}{(G-1)\sum_{g=1}^{G}\sum_{i=1}^{n_g}(Z_{gi} - Z_g)^2} \tag{2.9}$$

where Z_{gi} are the absolute deviations ($|X_{gi} - \bar{X}_{g.}|$) where $\bar{X}_{g.}$ is the mean of group g, $Z_{..} = \frac{1}{N}\sum_{g=1}^{G}\sum_{i=1}^{n_g} Z_{gi}$ is the mean of all the absolute deviations (Z_{gi}), $Z_g = \frac{1}{n_g}\sum_{i=1}^{n_g} Z_{gi}$ is the mean of the absolute deviations (Z_{gi}) for group g. The test statistic F is approximately F-distributed with $(G-1)$ and $(N-G)$ degrees of freedom.

Brown and Forsythe's test

Brown and Forsythe (1974) extended the Levene's test by using the median or a trimmed mean instead of the mean when calculating the deviations within each group. The trimmed mean is a method that calculates the average by excluding a small percentage (e.g., trim 10%) of the largest and smallest values of the data. A trimmed mean aims to reduce the impact of statistical outliers. In the above Equation 2.9, the $Z_{gi} = |X_{gi} - \tilde{X}_{g.}|$ where $\tilde{X}_{g.}$ is the location of the median or trimmed mean of a variable in a group g. This extension makes the test more robust to deviation from normality. A general discussion can be found in Gastwirth et al. (2009).

The tests mentioned above were used in Chapter 3 to decide whether QDA is preferable to LDA. Since the test of homogeneity of variances-covariances matrices is not the focus of this thesis, this aspect will not be discussed further.

2.2.4 Classification rule

Classification methods of discriminant analysis are used to predict the future diagnosis of new patients by assigning them into one of G prognostic groups. Let $f_g(x)$ represent the density function of X for observations that belong to the g-th group, $f_g(x) = p(X = x|G = g)$.

The posterior probabilities based on Bayes' rule are defined as:

$$p(G = g|X = x) = \frac{\pi_g f_g(x)}{\sum_{g=1}^{G} \pi_g f_g(x)} \tag{2.10}$$

where the multivariate normal density function can be written as:

$$f_g(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\, exp\Big(-\frac{1}{2}(x-\mu_g)'\Sigma^{-1}(x-\mu_g)\Big). \tag{2.11}$$

The group membership probabilities ($p_g(x)$) of LDA and QDA for an individual can be calculated using Equation 2.10. Then, the patient is assigned to the group if the probability is the maximal across all groups and if it is larger than a defined cutoff c (optimal cutoff).

2.3 Longitudinal discriminant analysis (LoDA) approaches

In recent years, longitudinal discriminant analysis has been used for classification by researchers (e.g., Marshall et al. (2009), Kohlmann et al. (2009), Komárek et al. (2010), Hughes et al. (2018b)). Methods that can be used to develop a discriminant tool using longitudinal data are the covariance pattern and mixed-effects models (Lix and Sajobi, 2010). However, in this thesis, only mixed models will be considered, as mixed models are most frequently used in longitudinal discriminant analysis.

The longitudinal study generally yields multiple measurements (or variables) on each subject. These variables are also referred to as markers. Those markers may be of different types, such as continuous, binary or counts.

Brant et al. (2003) used a single continuous longitudinal marker of the prostate-specific antigen (PSA) of 342 males who were diagnosed as prostate cancer-free at the first scan and were followed up at least ten years to classify them into four prognostic groups. Further application with a single continuous response had been presented by Wernecke et al. (2004). Marshall and Barón (2000) used classical discriminant approaches based on a linear/non-linear mixed model to classify pregnant women into normal or abnormal pregnancy.

In 2015, Arribas-Gil et al. (2015) introduced a classification method to predict normal and abnormal pregnancy using a semi-parametric linear mixed-effects model (SLMM) for the longitudinal data of each group (hormone levels measured at early stages of pregnancy). The authors proposed a unified procedure of estimation based on a penalised algorithm, involving a LASSO approach (least absolute shrinkage and selection operator) for the estimation of the model parameters. Another method used to analyse this pregnancy dataset was proposed by De la Cruz et al. (2017). They suggest using penalised splines within a semiparametric non-linear mixed-effects model (SNMM) for the longitudinal data. Rubin et al. (2017) introduced a joint logistic regression and Markov chain model based on continuous-time Markov chains (CTMCs) to model a cross-sectional binary outcome as a function of a longitudinal covariate process. They applied their approach to a dataset of patients with traumatic brain injury to predict a 6-month outcome (as favorable (good recovery and moderate disability) or unfavorable (severe disability, vegetative, or dead)) based on baseline data and physiological information collected every hour after injury.

Li and Gatsonis (2019) developed methods that combine multiple biomarker trajectories to achieve a composite diagnostic marker using functional data analysis (FDA). Their approach was applied to data from the Religious Orders Study and a non-small cell lung cancer trial to distinguish diseased from non-diseased patients. Lukasiewicz et al. (2011) introduced an approach based on longitudinal discriminant analysis using multiple biomarkers and partial area under the receiver operating characteristic (ROC) curve (pAUC) to predict non-response to treatment for hepatitis C virus. They also

used the partial area under the ROC curve index to evaluate the performance of their approach. Kim and Kong (2016*b*) propose a linear combination of longitudinal measurements to improve classification accuracy. These three approaches all contributed to the classification of longitudinal data literature by proposing methods where data from multiple biomarkers are incorporated into a single score to be used in a classification algorithm.

Multivariate longitudinal data using pseudo-likelihood methods for estimation was presented in Bruckers et al. (2016) in the context of cluster analysis. They fitted joint mixed-effects models and used ideas from k-means clustering to reveal homogenous subgroups. Their algorithm was applied to electro-encephalogram (EEG) data. Reddy et al. (2016) proposed an approach that involves a linear mixed model to fit the longitudinal measurements as the first step. In the second step, they estimated the time to attainment of two consecutive measurements less than a meaningful threshold, which takes into account the patient-specific trajectory and measurement error. De la Cruz et al. (2018) applied longitudinal discriminant analysis based on a non-linear mixed model to predict normal versus abnormal pregnancy outcomes. They compared the misclassification error rates by using several methods such as cross-validation and bootstrap algorithms.

In some clinical studies, multiple longitudinal variables are recorded, and this extra information can be used in the discriminant analysis to improve the predictive accuracy of the discriminant model. A comprehensive review of multivariate longitudinal data analysis based on mixed models can be found in Verbeke et al. (2014) and Bandyopadhyay et al. (2011). Marshall et al. (2009) used multiple continuous responses for classification based on multivariate non-linear mixed models. Moreover, multiple continuous markers had also been used by Komárek et al. (2010) to predict patients with primary biliary cirrhosis by fitting a multivariate linear mixed model to the Dutch Multicenter Primary Biliary Cirrhosis data. Morrell et al. (2005, 2012) also used discriminant analysis with three continuous markers to predict the development of cancer.

The applications described above involve multivariate linear/non-linear mixed-effects models with continuous multiple markers. However, fewer developments have been produced for different types of markers. An extension of the multivariate longitudinal data analysis to allow continuous and binary longitudinal markers were proposed by Fieuws et al. (2008). In their paper, they used a combination of linear, nonlinear and generalised linear mixed-effects models to predict renal graft failure in renal transplantation patients. Univariate mixed models were combined in a multivariate mixed model by specifying a joint distribution for the random effects. They used a pairwise approach presented by Fieuws and Verbeke (2006) to fit the mixed models. Moreover, Hughes et al. (2018b) developed a multivariate longitudinal discriminant analysis approach to allow longitudinal markers of three different type (continuous, binary and counts) and this was applied to identify patients with epilepsy who do not benefit from antiepileptic drugs. Their work provided additional flexibility by using a mixture distribution for the random effects.

As mentioned in the previous section, classical discriminant analysis can be applied to longitudinal data with a balanced dataset (where time points are identical for all patients). Tomasko et al. (1999) applied classical discriminant analysis using a mixed model to analyse unbalanced longitudinal data. The basic idea of their modified approach (the name they used in their paper) is first to generate the mixed model for each group to describe the change of a single marker over time, and then construct the quadratic discriminant function to predict the outcome. Details of the modified discriminant analysis are given in next section.

LoDA methods attempt to use a patient's clinical history to predict the future status of a patient. Three different approaches were proposed by Morrell et al. (2007) to classify patients using their longitudinal clinical records namely: marginal prediction, conditional prediction and random-effects prediction using the posterior probabilities.

Similar to the modified approach, the LoDA marginal, conditional and random-effects approaches are based on the mixed model. However, the modified approach uses the mixed model to directly estimate the means and covariance matrices that

are subsequently used in the classical discriminant analysis to predict a patient status. Meanwhile, the marginal, conditional and random-effects used the distribution of the marginal, conditional or random effects (respectively) of the mixed models for prediction.

These three approaches estimate a patient's posterior group membership probabilities differently. The marginal prediction focuses on the mean development of the markers over time. This prediction uses the marginal distribution of the new patient's observed longitudinal data for prediction. The conditional prediction is based on the patient-specific development of markers over time but does not take error in the variability of the patient's estimated random effects into account. In this case, the conditional density of the new patient's longitudinal data is used to predict their prognostic group. The conditional approach estimates conditional profiles of a new patient, given an estimate of their patient-specific deviations from the average longitudinal profile, then compares it with the overall mean longitudinal profiles of patient's with similar estimated random effects in each group. Finally, the random-effects prediction approach focuses on the patient-specific development of the markers, where the density of the patient's estimated random effects is utilised to predict patients status. These approaches have been compared by researchers such as Morrell et al. (2007, 2011), Komárek et al. (2010), Hughes et al. (2018b).

Comparisons of the marginal, conditional and random-effects approaches have been explored using particular datasets. For example, Morrell et al. (2011) used data from the Baltimore Longitudinal Study of Aging to predict prostate cancer of future patients by using the three approaches. Their comparison was based on a number of PSA measurements which concluded that the random-effects prediction approach provided the best prediction in terms of specificity and efficiency when the optimal cutoff value was used, while when the cutoff was 0.5 the marginal prediction gave the highest sensitivity and lead-time (time before patients were identified to the cancer group correctly). Komárek et al. (2010) compared the three approaches to predict the future prognosis of patients with Primary Biliary Cirrhosis and concluded that the random-

effects approach outperformed the other two approaches. Hughes et al. (2018b) applied the three approaches to classify 1752 patients with epilepsy who do not benefit from antiepileptic drugs using data from the Standard and New Antiepileptic Drugs study. In their comparison, the random-effects approach was inferior when compared to the marginal and conditional approaches in terms of the probability of correct classification (PCC).

2.3.1 Modified discriminant analysis

Discriminant analysis based on mixed models was introduced by Tomasko et al. (1999) who provided a discriminant analysis approach that was able to deal with missing data. Data from a large number of patients could potentially be excluded from the analysis if some of their values are missing. The other benefit of applying this modified approach is that the random subject effect can be embedded in the covariance matrix model. In other words, the mixed models allow modelling of the relationship between measurements collected for the same patient at different time points of the same marker and also among markers.

The modified discriminant analysis proceeds in two steps. In the first step, the linear mixed model is used to estimate the parameters (means and covariance matrices) using maximum likelihood (ML). In the second step, these parameters are used to build the discriminant function that will be used to classify individuals. In other words, the means and covariance matrices are generated from the mixed model, and then used to build the linear discriminant tool (Equation 2.2 or 2.5) or the quadratic discriminant tool (Equation 2.6).

In the modified prediction, the Expectation-Maximization (EM) algorithm (Laird et al. (1982)) is used to get the maximum likelihood parameters of the multivariate mixed model (see Tomasko et al. (1999); Marshall and Barón (2000) for more detail). Also, I consider a single normal distribution (i.e., $K = 1$, where K is number of mixture components of the random effects distribution, since the number of mixture components

K is assumed to be known and to be the same across groups) of the random effects.

The next section, consider a situation where there is multiple markers by using a multivariate generalised linear mixed-effects model.

2.3.2 Mixture multivariate generalised linear mixed model (MMGLMM)

A linear mixed-effects model to analyse longitudinal data that takes the correlation between the repeated measurements obtained from the same patient into account was considered by Laird et al. (1982) who used a random-effects approach. A combination of empirical Bayes and maximum likelihood were used to estimate the model using the EM algorithm.

Tomasko et al. (1999) applied their approach to incomplete data by using a univariate mixed model. Then, Roy (2006) modified the discriminant analysis approach to allow for the multivariate linear mixed models (MLMMs) and assumed a Kronecker product structure for the residual errors covariance matrix. Reinsel (1984) fitted the multivariate random-effects model for continuous outcomes to complete and balanced multivariate longitudinal designs in which all subjects have observations at the same time points. However, Shah et al. (1997) extended the work of Laird et al. (1982) from univariate longitudinal data to multivariate longitudinal data, and discussed the problem of unbalanced design by applying an extension of the EM algorithm method. Their approach dealt with unbalanced multivariate longitudinal designs and assumes that the relationship between different responses from the same subject is correlated.

The LoDA procedure as described in Hughes et al. (2018b) is based on a multivariate generalised linear mixed model with a normal mixture in the random effects distribution. This is the approach I describe in this thesis, although the other models I have described follow a similar framework.

Suppose that for each patient there are r markers (where $r = 1, \ldots, R$) measured at times $t_r = (t_{r,1}, \ldots, t_{r,n_r})$ in each prognostic group $g = 1, \ldots, G$. Let $\mathbf{Y}_i =$

$[\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{ir}]$ be the response matrix for individual i where $\mathbf{y}_{ir}$ is an $n_i \times 1$ response vector for marker r. Additional covariate vectors $\mathbf{v}_{r,1}, \ldots, \mathbf{v}_{r,n_r} \in \mathbb{R}^{p_r}$ which are denoted as $\mathcal{C}$ could available for longitudinal evolution of each marker. Let $\mathbf{y}_i = Vec(\mathbf{Y}_i)$ be a $rn_i \times 1$ vector of all response variables for individual i. In the same way, suppose the error term matrix is $\mathbf{E}_i = [\mathbf{e}_{i1}, \mathbf{e}_{i2}, \ldots, \mathbf{e}_{ir}]$ and $\mathbf{e}_i = Vec(\mathbf{E}_i)$ be a $rn_i \times 1$ vector of error terms. Also, $\mathbf{e}_i \sim N(0, \Sigma_i)$ where Σ_i is a $rn_i \times rn_i$ block-diagonal covariance matrix of the error term, where in each diagonal block equal to $\sigma_r^2 I_{rn_i}$, σ_r^2 is a r-th marker residual error. Moreover, it is assumed that observations across patients are independent and errors are independent for given patient ($cov(\mathbf{e}_i, \mathbf{b}_i) = 0$, where $i \neq i'$). MGLMMs are fitted to the longitudinal data for each group separately where the distribution of marker r-th belongs to an exponential family (e.g. normal, exponential, Poisson) for j'th longitudinal observation ($j = 1, \cdots, n_r$) is given by:

$$h_r^{-1}\Big\{\mathsf{E}\big(Y_{r,j} \,\big|\, \mathbf{b},\, g\big)\Big\} = \mathbf{x}_{r,j}^{g\top}\boldsymbol{\alpha}_r^g + \mathbf{z}_{r,j}^{g\top}\mathbf{b}_r, \qquad r = 1, \ldots, R,\ j = 1, \ldots, n_r, \tag{2.12}$$

where h_r^{-1} is a chosen link function (with possible dispersion parameters ϕ_r^g), covariate vectors $\mathbf{x}_{r,j}^g = \mathbf{x}_{r,j}^g(\mathcal{C})$ and $\mathbf{z}_{r,j}^g = \mathbf{z}_{r,j}^g(\mathcal{C})$ are used in a model for the group g and $\boldsymbol{\alpha}_r^g$, $r = 1, \ldots, R$, $g = 1, \ldots, G$ indicates unknown regression coefficients. Correlation between repeated observations of the same marker and between values of different markers on the same patients are accounted for by the formation of the random effects vector $\mathbf{b} = \big(\mathbf{b}_1, \ldots, \mathbf{b}_R\big)$. Normally, it is supposed that the random effects vector follows a normal distribution. However, Verbeke and Lesaffre (1996) suggested relaxing the normality assumption of the random effects distribution by introducing a univariate linear mixed-effects model with a normal mixture in the random effects distribution. Also, Komárek et al. (2010) and Hughes et al. (2018b) provided extra flexibility to the joint distribution of the random effects vector in each prognostic group by defining a normal mixture for the random effects distributions. Assuming that a multivariate normal mixture in the distribution of random effects is:

$$\mathbf{b} \,|\, g \;\sim\; \sum_{k=1}^{K} w_k^g\, \mathcal{MVN}\big(\boldsymbol{\mu}_k^g,\, \mathbb{D}_k^g\big), \tag{2.13}$$

with the mean vector $\boldsymbol{\mu}_k^g$ and a covariance matrix $\mathbb{D}_k^g$. Where $w_k, (k = 1, \ldots, K, 0 < w_k^g < 1, \sum_{k=1}^{K^g} w_k^g = 1)$ is a vector of weights for the mixture distributions. The number of mixture components K is assumed to be known and to be the same across groups. A density function of the multivariate normal distribution is indicated as $\varphi(\cdot; \boldsymbol{\mu}_k^g, \mathbb{D}_k^g)$. Let $\boldsymbol{\psi}^g := \left(\boldsymbol{\alpha}_1^g, \ldots, \boldsymbol{\alpha}_R^g, \phi_1^g, \ldots, \phi_R^g\right)$ indicate a vector of unknown parameters of the GLMM model (2.12) and additionally $\boldsymbol{\theta}^g := \left(\mathbf{w}^g, \boldsymbol{\mu}_1^g, \ldots, \boldsymbol{\mu}_{K^g}^g, \mathbb{D}_1^g, \ldots, \mathbb{D}_{K^g}^g\right)$ is a vector of mixture parameters (from 2.13) in the distribution of random effects in group g.

2.3.3 Bayesian method and Markov chain Monte Carlo (MCMC) method

Bayesian methods is a term which refers to any mathematical tools that are associated in some way to Bayesian inference. Bayesian inference considers model parameters as random variables, in contrast to the classical inference which deals with parameters as constants (Puza (2015)).

The Bayesian approach has been used in the analysis of longitudinal data to estimate parameters. Brown et al. (2001) proposed a Bayesian approach to classifying multivariate longitudinal data, that were only measured at fixed time points, and used an expectation-conditional maximization algorithm within a Bayesian Gaussian discrimination model. De La Cruz-Mesia and Quintana (2006) introduced a general Bayesian framework for the classification of unbalanced longitudinal data (where the number or timing of the observations differs) and used Markov chain Monte Carlo methods to estimate the parameters.

The Bayesian approach with MCMC estimation was used in a multivariate linear mixed model with mixture models in random-effects distribution by Komárek et al. (2010). The MCMC approach is a beneficial method which can deal with problems involving hierarchically specified models (for example linear mixed models).

2.3.4 Estimation

Markov Chain Monte Carlo is used to estimate the model parameters. I assume that the mixture associated parameters θ and the GLMM associated parameters ψ are independent. The prior distribution $p(\psi, \theta)$ is based on the factorisation $p(\psi)$ and $p(\theta)$. This factorization is standard in generalized linear mixed models.

Prior distributions

In order to perform an MCMC procedure to estimate the parameters of a model, I must first specify prior distributions for the parameters in my model. In Bayesian statistical inference, a prior probability is a distribution specified in advance reflecting the users beliefs about the possible values of a particular parameter. This is then combined with collected data to derive a posterior distribution, which reflects the possible values of a parameter once both, prior beliefs, and knowledge from data have been combined. To obtain the weakly informative prior distribution, I follow the suggestions of Richardson and Green (1997) and Komárek and Komárková (2013) for the choices of the fixed hyperparameters. I have to set some initial values of the random effects in GLMM by using maximum-likelihood estimates in each of markers of GLMM models separately assuming that the distribution of random effects is normal. Moreover, let for each marker, $b^0_{i,r}, i = 1, \ldots, N, r = 1, \ldots, R$ be empirical Bayes estimates of individual values of random effects in the rth ML estimated GLMM assuming that the random effects follow the normal distribution. The prior distributions for the model parameter (ψ, θ, B, k) are specified so as to be weakly informative as follows. The random effects prior distribution $p(b_i|\theta, k)$ is a (multivariate) normal distribution $N(\mu_k, \mathbb{D}_k)$, which is

$$p(b_i|\theta, k) \propto |\mathbb{D}_k|^{-\frac{1}{2}} exp\{-\frac{1}{2}(b_i - \mu_k)'\mathbb{D}_k^{-1}(b_i - \mu_k)\}.$$

The prior distribution for the component allocations for each mixture component in the assumed random effects distribution, $p(k_i|w), i = 1, \ldots, N$, is assumed to be a discrete distribution with parameter w_k

$$P(k_i = k|w) = w_k, \;\; k = 1, \ldots, K,$$

with a Dirichlet distribution $D(\delta, \cdots, \delta)$, as the prior for the mixture weights, where δ is a fixed hyperparameter. That is,

$$p(w) = \frac{\Gamma(K\delta)}{\Gamma^k(\delta)} \prod_{k=1}^{K} w_k^{\delta-1},$$

where δ set to be a small positive number to obtain weakly informative prior, e.g., $\delta = 1$. The prior for the means of the mixture components, $p(\mu_k), k = 1, \ldots, K$, is a (multivariate) normal distribution

$$p(\mu_k) \propto |C_b|^{-\frac{1}{2}} exp\{-\frac{1}{2}(\mu_k - \xi_b)' C_b^{-1}(\mu_k - \xi_b)\},$$

where C_b and ξ_b are fixed prior mean and covariance matrix, respectively.
The prior distribution for the random effects mixture covariance matrices, $p(\mathbb{D}_k^{-1}|\gamma_b)$ follow the Wishart distribution, which can be expressed as

$$p(\mathbb{D}_k^{-1}|\gamma_b) \propto |\Xi_b|^{\frac{-\zeta_b}{2}} |\mathbb{D}_k^{-1}|^{\frac{\zeta_b - d - 1}{2}} exp\{-\frac{1}{2} tr(\Xi_b^{-1}\mathbb{D}_k^{-1})\},$$

where ζ_b are fixed prior degrees of freedom and set to be a small positive number, e.g., $\zeta_b = d + 1$ to get a weakly informative prior. $\Xi_b = diag(\gamma_b)$ is a diagonal scale matrix with random diagonal components, and γ follows a gamma distribution as

$$p(\gamma_b^{-1}) = \prod_{l=1}^{d} p(\gamma_{b,l}^{-1}) = \prod_{l=1}^{d} \{\frac{h_{b,l}^{gb,l}}{\Gamma(g_{b,l})} (\gamma_{b,l}^{-1})^{gb,l-1} exp(-h_{b,l}\gamma_{b,l}^{-1})\},$$

where $h_{b,l}$ and $g_{b,l}, l = 1, \ldots, d$ are fixed hyperparameters. To obtain weakly informative prior distribution as recommended by Richardson and Green (1997), $g_{b,l}$ set to be a small positive number and $h_{b,l} = \frac{1}{R^2}$ where R is a range of residuals from initial maximum-likelihood fits for each marker. Prior for GLMM dispersion parameters $p(\phi_r^{-1}|\gamma_{\phi,r}), r = 1, \ldots, R$ which ϕ_r in my case is residual variance σ_r^2 follows gamma distribution as

$$p(\phi_r^{-1}|\gamma_{\phi,r}) = \frac{2^{\frac{-\zeta_{\phi,r}}{2}}}{\Gamma(\frac{\zeta_{\phi,r}}{2})} \gamma_{\phi,r}^{\frac{-\zeta_{\phi,r}}{2}} (\phi_r^{-1})^{\frac{\zeta_{\phi,r}-2}{2}} exp\{\frac{-1}{2}\gamma_{\phi,r}^{-1}\phi_r^{-1}\}.$$

To get weakly informative prior, $\zeta_{\phi,r}$ set to be a small positive number, e.g., $\zeta_{\phi,r} = 2$. Fixed effects prior distribution $p(\alpha)$ is a (multivariate) normal distribution which is

$$p(\alpha) \propto |C_\alpha|^{-\frac{1}{2}} exp\{-\frac{1}{2}(\alpha - \xi_\alpha)' C_\alpha^{-1}(\alpha - \xi_\alpha)\},$$

where ξ_α and C_α are mean and variance fixed prior hyperparameters. To obtain weakly informative for fixed effects α, the mean ξ_α set to be zero, and the variance C_α set to be a large positive number (e.g., 10,000).

The joint prior distribution for the MLMM model is

$$\begin{aligned} p(\psi, \theta, B, k, \gamma_b, \gamma_\phi) &= p(B|\theta, k) \times p(k|\theta) \times p(\theta|\gamma_b) \times p(\gamma_b^{-1}) \times p(\psi|\gamma_\phi) \times p(\gamma_\theta^{-1}) \\ &= \prod_{i=1}^{N} \{p(b_i|\theta, k) \times p(k_i|w)\} \times p(w) \times \prod_{k=1}^{K} \{p(\mu_k) \times p(\mathbb{D}_k^{-1}|\gamma_b)\} \times \\ &\prod_{l=1}^{d} p(\gamma_{b,l}^{-1}) \times \prod_{r=1}^{R} \{p(\theta_r^{-1}|\gamma_{\theta,r}) \times p(\gamma_{\phi,r}^{-1})\} \times p(\alpha). \end{aligned}$$

The likelihood function of the multivariate mixture GLMM (Equation 2.12) is

$$\mathbf{L}_g(\boldsymbol{\psi}^g, \boldsymbol{\theta}^g) = \prod_{i:g} p(y_i^g|\boldsymbol{\psi}^g, \boldsymbol{\theta}^g) = \prod_{i:g} \prod_{r=1}^{R} p(y_i^g|\boldsymbol{\psi}^g, \boldsymbol{\theta}^g) \tag{2.14}$$

The likelihood function can be written as:

$$\begin{aligned} \mathbf{L}_g(\boldsymbol{\psi}^g, \boldsymbol{\theta}^g) &= \prod_{i:g} f_g^{marg}\big(\mathbf{y}_{i,1}, \ldots, \mathbf{y}_{i,R};\, \boldsymbol{\psi}^g,\, \boldsymbol{\theta}^g,\, \mathcal{C}_i\big), \\ &= \prod_{i:g} \int f_g^{cond}\big(\mathbf{y}_{i,1}, \ldots, \mathbf{y}_{i,R} \,\big|\, \mathbf{b}_i;\, \boldsymbol{\psi}^g,\, \mathcal{C}_i\big)\, f_g^{ranef}\big(\mathbf{b}_i;\, \boldsymbol{\theta}^g\big)\, \mathrm{d}\mathbf{b}_i, \\ &= \prod_{i:g} \int \prod_{r=1}^{R} \prod_{j=1}^{n_{i,r}} p_r\big(y_{i,r,j} \,\big|\, \mathbf{b}_i;\, \boldsymbol{\psi}^g,\, \mathcal{C}_i\big) \left\{ \sum_{k=1}^{K^g} w_k^g\, \varphi(\mathbf{b};\, \boldsymbol{\mu}_k^g,\, \mathbb{D}_k^g) \right\} \mathrm{d}\mathbf{b}_i, \\ &= \prod_{i:g} \left\{ \sum_{k=1}^{K^g} w_k^g \int \prod_{r=1}^{R} \prod_{j=1}^{n_{i,r}} p_r\big(y_{i,r,j} \,\big|\, \mathbf{b}_i;\, \boldsymbol{\psi}^g,\, \mathcal{C}_i\big)\, \varphi(\mathbf{b};\, \boldsymbol{\mu}_k^g,\, \mathbb{D}_k^g)\, \mathrm{d}\mathbf{b}_i \right\} \end{aligned} \tag{2.15}$$

where $f_g^{marg}\big(\mathbf{y}_{i,1}, \ldots, \mathbf{y}_{i,R};\, \boldsymbol{\psi}^g,\, \boldsymbol{\theta}^g,\, \mathcal{C}_i\big)$ is the marginal density of the observed markers, where the term marginal indicates that the random effects are integrated out, $f_g^{cond}\big(\mathbf{y}_{i,1}, \ldots, \mathbf{y}_{i,R} \,\big|\, \mathbf{b}_i;\, \boldsymbol{\psi}^g,\, \mathcal{C}_i\big)$ is the conditional density of the observed markers given

the random effects vectors which presents as:

$$f_g^{cond}\big(\mathbf{y}_1, \ldots, \mathbf{y}_R \,\big|\, \mathbf{b};\, \boldsymbol{\psi}^g,\, \mathcal{C}\big) = \prod_{r=1}^{R} \prod_{j=1}^{n_r} p_r\big(y_{r,j} \,\big|\, \mathbf{b};\, \boldsymbol{\psi}^g,\, \mathcal{C}\big). \tag{2.16}$$

Here, $p_r\big(y_{r,j} \,\big|\, \mathbf{b};\, \boldsymbol{\psi}^g,\, \mathcal{C}\big)$ denotes a density function of the exponential family distribution of the random marker $\mathbf{Y}_{r,j}$ whose expectation relies on the random effects vector $\mathbf{b}$, the fixed effects $\boldsymbol{\alpha}_r^g$ and covariate information $\mathcal{C}$ by the GLMM model 2.12. The random effects density f_g^{ranef} in 2.15 in the prognostic group g is presented as:

$$f_g^{ranef}\big(\mathbf{b};\, \boldsymbol{\theta}^g\big) = \sum_{k=1}^{K^g} w_k^g\, \varphi(\mathbf{b};\, \boldsymbol{\mu}_k^g,\, \mathbb{D}_k^g) \tag{2.17}$$

where $\varphi(\mathbf{b};\, \boldsymbol{\mu}_k^g,\, \mathbb{D}_k^g)$ is a multivariate normal distribution density function.

Komárek et al. (2013) suggested using Markov Chain Monte Carlo (MCMC) methodology within a Bayesian framework for models with high dimension on random effects instead of using maximum-likelihood through the EM algorithm.

Basically, in a situation when the rth GLMM distribution is normal distribution (continuous markers) a block Gibbs algorithm is used to obtain a sample from the posterior distribution of model parameters. However, Komárek et al. (2013) suggested using the Metropolis–Hastings steps proposed by Gamerman (1997) in the case where there are different nature (i.e., continuous, discrete or dichotomous) of markers involved in the model.

A sample $S_M = \{(\boldsymbol{\psi}^{g,(m)}, \boldsymbol{\theta}^{g,(m)}) \,:\, m = 1 \ldots M\}$ is generated from the posterior distribution $p(\boldsymbol{\psi}^g, \boldsymbol{\theta}^g | y_g) \propto L_g(\boldsymbol{\psi}^g, \boldsymbol{\theta}^g) \times p(\boldsymbol{\psi}^g, \theta^g)$ (where $L_g(\boldsymbol{\psi}^g, \boldsymbol{\theta}^g)$ is the likelihood of the Bayesian model and $p(\boldsymbol{\psi}^g, \theta^g)$ is the prior distribution of the model parameters) by using the MCMC methods, where

$$\boldsymbol{\psi}^{g,(m)} = (\boldsymbol{\alpha}_1^{g,(m)}, \ldots \boldsymbol{\alpha}_R^{g,(m)}, \phi_1^{g,(m)}, \ldots, \phi_R^{g,(m)}),$$

and

$$\boldsymbol{\theta}^{g,(m)} = (w_1^{g,(m)}, \ldots, w_{K^g}^{g,(m)}, \boldsymbol{\mu}_1^{g,(m)}, \ldots, \boldsymbol{\mu}_{K^g}^{g,(m)}, \mathbb{D}_1^{g,(m)}, \ldots, \mathbb{D}_{K^g}^{g,(m)}).$$

The model parameters, $\boldsymbol{\psi}^g$ and $\boldsymbol{\theta}^g$ estimated from the multivariate mixture GLMM in each group are used in the discriminant analysis. In Equation 2.12, h_r^{-1} is a canonical like function which assumes that

$$\begin{aligned} p(y_{i,r}|\phi_r^g, \boldsymbol{\alpha}_r^g, \mathbf{b}_{i,r}) &= p(y_{i,r}|\phi_r^g, \eta_{i,r}^g) \\ &= exp\{\frac{y_{i,r}^g \eta_{i,r}^g - q_r(\eta_{i,r}^g)}{\phi_r^g} + k_r(y_{i,r}^g \phi_r^g)\}, \end{aligned}$$

where k_r and q_r are functions for appropriate distribution. The three common distributions are Gaussian with a linear mixed model for the rth marker where the dispersion parameter ϕ_r^g is the unknown residual variance, Bernoulli with the logit link where $\phi_r^g = 1$ and Poisson with the log link where $\phi_r^g = 1$ (see Komárek and Komárková (2013) for more detail).

There are many ways to evaluate convergence of the performed MCMC simulation. One possible way is that autocorrelations can be estimated in Markov chain of the model deviances $D(\psi, \theta) = -2log\{L(\psi, \theta)\}$ by using `autocorr` function from the `R` package (version 3.4.3) `coda` (Plummer et al. (2006)). Another way that can be followed to evaluate the performance of MCMC samples is to consider the trace plots of model parameters such as w, μ, $\mathbb{D}$ and model deviance $D(\theta)$. That can be accomplished by using `tracePlots` function from `R` package `mixAK`. See Figure 4.1 of Chapter 4, for an example of using trace plots to assess MCMC convergence.

To classify a new subject using their longitudinal history. In a Bayesian method, estimating the predictive density $\hat{f}_{g,new}$ is the average across all posterior samples as

$$\hat{f}_{g,new} = \frac{1}{M} \sum_{m=1}^{M} f(\mathbf{y}_{new};\, \boldsymbol{\psi}^{g,(m)}, \boldsymbol{\theta}^{g,(m)}). \tag{2.18}$$

2.3.5 Marginal approach

In the literature on LoDA, the marginal approach is the most commonly used approach (for example, Brant et al. (2003, 2005) and Morrell et al. (2005)).

The marginal approach aims to use the longitudinal profiles of a new patient to classify them by comparing them with the group-specific average profiles.

The marginal density $f^{marg}_{g,new}\big(.;\, \boldsymbol{\psi}^g,\, \boldsymbol{\theta}^g,\, \mathcal{C}\big)$ using the observed values $\mathbf{y}_{new,1} = \big(\mathbf{y}_{1,1},\, \ldots,\, \mathbf{y}_{1,n_1}\big),\, \ldots\, \mathbf{y}_{new,R} = \big(\mathbf{y}_{R,1},\, \ldots,\, \mathbf{y}_{R,n_R}\big)$ of longitudinal markers $\mathbb{Y}_{new} = \big(\mathbf{y}_{new,1},\, \ldots,\, \mathbf{y}_{new,R}\big)$ for a participant from group g is

$$f^{marg}_{g,new}\big(\mathbf{y}_1,\, \ldots,\, \mathbf{y}_R;\, \boldsymbol{\psi}^g,\, \boldsymbol{\theta}^g,\, \mathcal{C}\big) = \int f^{cond}_g\big(\mathbf{y}_1,\, \ldots,\, \mathbf{y}_R \,\big|\, \mathbf{b};\, \boldsymbol{\psi}^g,\, \mathcal{C}\big)\, f^{ranef}_g\big(\mathbf{b};\, \boldsymbol{\theta}^g\big)\, \mathrm{d}\mathbf{b}, \tag{2.19}$$

where $f^{ranef}_g\big(\mathbf{b};\, \boldsymbol{\theta}^g\big)$ denotes a density of the random effects and conditional density of the observed markers given the random effects vectors represents as $f^{cond}_g\big(.;\,|\,\mathbf{b};\, \boldsymbol{\psi}^g,\, \mathcal{C}\big)$. Equation 2.16 presents the density function of conditional of the observed markers and $f^{ranef}_g\big(\mathbf{b}_i;\, \boldsymbol{\theta}^g\big)$ denotes a density of the random effects (Equation 2.17 shows the random effects density function).

The predictive density $f_{g,new}$ of a new patient is equal to the marginal density of $\mathbb{Y}_{new} = \big(\mathbf{y}_{new,1},\, \ldots,\, \mathbf{y}_{new,R}\big)$.

$$\mathcal{P}^{marg}_{new,g}\big(\boldsymbol{\psi},\, \boldsymbol{\theta}\big) = \frac{\pi_g\, f^{marg}_{g,new}\big(\mathbf{y}_{new,1},\, \ldots,\, \mathbf{y}_{new,R};\, \boldsymbol{\psi}^g,\, \boldsymbol{\theta}^g,\, \mathcal{C}\big)}{\sum_{\widetilde{g}=1}^{G} \pi_{\widetilde{g}}\, f^{marg}_{\widetilde{g},new}\big(\mathbf{y}_{new,1},\, \ldots,\, \mathbf{y}_{new,R};\, \boldsymbol{\psi}^{\widetilde{g}},\, \boldsymbol{\theta}^{\widetilde{g}},\, \mathcal{C}\big)} \qquad g = 1,\, \ldots,\, G, \tag{2.20}$$

where $f^{marg}_{g,new}$ is the marginal density 2.19. Marginal group probabilities are measured as the average across all M samples in MCMC procedure.

$$\widehat{\mathcal{P}}^{marg}_{new,g} \;=\; \frac{1}{M} \sum_{m=1}^{M} \mathcal{P}^{marg}_{new,g}\big(\boldsymbol{\psi}^{(m)},\, \boldsymbol{\theta}^{(m)}\big), \qquad g = 1,\, \ldots,\, G,$$

In the case of continuous markers and a single normal distribution of random effects (K = 1), the marginal prediction is equivalent to the modified discriminant analysis except that different approaches are used to estimate the parameters (with a maximum likelihood (ML) approach used for the modified discriminant analysis and a Bayesian MCMC approach for the marginal approach). However, this equivalence does not hold in the case of violations of the normality assumption (such as binary markers).

2.3.6 Conditional approach

In the conditional prediction approach, the group membership probabilities are calculated from the average of M samples from the MCMC routine as follows:

$$\widehat{\mathcal{P}}^{cond}_{new,g} = \frac{1}{M} \sum_{m=1}^{M} \mathcal{P}^{cond}_{new,g}\big(\mathbf{b}^{1,(m)}_{new}, \ldots, \mathbf{b}^{G,(m)}_{new}, \boldsymbol{\psi}^{(m)}, \boldsymbol{\theta}^{(m)}\big), \qquad g = 1, \ldots, G.$$

where

$$\mathcal{P}^{cond}_{new,g}(\mathbf{b}^{1}_{new}, \ldots, \mathbf{b}^{G}_{new}, \boldsymbol{\psi}, \boldsymbol{\theta}) \;:=\; \frac{\pi_g \, f^{cond}_{g}\big(\mathbf{y}_{new,1}, \ldots, \mathbf{y}_{new,R} \,\big|\, \mathbf{b}^{g}_{new}; \boldsymbol{\psi}^{g}\big)}{\sum_{\widetilde{g}=1}^{G} \pi_{\widetilde{g}} \, f^{cond}_{\widetilde{g}}\big(\mathbf{y}_{new,1}, \ldots, \mathbf{y}_{new,R} \,\big|\, \mathbf{b}^{\widetilde{g}}_{new}; \boldsymbol{\psi}^{\widetilde{g}}\big)}$$

where $\mathbf{b}_{new}$ is the values of the unobserved random effects vector for the new individual.

The advantage of conditional prediction is that it is based on the patient-specific development of the markers over time. However, the error variance in the estimation of individual random effects is not taken into account (Komárek et al. (2010)).

2.3.7 Random-effects approach

For the random effects approach, prediction is based on the patient-specific growth of the longitudinal markers. The posterior probability is computed by using the distribution of the individual random effects. The group membership probabilities for the random effects prediction are calculated by averaging over the MCMC samples.

$$\widehat{\mathcal{P}}^{ranef}_{new,g} = \frac{1}{M} \sum_{m=1}^{M} \mathcal{P}^{ranef}_{new,g}\big(\mathbf{b}^{1,(m)}_{new}, \ldots, \mathbf{b}^{G,(m)}_{new}, \boldsymbol{\psi}^{(m)}, \boldsymbol{\theta}^{(m)}\big), \qquad g = 1, \ldots, G.$$

where

$$\mathcal{P}^{ranef}_{new,g}(\mathbf{b}^{1}_{new}, \ldots, \mathbf{b}^{G}_{new}, \boldsymbol{\psi}, \boldsymbol{\theta}) \;:=\; \frac{\pi_g \, f^{ranef}_{g}\big(\mathbf{b}^{g}_{new}; \boldsymbol{\theta}^{g}\big)}{\sum_{\widetilde{g}=1}^{G} \pi_{\widetilde{g}} \, f^{ranef}_{\widetilde{g}}\big(\mathbf{b}^{\widetilde{g}}_{new}; \boldsymbol{\theta}^{\widetilde{g}}\big)}.$$

The random effects approach uses the parameters of the random effects distribution.

2.3.8 Classification rules

As described earlier in the modified discriminant analysis, linear mixed models are used to estimate the parameters (means and covariance matrices), and these parameter estimates are then used in LDA and QDA to classify a patient.

For classification, the group membership probabilities of the modified discriminant analysis are calculated (see Section 2.2.4 for detail). Then, by using an optimal cutoff, the patient is allocated to the group if the probability of belonging to that group is larger than the optimal cutoff c.

Similarly, the group membership probabilities from the marginal, conditional and random-effects approaches are estimated first. The classification of a new subject to group g is based on a combination of the prior probabilities $\pi_1, \ldots, \pi_G$ ($0 < \pi_g < 1, \sum_{g=1}^{G} \pi_g = 1$), and then the use of Bayes' rule to estimate the group membership probabilities by:

$$\mathcal{P}_{g,new} = \frac{\pi_g \hat{f}_{g,new}}{\sum_{\tilde{g}=1}^{G} \pi_{\tilde{g}} \hat{f}_{\tilde{g},new}} \qquad g = 1, \ldots, G. \tag{2.21}$$

Where $\hat{f}_{g,new}$ is defined in Equations 2.15 and 2.18. Then, the patient is allocated to the group if the probability of belonging to that group is larger than a determined cutoff c (optimal cutoff). Typically, this optimal cutoff is selected by analysis of a receiver operating characteristic (ROC) curve.

2.4 Assessment of classification accuracy

2.4.1 Validation Method

In order to assess the performance of a predictive model, an investigation of the model accuracy is needed. There are two approaches used in this thesis: splitting the sample and leave-one-out cross-validation.

One option is that the sample is divided into two sample sets (e.g., 70:30 or 90:10). One sample is used to develop the discriminant model (known as training sample) and the second sample is used for assessing the accuracy of the discriminant model (known as validation sample). Generally, the two datasets are generated by randomly splitting the original data. In the case of discriminant analysis, I split the sample dataset into the training set (and also in the validation set) based on the group prevalence. This means that I randomly split each group into a training set and testing set, then I combine the training sets from each group together to have one training set and the same procedure is followed for the testing sets.

The idea of using a testing sample for validation is to examine how well the discriminant model works on the sample of observations not used to generate the discriminant model which helps to reduce potential over-fitting. Splitting the sample is repeated a number of times (e.g., 100) to produce sets of training and validation samples. The several accuracy measures are then averaged to obtain a representative accuracy measure, and ensure that accuracy measures are not overly influenced by the selection of training/test sets. Hair et al. (1994) pointed out that the motivation for dividing the total sample into two sets is to avoid an upward bias in the prediction accuracy of the discriminant model. This upward bias occurs when the subjects used in building the classification model are the same as those used in assessing is accuracy.

A second internal validation approach used to evaluate predictions methods internally is the leave-one-out cross-validation (LOOCV) method. In LOOCV, each observation is omitted as a testing sample, and the remaining $n - 1$ observations are a training sample. In other words, $n-1$ observations are used to develop the discriminant model, and the remaining observation is used to assess the model.

Sensitivity and specificity

Sensitivity and specificity are common indicators of the diagnostic ability of the test which was introduced early by Yerushalmy (1947). Table 2.1 represents a 2×2 contin-

gency table for a dichotomous outcome that summarised the results in terms of true and false positives and true and false negatives. There are four possible results. If the true status is positive and a subject is classified as positive, this is called a true positive (TP), whilst if they are classified as a negative, they are referred to as a false negative (FN). If the true status is negative and a patient is classified as negative, they are called a true negative (TN), while if they are classified as positive they are referred to as false positives (FP) (Fawcett (2006)).

Table 2.1: Classification table: Predicted versus true outcome.

		Predicted disease status		
		Positive (1)	Negative (0)	Total
True disease status	Positive (1)	True positive (TP)	False negative (FN)	all true positive
	Negative (0)	False positive (FP)	True Negative (TN)	all true negative
Total		all positive predicted	all negative predicted	Total sample size

Source: Kumar and Indrayan, Receiver operating characteristic (ROC) curve for medical researchers, 2011; Fawcett, An introduction to ROC analysis, 2006.

In clinical applications where the aim is to predict disease, the true positive rate (TPR) or sensitivity is the probability that someone who has a disease is classified as having the disease, and the true negative rate (TNR) or specificity is the probability that someone who does not have have the disease is classified as not having the disease. Therefore, sensitivity can be calculated as $\frac{TP}{TP+FN}$ and specificity as $\frac{TN}{TN+FP}$. Furthermore, positive predictive value (PPV) and negative predictive value (NPV) are two additional measurements used to evaluate the discriminant ability of the model. PPV is defined as the probability that the disease is present when classified as having the disease, and NPV is the probability that the disease is not present when classified as not having the disease. The estimation of these probabilities can be presented as $PPV = \frac{TP}{TP+FP}$, and $NPV = \frac{TN}{FN+TN}$ (Kumar and Indrayan (2011), Fawcett (2006)).

2.4.2 Receiver operating characteristic curve (ROC)

There is a large volume of published studies describing the role of the ROC curve and AUC in medical decision making (Goodenough et al. (1974), Metz (1978); Zou (2002),

Hajian-Tilaki (2013)). The receiver operating characteristic (ROC) curve (Lusted (1971)), is a two-dimensional plot that presents the full picture of all possible points of sensitivity (true positive rate) on the y-axis and (1 - specificity; false positive rate) on the x-axis across a series of cut-off points. Sensitivity has an inverse relationship with specificity, which means that sensitivity increases as specificity decreases across different thresholds.

The ROC curve can be valuable in three instances: (i) obtaining the optimal cut-off point for minimally misclassifying either diseased or healthy while also permitting all possible cut-off points to be presented on the ROC subjects, (ii) evaluating the discriminatory capability of a test to separate diseased from healthy subjects, and (iii) comparing the ability of two or more tests in evaluating the same disease (Kumar and Indrayan (2011)).

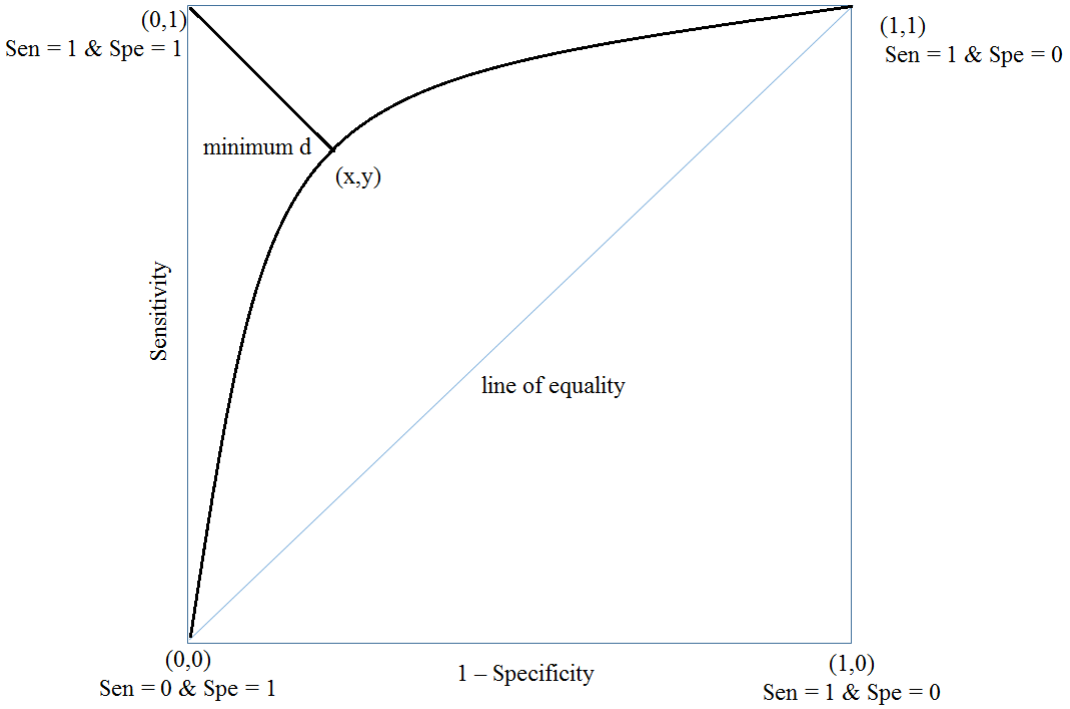

Figure 2.1: Finding best cut-off from the ROC curve

Source: Kumar and Indrayan, Receiver operating characteristic (ROC) curve for medical researchers, 2011.

Figure 2.1 below depicts a ROC curve and demonstrates how to choose an optimal cut-off point. The area under the ROC curve (AUC) is an effective measure of perfor-

mance for classification and diagnostic rules. The classification is excellent when the AUC is near to 1. In Figure 2.1, the line between points (0,0) and (1,1) divides the square into two equal areas, and each one is equal to 0.5. ROC closer to the left-hand side of the graph (i.e., close to the point (0,1)) indicates the excellent performance of the model.

Optimal threshold point

The optimal threshold is the point that gives the best balance between sensitivity and specificity values. I followed a criterion that uses the distance between the points (0,1) and any other point on the ROC curve to select the optimal cut-off point (Kumar and Indrayan (2011)). The distance can be calculated as:

$$d^2 = [(1 - Sen)^2 + (1 - Spe)^2].$$

This distance can be calculated for each observed cut-off point, and the point that provides minimum distance can be determined. This is shown in Figure 2.1.

There are other procedures available for choosing a threshold, for example, Freeman and Moisen (2008) and Hughes et al. (2017) presented comparisons of different choices for determining a threshold.

2.4.3 Area under curve AUC

The area under the receiver operating characteristic curve (AUC) is one of the common quantities used to evaluate the performance of classification rules (Hanley and McNeil (1982)). The AUC uses a range of thresholds points to summarise the ROC into a single value. Hanley and McNeil (1982) pointed out that Wilcoxon test of the ranks is equivalent to AUC. Also, the Gini coefficient is related to the AUC which is twice the area between the ROC curve and the diagonal line, and it is therefore determined as $Gini = 2.AUC - 1$ (Hand (2012), Hand and Till (2001)). Suppose that F_0 and F_1

are the cumulative score distributions of the two classes, respectively. The AUC can be measured by:

$$AUC = \int F_0(t)dF_1(t) = \int F_1(t)dF_0(t)$$

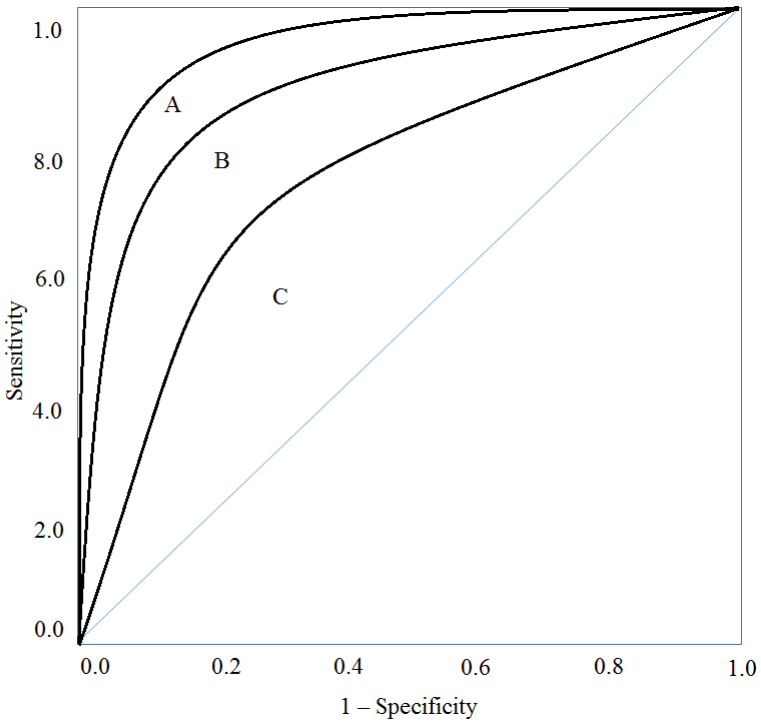

Figure 2.2: Comparison of three ROC curves with different areas.

Source: Kumar and Indrayan, Receiver operating characteristic (ROC) curve for medical researchers, 2011.

The AUC value ranges from 0 to 1.0. A perfect classification is obtained when the value of the AUC is equal to 1.0. Figure 2.2 shows three ROC curves and their different AUC. The AUC of ROC curve A shows a better classification accuracy when compared to ROC curves B and C. In general, the discrimination of a test is better when the ROC curve is closer to the left-hand side (Kumar and Indrayan (2011)). Hand (2009) pointed out that the AUC has a well-known weakness. When the ROC curves cross, the AUC might provide misleading results.

2.5 Summary

In this section, I have presented the approaches that are used in this thesis. Two well-known approaches of discriminant analysis are linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). LDA works well under the assumption of multivariate normality of the explanatory variables with equal covariance matrices across groups. QDA is recommended when the assumption of equal covariance is not met and it tends to require larger sample sizes to cope with the larger number of parameters that need to be estimates compared to LDA. Box's M, Bartlett's, Brown and Forsythe's tests are typically used in a discriminant analysis to test of equality of variance-covariance matrices.

A limitation of classical discriminant analysis when longitudinal data is involved is that they cannot handle an unbalanced design (i.e., where the longitudinal marker is not measured from all patients at the same time points). In other words, applying classical LDA or QDA to longitudinal data usually leads to the exclusion of patients. Further, classical discriminant analysis is ignoring the correlation between repeated measurements on the same patient.

To overcome the problems associated with the missing values, a modified discriminant analysis based on the mixed model is proposed by Tomasko et al. (1999). Additionally, in the modified discriminant analysis the correlation between repeated measurements on the same patient is taken into account. The maximum likelihood approach (ML) is used to estimate the parameters of the mixed model.

A more flexible approach for longitudinal discriminant analysis (LoDA) has been recently developed by Hughes et al. (2018b). This approach uses a multivariate generalised linear mixed model with a normal mixture for the random-effects distribution, and a Bayesian approach is used to estimate the model parameters.

In this chapter, two internal validation approaches: splitting the sample and leave-one-out cross-validation to assess the model accuracy are described. Some statistical

measurements, such as sensitivity, specificity, a probability of correct classification, AUC, positive predictive value and negative predictive value are explained.

In the next chapter, the classical and modified discriminant analysis methodologies will be applied to a clinical dataset. LoDA approaches will be applied to the clinical dataset (PBC dataset) and simulation study in Chapters 4 and 5.

Chapter 3

Analysis of clinical data: Ophthalmic application

3.1 Introduction

In this chapter, I apply a range of discriminant analysis methods to the longitudinal ophthalmology data described in Chapter 1 in order to predict treatment success or treatment failure in patients with age-related macular degeneration (AMD). The chapter starts with a description of the structure of the data that was used. This is followed by an investigation of two groups of patients, those whose vision improved and those whose vision did not improve. I will study how these groups can be identified early by looking at their values of visual acuity (VA) and contrast sensitivity (CS), individually or together by using the univariate and multivariate classical linear and quadratic discriminant analysis, respectively. A comparison of two different approaches to discriminant analysis will be considered. The first, which I call a classical discriminant analysis assumes a complete and balanced dataset and each visit is treated as a separate variable. The second approach, which is called a modified discriminant analysis in this thesis is based on the linear mixed model (Tomasko et al. (1999)).

The structure of this chapter is as follows. Section 3.2 describes the data used in

this chapter. The classical discriminant analysis approach is addressed in Section 3.3, whilst the modified approach is shown in Section 3.4. Finally, the chapter concludes with a discussion of discriminant analysis approaches.

3.2 Description of data types

The Ophthalmic data is briefly described in Chapter 1. In this chapter, two discriminant analysis approaches were applied to predict treatment success or treatment failure in patients who had neovascular age-related macular degeneration (nAMD) at Paul's Eye Unit, Royal Liverpool University Hospital. Patients who had nAMD were treated with verteporfin photodynamic therapy at baseline and were then followed for a year. The clinical protocol was to examine the patient every three months, specifically at 3, 6, 9 and 12 months, and measurements were taken of CS and VA. These markers have been described in Section 1.3.1 and are important to assess vision function. The data consists of 1008 patients (with an average of 5.77 visits per patient) with a large number of clinical measurements and a median age of 78 years at baseline.

The purpose of this chapter is to use only data gathered before 12 months to identify patients whose vision will improve and those whose vision will not improve. For this reason, only patients who were followed for at least 12 months were included in this analysis. All patients must have had a baseline visit and had at least one follow-up visit to be included in this analysis. Although patients were expected to be examined every three months, roughly at 3, 6, 9 and 12 months, this routine was hard to achieve in practice. For example, some patients had at least one appointment missing. These are patients who missed some of their appointments, and they may be seen only twice or three times.

Therefore, the ophthalmic sample consisted of 447 patients (complete cases, i.e., all patients had complete markers profiles) who were followed up to a year. Before the analysis was achieved, several issues had to be considered: (i) the effect of approximating the actual time of the patient's visit to the clinic, to the nearest scheduled

visit, (ii) the benefit of using a modified discriminant method that uses mixed models and the exact visit times, (iii) the effects of imputing missing data. To achieve this, I considered four datasets for analysis.

- In the first dataset, an allocation process was followed so only the patients who were seen at least five times were used, and I used the actual visiting time. This dataset included 176 patients who had at least five visits during a year (all patients must have had a baseline visit and had at least four additional visits which may or may not be at these indented time points at 3, 6, 9 and 12 months). That means that the data is not balanced as each patient has a different number of visits (at least five observations), hence it is described as an unbalanced dataset (referred to as D1 in Figure 3.1). There were 176 such patients, and among them, there were a total of 72 patients whose treatment was classified as a failure at 12 months and 104 patients whose treatment was classified as success at 12 months. This unbalanced dataset (D1) is appropriate for the modified discriminant analysis based on the mixed model.

- As the time points of the visits of the 176 patients (D1) differ from each other, the data cannot be analysed with the classical discriminant analysis, which requires a balanced design (all patients must have five observations and identical visit times). Therefore, in the second dataset, the time point was approximated, all patients must have had a baseline visit and had four observations at a predefined time point or near to 3, 6, 9 and 12 months (referred to as D2 in the Figure 3.1). For example, if a patient was late to his/her second visit and had it 4-months instead of 3-months after baseline, his/her 4-months visit was treated as a 3-months visit in this dataset. It is important to note that, the unbalanced subset data (D1) and balanced subset data (D2) involved the same patients. The classical discriminant analysis can be applied to this balanced dataset (D2).

- A total of 447 patients (who may have missed one or more of their appointment)

was considered as the third dataset. This data might include a patient who visited once over a year or who attended the clinic more than five times. The times at which the visits occurred vary across patients. This dataset is referred to as D3 in Figure 3.2 which shows that D3 includes all patients from D1 and also those who missed one or more visits. There were a total of 264 patients whose treatment was classified as success at 12 months and 183 patients whose treatment was classified as the failure at 12 months. The modified discriminant analysis can be used for the full dataset (D3).

- In the last dataset, the missing values in D3 were imputed using the last observation carried forward. For example, assume a patient who received the treatment at month 0 and then came to his visits at time points 2, 5 and 12 months when his CS and VA were measured and recorded. He should have had his visits at 3, 6, 9 and 12 months, so he came to his 3-month and 6-month visits too early, at 2-month, and 5-month, and he missed his 9-month visit completely. To deal with such a situation, I used the values at 2 and 5 months to populate the values at 3 and 6, and I imputed the 9-months visit by carrying the last observation from the last time point from month 5. This dataset was created in order to facilitate the use of classical discriminant analysis and hence to see the effect of time approximation and data imputation. The Figure 3.2 has referred to this dataset as D4 dataset. The aim of applying this imputation method is that the classical method will be able to predict a patient who was not seen as required by the doctors (at 0, 3, 6, 9, 12 months). This dataset includes the same patients as in the unbalanced dataset (D3) (i.e., 447 patients). This imputed dataset will be analysed using the classical discriminant analysis approach.

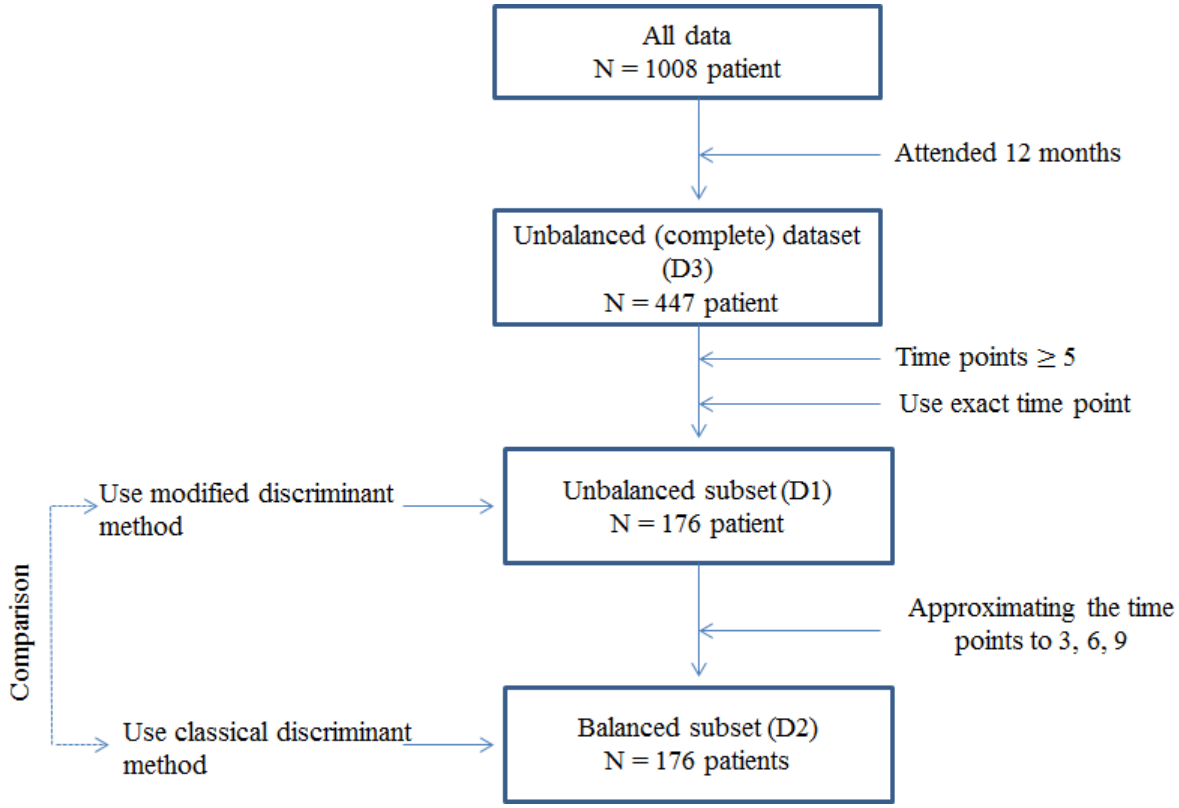

Figure 3.1: Flow chart illustrating the comparisons between an unbalanced dataset (D1) and a balanced dataset (D2) and the time approximation.

3.3 Classical linear and quadratic discriminant analysis

The methodology of the classical linear and quadratic discriminant analysis is described in Chapter 2, Section 2.2. Linear discriminant analysis assumes the equality of the variance-covariance matrices between the two groups, while quadratic discriminant analysis assumes that the variance-covariance matrices are unequal. These traditional discriminant methods can be used for longitudinal data only if the data is complete and balanced. In other words, they cannot handle missing values, and they cannot handle if patients measurements were arranged at different time points. Each visit date is considered as an independent variable and the correlation between repeated measurements on the same patient is not taken into account in the traditional discriminant methods. Tomasko et al. (1999) mentioned that linear/quadratic discriminant analysis can be called classical linear/quadratic discriminant analysis when there is not a tool to join the random subject effect into a covariance matrix model.

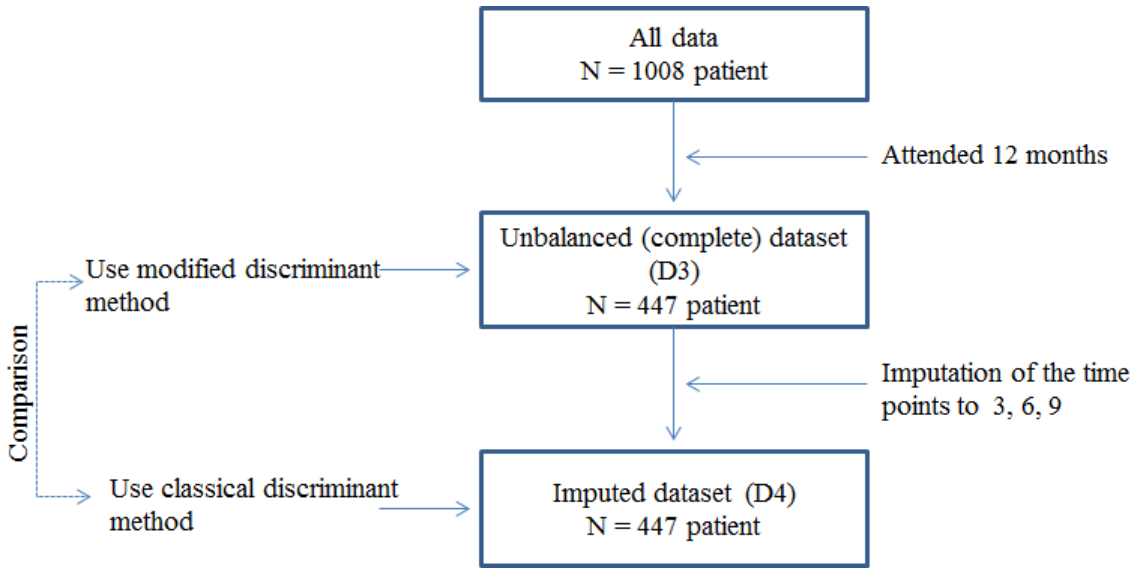

Figure 3.2: Flow chart illustrating an unbalanced dataset (D3) and a balanced and imputed dataset (D4).

This section will show a comparison of of classical LDA and classical QDA using the Ophthalmology data. In particular, I explore if QDA brings a benefit in terms of the area under the receiver operating characteristic (ROC) curve (AUC).

For this study, a balanced dataset on all markers of interest at specific visits was used. This dataset (D2) is described above (Figure 3.1). There were a total of 72 patients whose treatment was classified as a failure at 12 months and 104 patients whose treatment was classified as treatments success at 12 months. The two markers chosen for this study were: CS and VA, and four time points (0, 3, 6 and 9 months) for each patient were considered. These four time points for each of the two markers were entered as variables in LDA/QDA. The last visit (at 12 months) was used to determine who succeeds or fails the treatment (group membership). .

The aim of this application was to investigate the effect of CS and VA individually and together on treatment success by using the univariate and multivariate discriminant methods, respectively.

The longitudinal data were used in the linear and quadratic discriminant analysis by treating each visit as a separate variable. In other words, LDA/QDA was performed on each of the datasets where the visits to the particular time points were aligned.

The procedures that were applied for the univariate and multivariate longitudinal data are as follows:

(i) For each of the univariate discriminant models, there were four models involving CS. The first discriminant model included CS at baseline and age. Then, the second model used CS at baseline, CS at 3 months and age at baseline. The third model used CS measurements at baseline, 3 and 6 months and age. In the final model, CS at 9 months was added to the previous model (i.e., the final model contained age at baseline, CS at 0, 3, 6 and 9 months). The univariate discriminant equations are described in Chapter 2, Equation 2.2.

(ii) For each of the univariate discriminant models, there were four models involving VA. The VA models were built similarly as the models involving CS, above. The first model used the information at baseline of VA and age. Information of VA at 3 months was added to the first model to build the second VA model. Next, VA at baseline, 3 and 6 months and age was used to build the third model. VA at 9 months was added to the previous model to build the final VA model (i.e., this model included 5 variables, VA at baseline, 3, 6 and 9 months and age at baseline).

(iii) For the multivariate discriminant model involving both VA and CS, I followed the same process as in the univariate approach above. Four multivariate discriminant models were applied to the balanced dataset (D2). The first model included CS, VA and age at baseline. Then the second model was developed using the information at baseline and 3 months of the two markers (CS and VA). The third model involved information on CS and VA at baseline, 3 and 6 months. The final model contained the data on CS and VA for up to 9 months. Equation 3.1 illustrates all the models for the multivariate discriminant analysis.

$$
\begin{aligned}
\text{Model 1} &= a_1 CS0_i + a_2 VA0_i + a_3 Age_i \\
\text{Model 2} &= a_1 CS0_i + a_2 VA0_i + a_3 CS3_i + a_4 VA3_i + a_5 Age_i \\
\text{Model 3} &= a_1 CS0_i + a_2 VA0_i + a_3 CS3_i + a_4 VA3_i + a_5 CS6_i + a_6 VA6_i + a_7 Age_i \\
\text{Model 4} &= a_1 CS0_i + a_2 VA0_i + a_3 CS3_i + a_4 VA3_i + a_5 CS6_i + a_6 VA6_i + a_7 CS9_i + \\
&\quad a_8 VA9_i + a_9 Age_i
\end{aligned}
\tag{3.1}
$$

Other covariates such as lesion type and gender were also included in the model. However, including these covariates did not show improvement in classification accuracy of the model (i.e., the overall levels of PCC, sensitivity, specificity and AUC do not improve when these covariates are incorporated). Linear and quadratic discriminant analyses are common discriminant methods which have been used to classify patients into two or more groups. To decide whether the linear (LDA) or quadratic discriminant analysis (QDA) should be applied in the ophthalmic dataset, a preliminary test called Box's M test is often used in the discriminant analysis to test for equality of covariance matrices.

The standard test statistics of the M test is based on the assumption that the within-group covariance matrices for two groups (denoted by Σ_0 and Σ_1) are equal, where the null hypothesis is presented as:

$$H_0 : \Sigma_0 = \Sigma_1 \text{ versus } H_1 : \Sigma_0 \neq \Sigma_1.$$

The Box's M tests were performed to each model using `boxM` function from the `R` package `heplots` (Fox et al. (2018)). For each model, I calculated the proportion of p-values of chi-square test and compared with a predefined significance level $\alpha = 0.05$. If the p-value is less than α, then the null hypothesis is rejected. When the p-value of Box's M test is less than 0.05 the two covariance matrices are considered to be significantly different at the 0.05 level.

Table 3.1 provides information of χ^2 approximation, degrees of freedom and p-values.

Table 3.1: Statistics results of the Box's test that uses the univariate markers separately (CS, VA) and the multivariate markers (CS and VA together). Each model included age at baseline.

Marker	Time	χ^2 value	d.f.	p.value
CS	0	10.56	3.00	0.01
	0, 3	19.74	6.00	< 0.01
	0,3,6	54.07	10.00	< 0.01
	0,3,6,9	90.43	15.00	< 0.01
VA	0	8.21	3.00	0.04
	0,3	13.32	6.00	0.04
	0,3,6	19.91	10.00	0.03
	0,3,6,9	23.70	15.00	0.07
CS + VA	0	13.69	6.00	0.03
	0,3	28.46	15.00	0.02
	0,3,6	70.28	28.00	< 0.01
	0,3,6,9	136.78	45.00	< 0.01

It can be seen for the CS models, the p-values obtained by the Box's M test give strong evidence to reject the null hypothesis at level 0.05. This means that the covariance matrix of the patients who failed treatment are not be equal to the covariance matrix of patients who succeed the treatment. A similar conclusion can be reached when testing the equality of covariance matrices for the multivariate models the p-values give evidence the covariance matrices are unequal. For the VA univariate model that included four time points (at 0, 3, 6, 9 months on the Table 3.1), the p-value (0.07) is not small enough to give evidence that the covariances are significantly different. Similarly, the p-values of remaining models (p-values = 0.04, 0.04, 0.03) are close to the significant level, although they are statistically significant.

As explained in Mardia et al. (1979), Box's M test seems to work less well if the number of variables is more than 5 (i.e., the Box's M test is not sufficiently robust to imbalances in group sizes, or departure from multivariate normality.) There are other tests that can be used to check the homogeneity of variance including: Bartlett's Test and Levene's Test (Brown and Forsythe (1974)). These tests have been described in Chapter 2. The results obtained from the Levene's tests (based on mean and median

Table 3.2: Results of the Levene's test that uses the univariate markers separately (CS, VA) and the multivariate markers (CS and VA together). Each model included age at baseline.

Marker	Time	d.f. (v_1, v_2)	Levene's test (mean) F value	p.value	Levene's test (median) F value	p.value
CS	0	(1, 174)	6.34	0.01	6.17	0.01
	0, 3	(1, 174)	1.07	0.30	1.05	0.31
	0,3,6	(1, 174)	0.87	0.35	0.61	0.43
	0,3,6,9	(1, 174)	0.001	0.99	0.001	0.99
VA	0	(1, 174)	0.03	0.95	0.001	0.97
	0,3	(1, 174)	0.02	0.87	0.02	0.89
	0,3,6	(1, 174)	1.49	0.22	1.47	0.23
	0,3,6,9	(1, 174)	5.45	0.02	5.59	0.02
CS + VA	0	(1, 174)	0.30	0.58	0.31	0.58
	0,3	(1, 174)	0.05	0.82	0.20	0.89
	0,3,6	(1, 174)	1.74	0.20	1.69	0.19
	0,3,6,9	(1, 174)	1.73	0.19	1.73	0.19

locations) can be seen in Table 3.2. The p-value of CS at baseline is 0.01, and of VA at 0,3,6,9 months is 0.02 in both tests. Based on these results, there is a significant difference between the two variance matrices at a significance level of 0.05. The remaining results, as shown in Table 3.2, indicate that there is no evidence that the two variance matrices differ between the groups.

In contrast to the results of Levene's test, the results of Bartlett's test (as summarised in Table 3.3) show that there is evidence that the two variances matrices differ. The Bartlett's test assumes that the data comes from a normal distribution, while Levene's test allows for non-normal distributions. The Bartlett's test assess the variances only. If the variances are unequal while correlations are the same, this would be a scenario where applying QDA may be more beneficial than using LDA.

According to these results, it is possible that applying QDA will give better results than LDA, although there is some evidence for QDA in models with lower number of discriminatory variables. A considerable amount of literature has been published for deciding whether LDA will outperform QDA (see e.g., Huberty and Curry (1978), McLachlan (2004), Meshbane and Morris (1995)). They have reported that if the

Table 3.3: The test of equality of variances results of the Bartlett's test that uses the univariate markers separately (CS, VA) and the multivariate markers (CS and VA together). Each model included the age at baseline.

Marker	Time	K-squared	d.f.	p.value
CS	0	10.44	1.00	<0.01
	0,3	10.46	2.00	<0.01
	0,3,6	10.97	3.00	<0.01
	0,3,6,9	11.57	4.00	<0.01
VA	0	26.34	1.00	<0.01
	0,3	72.71	2.00	<0.01
	0,3,6	105.36	3.00	<0.01
	0,3,6,9	132.22	4.00	<0.01
CS + VA	0	73.38	2.00	<0.01
	0,3	208.19	4.00	<0.01
	0,3,6	346.55	6.00	<0.01
	0,3,6,9	493.92	8.00	<0.01

sample sizes of groups n_g is smaller than p (number of predictors), then LDA is preferred with ignoring heterogeneity of covariance matrices. While if the sample sizes of groups are large compared to p and the covariance matrices are heterogeneous, then QDA is suggested. There is very little guidance on how large sample sizes need to be (Huberty and Olejnik (2006)).

Therefore, in this thesis both the LDA and QDA were used to build the univariate and multivariate models. Each model (as described in Table 3.1) was analysed using LDA and QDA to predict the patient's status at 12 months, i.e., success of failure of the treatment. In total, there are 12 models to predict patient's status (i.e., whether they benefit of using the treatment and improve their vision or do not improve their vision). The classical discriminant function is basically determined based on the mean and covariance estimates obtained from each dataset. The classical LDA and QDA approaches were fitted using the `lda` and `qda` functions from the `R` package (version 3.4.3) `MASS`, and the prediction into the two treatment response groups was performed using `predict` function from the `R` package (version 3.4.3) `MASS` (Venables and Ripley (2002)).

Table 3.4: Results of the linear and quadratic discriminant analyses that uses a univariate marker (CS and VA, separately) and multivariate markers CS and VA together.

		LDA							QDA						
Marker	Prediction Time	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
CS	0	0.40	0.63	0.59	0.61	0.58	0.53	0.70	0.42	0.60	0.65	0.63	0.61	0.56	0.69
	0, 3	0.39	0.69	0.64	0.66	0.62	0.58	0.75	0.38	0.69	0.67	0.68	0.66	0.61	0.76
	0, 3, 6	0.42	0.65	0.77	0.72	0.69	0.68	0.75	0.34	0.76	0.74	0.75	0.76	0.68	0.82
	0, 3 ,6, 9	0.39	0.69	0.74	0.72	0.73	0.66	0.77	0.32	0.75	0.76	0.76	0.77	0.70	0.82
VA	0	0.42	0.60	0.67	0.64	0.62	0.57	0.71	0.39	0.70	0.61	0.64	0.63	0.57	0.74
	0, 3	0.40	0.67	0.69	0.68	0.68	0.62	0.75	0.38	0.68	0.66	0.67	0.69	0.60	0.75
	0, 3, 6	0.42	0.78	0.80	0.79	0.83	0.74	0.84	0.39	0.78	0.78	0.78	0.82	0.72	0.84
	0, 3, 6, 9	0.35	0.87	0.89	0.88	0.93	0.85	0.91	0.38	0.85	0.86	0.86	0.91	0.82	0.89
CS + VA	0	0.39	0.70	0.62	0.66	0.66	0.58	0.75	0.39	0.71	0.66	0.68	0.68	0.61	0.77
	0, 3	0.37	0.74	0.67	0.70	0.71	0.63	0.79	0.33	0.74	0.68	0.71	0.71	0.63	0.79
	0, 3, 6	0.42	0.77	0.82	0.79	0.82	0.75	0.83	0.29	0.83	0.75	0.78	0.82	0.71	0.86
	0, 3, 6, 9	0.34	0.89	0.88	0.88	0.93	0.85	0.92	0.33	0.86	0.85	0.85	0.89	0.81	0.89

The classification accuracy of the univariate and multivariate discriminant analysis applied to the balanced dataset (D2) is provided in Table 3.4. The results were evaluated by splitting the data into 90% for training (i.e., used to derive the discriminant function) and 10% for testing the model. Then this was repeated 100 times to get stable results. The column labelled Model indicates the model type that has been used, and the column called Prediction Time shows the prediction times used (response to the treatment is known at time 12 months). Table 3.4 provides the following performance metrics of the prediction: the sensitivity, specificity, probability of correct classification (PCC), AUC, positive predictive value (PPV) and negative predictive value (NPV) of the linear and quadratic approaches applied to the balanced data. The optimal cut-off point, defined as the point that gives the best balance between sensitivity and specificity values, is used to get the best predictions. These performance metrics are the averages of the 100 repeats.

Table 3.4 shows that using more longitudinal information (time prediction 0, 3, 6, 9 months) in the 12 models (the univariate models and multivariate model) gives the best values of PCC and AUC. Secondly, the univariate models that involved VA give better discriminant results compare with the univariate models that included CS. This result is not surprising, according to the definition of the patient's outcome (failure or success the treatment) which is based on the values of visual acuity (VA) at 12 months, which makes VA more informative at time points 6 and 9 months, since they are closer to 12 months.

Moreover, using the multivariate discriminant analysis does not show improvement in terms of the PCC and AUC after prediction time 6 months, at which point, the results are similar to the univariate model that used VA. However, the multivariate model improves the prediction at the early time points. In other words, the benefit of using the multivariate discriminant analysis, in this application, is in helping to identify early (using data from baseline and three months) the patients who will improve their eye vision (i.e., those who will have improved vision at 12 months). Figure 3.5 shows the ROC curves for the multivariate models using LDA and QDA and shows the improved

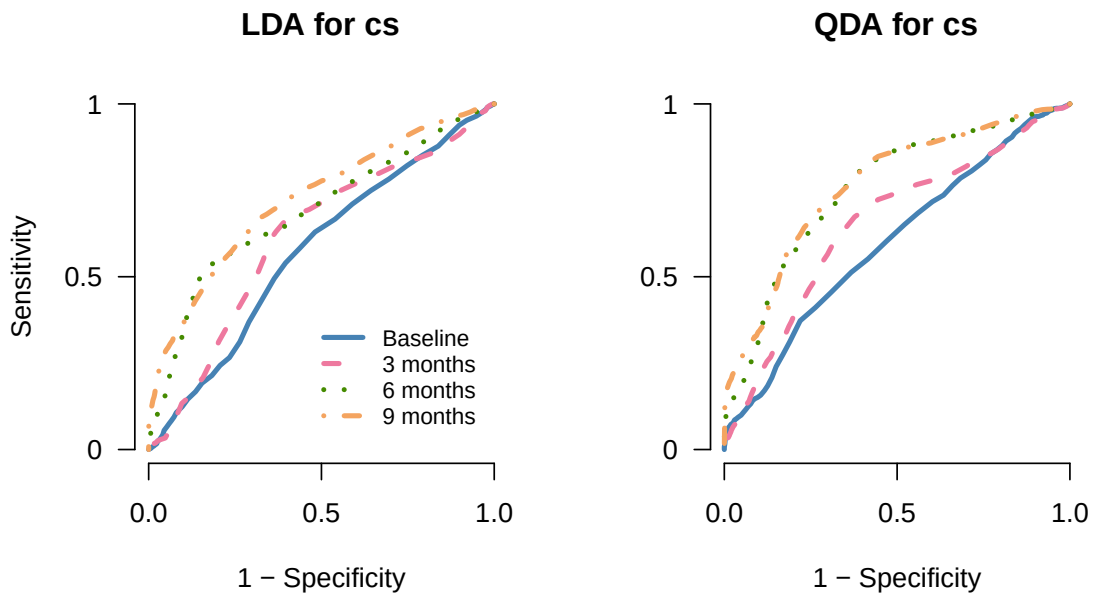

Figure 3.3: ROC curves for the linear discriminant analysis (left panel), and for the quadratic discriminant analysis(right panel) using contrast sensitivity.

performance of the classification tool with increased longitudinal information.

The quadratic discriminant model using CS data at 0, 3, 6 and 9 months shows improvement in AUC (77%) compared to the CS linear discriminant model (73%) (Table 3.4 and Figure 3.3). This result is consistent with the results of the box's M test and Bartlett's test presented in the Table 3.1 and Table 3.3. On the other hand, the discriminant results involving VA (Table 3.4) show that linear and quadratic functions provided very similar results in AUC, despite the test for equality of covariances was rejected for VA which suggest evidence of inequality of covariances and the suitability of QDA over LDA (Figure 3.4).

Since the definition of treatment success is based on VA, including VA provides better performance than if VA is not included. There are other examples in the literature where the longitudinal predictor is used (in combination with other biomarkers) to predict a clinical outcome that is defined based on the predictor itself. For example,

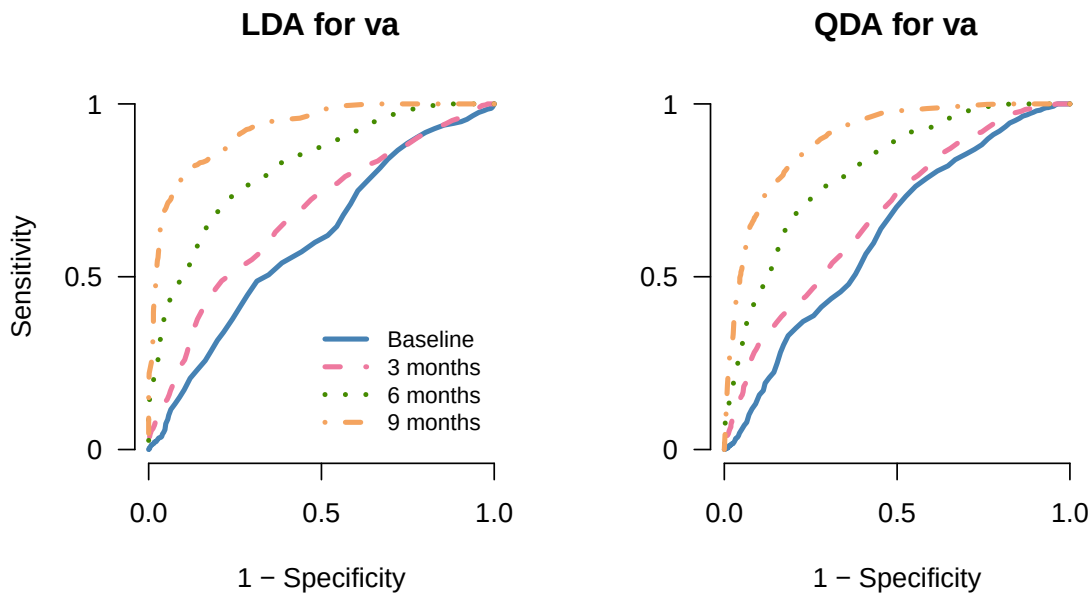

Figure 3.4: ROC curves for the linear discriminant analysis (left panel), and for the quadratic discriminant analysis(right panel) using visual acuity.

García-Fiñana et al. (2019) used the level of diabetic retinopathy (categorical variable) measured over time, within a discriminant model together with other risk factors (type of diabetes, ethnicity, HbA1c, etc.) to predict patients who will develop sight threatening diabetic retinopathy (STDR) within a year. STDR is defined when a certain level of diabetic retinopathy is achieved. The idea is that changes in the variable, that are less than the changes needed to define a group membership may be informative about progression. Nevertheless, when comparing the models described in Table 3.4, it is expected that models that make use of more recent data (e.g., data collected at 0, 3, 6 and 9 months to predict treatment success/failure at 12 months) show higher predictive accuracy levels.

Hastie et al. (2009) mentioned that increasing the number of parameters can affect the decision boundaries. The decision boundaries are functions of the parameters of the estimated densities. The QDA requires more parameters to be estimated than the LDA, which could be a reason why the LDA approach gives better or similar

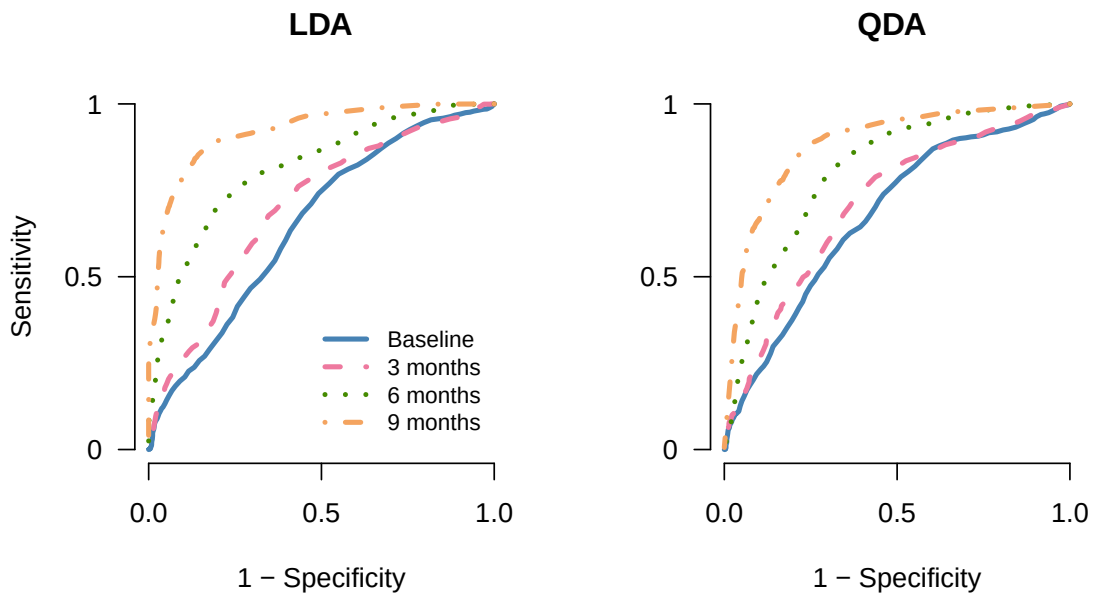

Figure 3.5: ROC curves for the multivariate models. ROC curves for the linear discriminant analysis (left panel) and for the quadratic discriminant analysis (right panel).

classification for some of the models when compared to QDA. There is another reason for LDA to have better performance in small sample sizes in that it might be expected to have greater across-sample stability of results (Huberty and Curry (1978); Michaelis (1973)). Huberty and Curry (1978) showed that both LDA and QDA performed with similar classification results, with LDA most often being the best in nearly all of seven situations from three sets of real data (equal and unequal covariance matrices and two and three criterion groups). Interestingly, the LDA was recorded between the best three classification approaches for 7 of the 22 datasets and QDA between three for four datasets in the STATLOG project (Michie et al. (1994), Hastie et al. (2009)).

On the other hand, in our first model of the univariate case with two variables (CS and age at baseline) used for prediction, the results show that LDA performs less well than the QDA (i.e., the levels of sensitivity, specificity, PCC, PPV and AUC for the LDA model are lower compared with the QDA model). In this case, three parameters need to be estimated in the LDA model, and there are six parameters that need to be

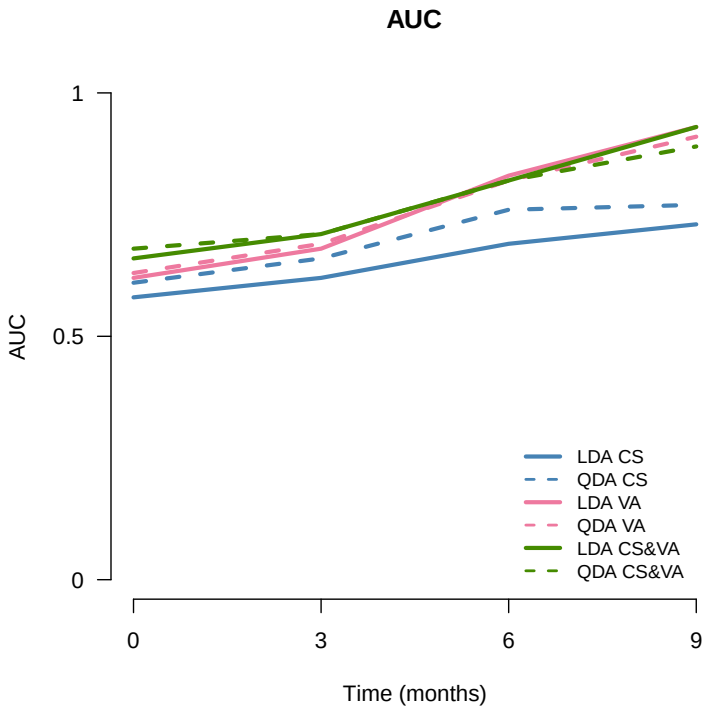

Figure 3.6: Comparison of AUC for linear and quadratic functions that used CS and VA (separately and together), measured using different follow-ups, to predict treatment success or treatment failure.

estimated in the QDA model. This is not a large difference in the number of parameters; hence it is not surprising that QDA performs better than LDA.

To further visualise our comparisons of LDA and QDA, Figure 3.6 was constructed. It shows the AUCs over time for the LDA and QDA models (univariate and multivariate) using the balanced dataset D2. It can be seen that using both CS and VA in one model (green lines in the Figure 3.6) improves the prediction at early time points.

The classical discriminant analysis (linear and quadratic) has limitations since the data must be complete (i.e., cannot handle missing data) and balanced (patients must all be examined at the same time points or their visits approximated to some fixed time points). Also, they do not deal with the correlation between repeated measurements on the same subject (Roy (2006)). One way to overcome the limitation is to model the longitudinal responses in a suitable longitudinal model that accounts for the correlation between repeated measurements on the same patient. This can be achieved through

the use of linear mixed models (see Tomasko et al. (1999)). I name such an approach to *modified discriminant analysis* and explore it in the next section.

3.4 Comparison of classical and modified discriminant analysis

In the previous section, both LDA and QDA performed with similar classification results (as shown in Table 3.4), and the combined evidence of Tables 3.2 and 3.3 gave an unclear picture about the appropriateness of assuming equal covariances. This section, focuses on a comparison of the classical discriminant methods and a modified discriminant approach. QDA is chosen in order to allow greater flexibility and to satisfy concerns about the equality of covariances assumption. The ophthalmology dataset is used to predict patients whose vision improve or do not improve (i.e., success or failure of treatment).

The classical method requires a complete and balanced dataset to predict treatment failure. The particular aim of this section is to examine whether using the exact visit times, and modelling correlation between repeated measurements gives more accurate predictions than simply using the classical approach with approximated times and imputed data.

Hence the sample data consisted of 447 patients (see Section 3.2 for a detailed description of the dataset). The comparison of this section can be summarised into two main points:

- Comparison between classical discriminant analysis using the balanced dataset (D2 in the the Figure 3.1), which approximated the time points to 0, 3, 6, 9 and 12 months, and a modified approach applied to the unbalanced dataset (D1) which used the exact time points. In other words, I evaluated the effect of approximating the visit times of patients followed in the clinic.

- The second comparison is shown in Figure 3.2, which compares the classical discriminant analysis applied to the balanced and imputed data (D4) with the modified discriminant analysis approach applied to the unbalanced data (D3) (which used the exact time points). In other words, I evaluated whether imputing missing data improves classification ability compared to simply using the observed data.

This comparison between two approaches was evaluated using several measures: sensitivity, specificity, PCC, AUC, PPV and NPV with each of these measures calculated at the optimal cut off value (described in Chapter 2). Additionally, I divided the sample into 80% for training, and 20% for testing the predictions for new patients. Both approaches (modified and classical QDA) were repeated 100 times to remove possible biases form selected training or testing sets, and the overall average of the 100 repeats was recorded.

3.4.1 Modified discriminant analysis

The modified discriminant analysis is a type of longitudinal discriminant analysis (LoDA) that uses a patient's longitudinal history to predict which group a patient belongs to. The modified discriminant analysis was applied to the unbalanced datasets (D3 and D1) to classify patients. The modified discriminant analysis approach consists of two steps.

First step: multivariate linear mixed effects model (MLMM)

The first step is the multivariate linear mixed-effects model which was used to account for the correlation between repeated observations of the same marker, and between markers, on the same patient. Here I displayed the models and parameters estimations for each diagnostic group (success or failure of the treatment). The MLMMs were applied to the unbalanced datasets (D1 in Figure 3.1 and D3 in Figure 3.2). The MLMMs were fitted to the training dataset (80%) using the `lme` function from the `R` package

`nlme` Pinheiro et al. (2014). Prediction of the testing dataset (20%) was performed using the `dMVN` function from the `mixAK` Komárek and Komárková (2014). The Equation 3.2 shows the multivariate linear mixed model (MLMM) based on longitudinal measurements of contrast sensitivity (CS) and visual acuity (VA).

$$\begin{pmatrix} CS_{ij}^{g} \\ VA_{ij}^{g} \end{pmatrix} = \begin{pmatrix} \beta_{0cs}^{g} \\ \beta_{0va}^{g} \end{pmatrix} + \begin{pmatrix} \beta_{1cs}^{g} \\ \beta_{1va}^{g} \end{pmatrix} t_{ij} + \begin{pmatrix} \beta_{2cs}^{g} \\ \beta_{2va}^{g} \end{pmatrix} \text{Age} + \begin{pmatrix} b_{0i_{cs}}^{g} \\ b_{0i_{va}}^{g} \end{pmatrix} + \begin{pmatrix} b_{1i_{cs}}^{g} \\ b_{1i_{va}}^{g} \end{pmatrix} t_{ij} + \begin{pmatrix} \varepsilon_{ij_{cs}}^{g} \\ \varepsilon_{ij_{va}}^{g} \end{pmatrix} \tag{3.2}$$

where g refers to the groups ($g = 0, 1$), $i = 1, \cdots, n$, $j = 1, \cdots, n_i$. In Equation 3.2 t_{ij} is the time in months in which marker was recorded (from baseline), and *Age* is the age in years of the patient at baseline. The MLMM consisted of a four dimensional vector of random effects (Equation 3.2), with a random intercept for each marker (b_{0ics}, b_{0iva}) and also a random slope for each marker (b_{1ics}, b_{1iva}). Also, the MLMM included six fixed effects (β_{0cs}, β_{0va}, β_{1cs}, β_{1va}, β_{2cs}, β_{2va}) which denote the fixed intercept, time and age effects for each marker. The residual errors (ε_{ijcs}, ε_{ijva}) were assumed to be independent and identically distributed following a Gaussian distribution.

The results of the longitudinal multivariate linear mixed model (MLMM) for the unbalanced dataset D3 to predict treatment success or treatment failure are provided in Table 3.5. For patients who had successful treatment there was no significant change in CS and VA over time (P-value = 0.28 and 0.86 respectively). Older patients at baseline generally had lower CS and VA scores ($\beta_{0cs} = 34.15$, $\beta_{0va} = 71.52$, $\beta_{2cs} = -0.14$, $\beta_{2va} = -0.31$). The profiles of both CS and VA remained relatively constant during the year of follow up in patients for whom the treatment was successful (see Figure 1.2 in Chapter 1). In addition, Table 3.5 shows that for patients whose treatment failed, there was a significant decrease in both CS and VA over time (β_1 = -0.38 and -2.859, P-values < 0.001, respectively) suggesting a worsening condition. Age at entry did not have a significant effect on VA measurements but in general older patients at baseline had worse CS scores (β_1 = -0.16 and P-value = 0.01).

Table 3.5: Fixed effects parameters of the MLMM based on contrast sensitivity (CS) and visual acuity (VA) for the prognostic groups (i.e., success and failure of treatment).

Fixed effect parameter for success of treatment $n_0 = 264$						
Markers	Covariates	Coefficient	Standard error	t-value	95%CI	P-value
CS	intercept β_0	34.15	3.13	10.90	28.01 40.29	< 0.001
	slope (time) β_1	0.04	0.04	1.08	-0.03 0.12	0.28
	age (year) β_2	-0.14	0.04	-3.48	-0.22 -0.06	< 0.001
VA	intercept β_0	71.52	6.68	10.69	58.40 84.64	< 0.001
	slope (time) β_1	-0.01	0.08	-0.17	-0.18 0.15	0.86
	age (year) β_2	-0.31	0.08	-3.58	-0.48 -0.14	<0.001
Fixed effect parameter for failure of treatment $n_1 = 183$						
CS	intercept β_0	33.04	5.17	6.38	22.89 43.19	<0.001
	slope (time) β_1	-0.38	0.071	-5.37	-0.52 -0.24	<0.001
	age (year) β_2	-0.16	0.06	-2.44	-0.29 -0.03	0.01
VA	intercept β_0	54.60	8.99	6.07	36.95 72.25	<0.001
	slope (time) β_1	-2.86	0.12	-23.09	-3.10 -2.61	< 0.001
	age (year) β_2	-0.05	0.11	-0.48	-0.28 0.17	0.63

Second step: quadratic discriminant analysis

In the previous section, the MLMM was developed for each group from the training dataset to describe the change of longitudinal markers (as the first step of the modified discriminant analysis approach). The second part of the modified discriminant analysis is to construct the quadratic discriminant analysis to predict the patient's eye status at 12 months. Appendix A shows how to set the second step of the modified discriminant analysis in `R`. In Section 2.2.2 the methodology of the quadratic discriminant analysis (QDA) was described. The discriminant function was developed based on the mean and covariance matrices estimates obtained from the MLMM.

For each comparison, sensitivity, specificity, PCC, PPV and NPV were computed using the optimal cut-off value. The AUC was also calculated for each comparison. Table 3.6 provides the results for classical discriminant analysis that used the balanced data (D2) and the modified discriminant analysis that used the unbalanced data (D1), while Table 3.7 displays the results for the unbalanced data (D3) using a modified discriminant analysis and the balanced and imputed data (D4) using a classical discriminant analysis.

In describing the comparison results, I first addressed whether approximating the

Table 3.6: Accuracy measures for the modified QDA that used the unbalanced dataset (D1) and for the classical QDA that used the balanced data (D2). Note that both dataset use the same patients and same measurements.

Modified quadratic discriminant analysis (D1, n = 176)

prediction time	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
0	0.40	0.63	0.66	0.65	0.64	0.58	0.72
0, 3	0.34	0.75	0.73	0.74	0.74	0.67	0.81
0, 3, 6	0.37	0.82	0.77	0.79	0.83	0.73	0.86
0, 3, 6, 9	0.34	0.89	0.84	0.86	0.91	0.80	0.92

Classical quadratic discriminant analysis (D2, n = 176)

prediction time	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
0	0.40	0.69	0.67	0.68	0.68	0.61	0.76
0, 3	0.34	0.73	0.70	0.71	0.72	0.64	0.79
0, 3, 6	0.33	0.79	0.77	0.78	0.81	0.72	0.84
0, 3, 6, 9	0.31	0.86	0.85	0.85	0.88	0.81	0.89

time point and applying the classical discriminant analysis improves in sensitivity, specificity and PCC of prediction (Table 3.6, Figure 3.7). The results show that the classical discriminant analysis has slightly better performance values: sensitivity, specificity, PCC, and AUC at baseline compared with the modified discriminant analysis that used unbalanced data (D1). Recall, in the unbalanced data (D1, $n = 176$) all patients must have had a visit at baseline and at least four follow-up visit near to 3, 6, 9 and 12 months and in the balanced data (D2, $n = 176$) I approximated the four visits measurements of the 176 patients (from D1) to 3, 6, 9 and 12 months. The classical quadratic discriminant model at baseline involved CS, VA and age at baseline and required fewer parameters than the modified quadratic discriminant analysis based on the MLMM (Equation 3.2), which can explain why classical QDA performed slightly better here than the modified QDA. However, the results slightly change when models involve more longitudinal information (more visit points 3, 6 and 9 months added); the modified discriminant analysis has better classification accuracy than the classical QDA in terms of AUC (91% and 88%, respectively).

When there is a small number of visits available (e.g., 1 or 2), then there is little

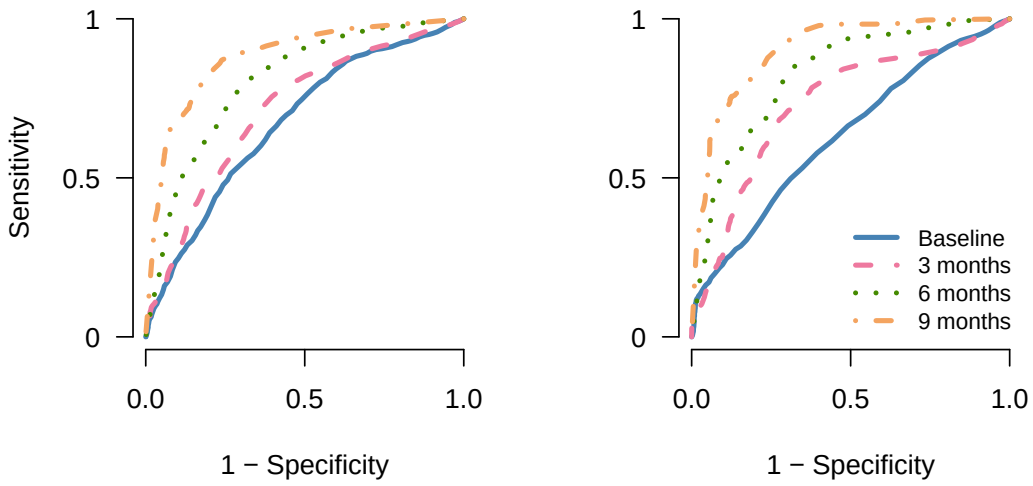

Figure 3.7: Receiver operating characteristic (ROC) curves of the classical discriminant model using the balanced data (D2, left plot) and of the modified discriminant model using the unbalanced data (D1, right plot) at four time points.

difference between approximating the visit times to the nearest 3 months scheduled visit and using a more accurate mixed model (see Table 3.6). With a small amount of data, more simple models are just as accurate. However, when more data is available for each patient, the more accurate mixed models outperform the classical techniques.

Figure 3.7 shows the four ROC curves for the classical QDA (left plot) applied to the balanced dataset (D2) and for the modified QDA (right plot) applied to the unbalanced dataset (D1) at different time points.

Figure 3.8 shows the AUCs for the two models and suggests that approximating the time points to have a balanced dataset does not seem to have much effect on the prediction accuracy.

Next, I compared the results of the modified approach applied to the unbalanced data (D3) and the classical approach applied to the balanced, imputed data (D4). In this case, the classical approach offers similar classification results to the modified approach at baseline prediction, with 64% of patients correctly classified (Table 3.7,

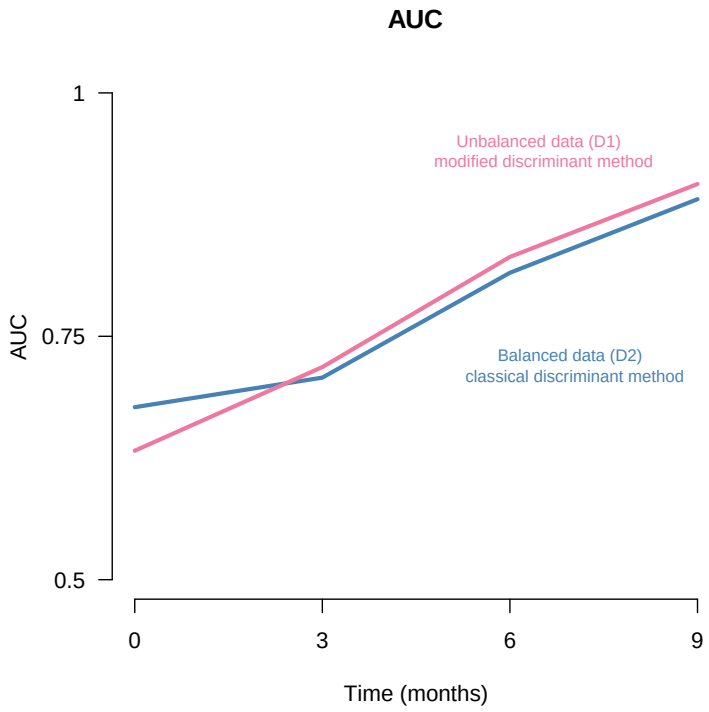

Figure 3.8: Area under ROC curve (AUC) for the modified and the classical discriminant analysis approaches to predict failure of treatment over time.

Figure 3.9). The prediction accuracy shows slight improvements when adding more longitudinal measurements (time points 3, 6 and 9 months) to the classical model, with 20% of patients classified incorrectly (using data up to 9 months) compared with prediction at baseline (36% of patients are misclassified).

The modified approach gives better predictions than the classical approach when the model involved all follow-up measurements (up to 9 months), with 89% sensitivity (89% of patients who did not improve their eye vision at 12 months were correctly classified), and 89% specificity (89% of patients who improved their eye vision at 12 months were correctly classified). The overall correct classification is 89% (PCC). The comparison in AUC between the modified approach and the classical approach at prediction time (0, 3, 6 and 9 months) shows that the modified approach performs better.

Figure 3.10 presents the AUC for the modified QDA and the classical QDA on longitudinal data with unbalanced design and imputed design. The classification using the modified discriminant analysis increases the AUC to approximately 93% compared

Table 3.7: Accuracy values of the modified QDA model that used the unbalanced dataset (D3) and of the classical QDA that used the balanced and imputed data (D4).

Modified quadratic discriminant analysis (D3, N = 447)

prediction time	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
0	0.40	0.64	0.64	0.64	0.65	0.56	0.72
0, 3	0.33	0.73	0.72	0.73	0.75	0.65	0.80
0, 3, 6	0.39	0.79	0.81	0.80	0.85	0.75	0.84
0, 3, 6, 9	0.34	0.89	0.89	0.89	0.93	0.85	0.92

Classical quadratic discriminant analysis (D4, N = 447)

prediction time	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
0	0.40	0.67	0.62	0.64	0.66	0.56	0.73
0, 3	0.32	0.70	0.68	0.69	0.74	0.61	0.77
0, 3, 6	0.37	0.74	0.80	0.77	0.82	0.73	0.81
0, 3, 6, 9	0.30	0.77	0.83	0.80	0.84	0.76	0.84

to AUC of the classical QDA of 84% when using the time points at 0, 3, 6 and 9 months. Also, there is no substantial advantage to imputing missing CA and VA value as the modified approach can handle missing values.

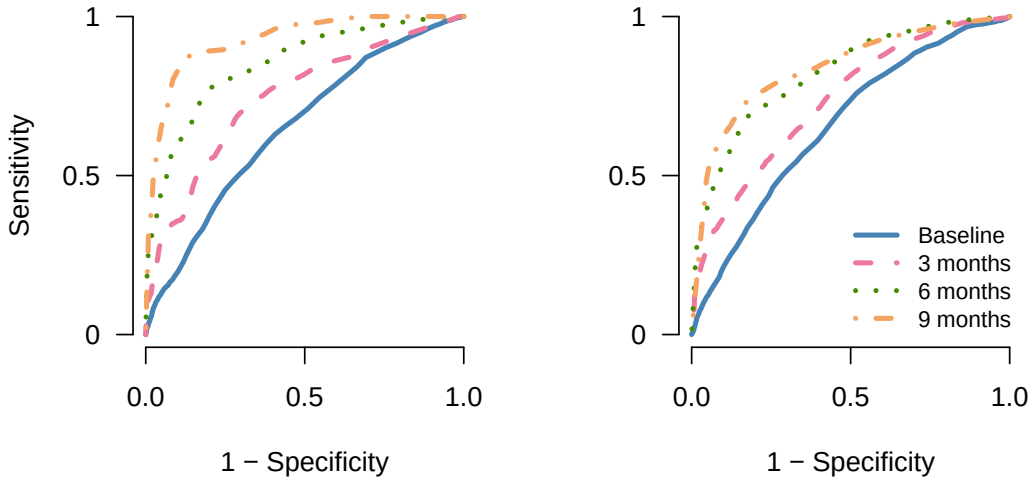

Figure 3.9: Receiver operating characteristic (ROC) curves of the modified discriminant analysis for the unbalanced dataset (D3, left plot) and of the classical discriminant analysis for the balanced, imputed dataset (D4, right plot) at four time points.

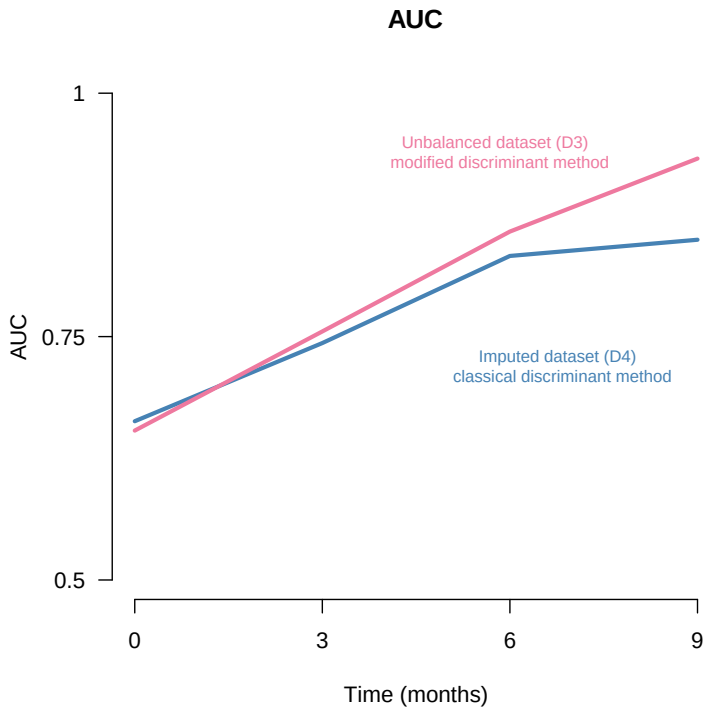

Figure 3.10: Area under ROC curve (AUC) for the modified and the classical approaches that applied to the unbalanced dataset (D3) and balanced, imputed dataset (D4), respectively to predict failure of treatment over time.

3.5 Simulation Study

In Section 3.4, I showed a linear mixed model to model the AMD longitudinal data allowed more accurate classification than imputing missing observations and using a classical discriminant analysis. I now use a simulation study to investigate whether the amount of missing observations influences the performances of the classical and modified discriminant analysis approaches.

This simulation study focuses on a comparison of the classical discriminant method and a modified discriminant approach. The classical method requires a complete and balanced dataset. In this simulation, I imputed the missing values using the last observation carried forward to predict treatment failure, while the modified discriminant analysis can deal with missing values. The simulation is based on the AMD dataset (which includes data collected over 12 months). I simulated a dataset consisting of two continuous markers: contrast sensitivity and visual acuity. I created three simulated scenarios. I assumed two groups for discrimination, Group 0 and Group 1. To be consistent with the AMD data, I assume that each simulation consisted of 200 patients (40% of patients) whose vision improved and 300 patients (60% of patients) whose vision did not improve before a year. The original data was collected at four-time points ($n_i = 4$) at baseline, then after approximately three months, six months and nine months. For each patient, the four times were generated as follows: the first time visit was set to 0 and uniform distributions in the intervals (70, 110), (160, 200) and (250, 290) days were used to generate the remaining visit times. The elements of mean vectors and variance-covariance matrices considered for the two markers for each group are presented in Table 3.8. At each time point I simulated values for each marker, by first generating random effects from a multivariate normal distribution with mean vector and covariance matrix given in Table 3.8. I also assumed that all patients had the same number of visits.

$$\begin{pmatrix} CS_{ij}^{g} \\ VA_{ij}^{g} \end{pmatrix} = \begin{pmatrix} b_{0i_{cs}}^{g} \\ b_{0i_{va}}^{g} \end{pmatrix} + \begin{pmatrix} b_{1i_{cs}}^{g} \\ b_{1i_{va}}^{g} \end{pmatrix} t_{ij} + \begin{pmatrix} \varepsilon_{ij_{cs}}^{g} \\ \varepsilon_{ij_{va}}^{g} \end{pmatrix} \tag{3.3}$$

A total of 100 simulated datasets were generated using the MLMM described in Equation 3.3. Three simulation scenarios studied the effect of missing data on classification accuracy. In the first scenario, 10% of the data are removed randomly by using the `sample` function from `R` (version 3.4.3). For the second and third scenarios, the percentage of missing data is increased to 20% and 40% respectively to investigate which approaches (modified discriminant analysis or classical discriminant analysis) are more robust to missing data. All patients must have a visit at baseline.

Table 3.8: Parameter estimates for each simulation scenario.

	Group 0	Group 1
Contrast sensitivity		
E(intercept:contrast sensitivity)	23.3	20.5
E(slope:contrast sensitivity)	0.0469	-0.379
SD(intercept:contrast sensitivity)	5.4	5.46
cor(intercept:contrast sensitivity), slope:contrast sensitivity)	-0.327	0.159
cor(intercept:contrast sensitivity, intercept:visual acuity)	0.493	0.622
cor(intercept:contrast sensitivity, slope:visual acuity)	0.0831	0.231
SD(slope:contrast sensitivity)	0.359	0.565
cor(slope:contrast sensitivity, intercept:platelet)	-0.338	0.184
cor(slope:contrast sensitivity,slope:platelet)	0.549	0.893
SD(contrast sensitivity:residual)	3.21	6.13
Visual acuity		
E(intercept:visual acuity)	47.7	50.3
E(slope:visual acuity)	-0.0188	-2.86
SD(intercept:visual acuity)	10.5	9.03
cor(intercept:visual acuity, slope:visual acuity)	0.345	0.241
SD(slope:visual acuity)	0.852	0.939
SD(visual acuity:residual)	4.31	7.68

The predictions were assessed using 90% of the data for training and 10% for testing. The sensitivity, specificity, PCC, PPV and NPV were measured for each simulated dataset using the optimal cutoff value. Also, the AUC was measured for each simulation. These measurements were therefore used to compare the discriminant approaches, based on the average of 100 simulated datasets. Both models (modified and classical

QDA) were generated 100 times to minimise possible biases affecting the selected training or testing sets for each simulated dataset.

Table 3.9: Accuracy values of the modified QDA model that used the unbalanced dataset and of the classical QDA that used the balanced and imputed data. (Simulation study when 10% of data are missing).

Classical quadratic discriminant analysis

prediction time	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
0	0.36	0.68	0.74	0.71	0.72	0.64	0.78
0, 3	0.36	0.76	0.80	0.78	0.81	0.72	0.83
0, 3, 6	0.37	0.80	0.84	0.83	0.86	0.78	0.87
0, 3, 6, 9	0.38	0.84	0.87	0.85	0.90	0.81	0.89

Modified quadratic discriminant analysis

prediction time	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
0	0.36	0.63	0.71	0.68	0.72	0.59	0.74
0, 3	0.37	0.72	0.77	0.75	0.81	0.68	0.81
0, 3, 6	0.37	0.78	0.82	0.80	0.87	0.74	0.85
0, 3, 6, 9	0.37	0.82	0.85	0.84	0.91	0.79	0.88

Table 3.9 provides the results of the comparison between the classical discriminant analysis and the modified discriminant analysis when 10% of the data are missing. The classical QDA used the balanced and imputed data and the modified discriminant analysis that used the unbalanced data. When 10% of the observations were missing the performance of the two approaches was largely similar based on the AUC (see Table 3.9). As expected, the addition of more visits improved the classification performance of each approach (Figure 3.11).

In the case where 20% of the data are missing, the results for the classical discriminant analysis are very similar to the modified discriminant analysis when models involve more longitudinal measurements (data up to 9 months) with a slight improvement in AUC for the modified approach (see Table 3.10). Figure 3.12 shows the four ROC curves for the classical QDA (left plot) applied to the balanced and imputed dataset and for the modified QDA (right plot) applied to the unbalanced dataset at different time points where 20% of the data are missing.

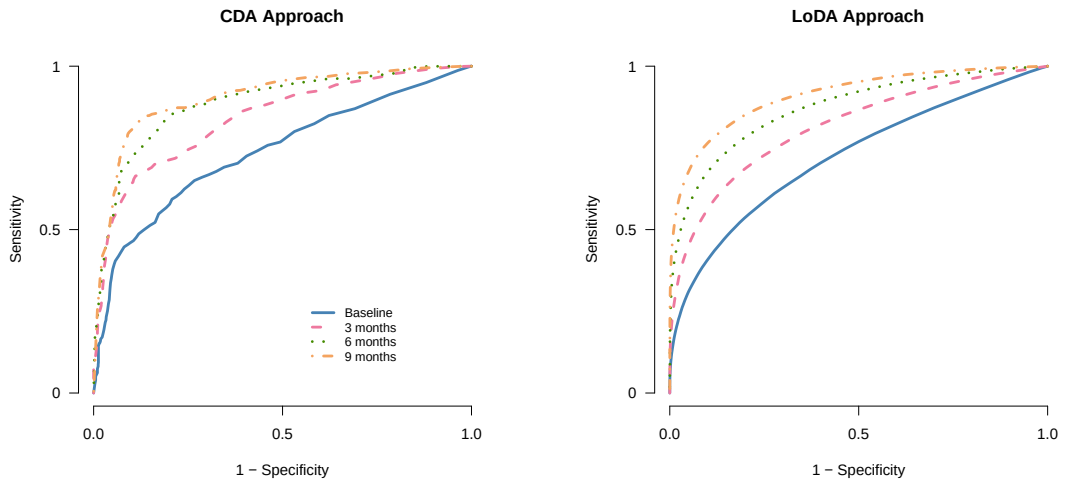

Figure 3.11: ROC curves of the modified discriminant analysis (right plot) and the classical discriminant analysis (left plot) at four time points (10% of data missing).

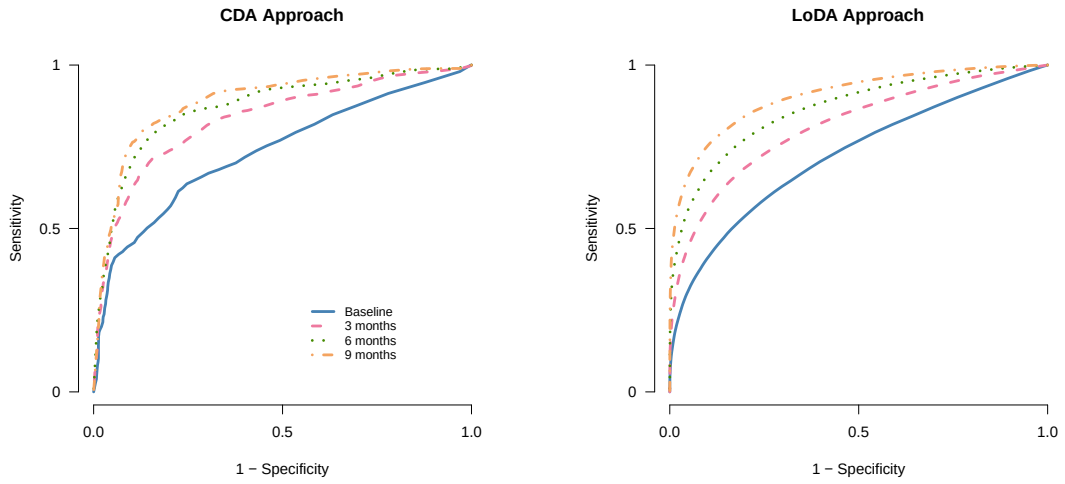

Figure 3.12: ROC curves of the modified discriminant analysis (right plot) and the classical discriminant analysis (left plot) at four time points (20% of data are missing).

The modified approach offers better predictions than the classical approach when 40% of the data are missing (Table 3.11) and the model incorporated all follow-up measurements (up to 9 months), with 80% of patients who did not improve their eye vision at 12 months being correctly classified (80% sensitivity), and 84% of patients who improved their eye vision at 12 months being correctly classified (84% specificity). The overall correct classification is 82% (PCC) and the AUC is 0.89.

Figure 3.13 shows the four ROC curves for the classical approach and the modified

Table 3.10: Accuracy values of the modified QDA model that used the unbalanced dataset and of the classical QDA that used the balanced and imputed data. (Simulation study when 20% of data are missing).

Classical quadratic discriminant analysis

prediction time	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
0	0.36	0.67	0.74	0.71	0.72	0.65	0.78
0, 3	0.36	0.75	0.80	0.78	0.81	0.72	0.83
0, 3, 6	0.37	0.79	0.83	0.81	0.85	0.76	0.86
0, 3, 6, 9	0.38	0.82	0.85	0.84	0.88	0.79	0.88

Modified quadratic discriminant analysis

prediction time	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
0	0.36	0.63	0.71	0.68	0.72	0.59	0.74
0, 3	0.36	0.72	0.77	0.75	0.81	0.68	0.81
0, 3, 6	0.36	0.78	0.81	0.80	0.87	0.73	0.84
0, 3, 6, 9	0.37	0.82	0.85	0.84	0.91	0.78	0.87

approach at four different time points when 40% of the data are missing.

Figure 3.14 presents the comparison of the AUC for the modified QDA and the classical QDA on longitudinal data with 10% (left plot), 20% (middle plot) and 40% (right plot) of the data missing. Figure 3.14 shows that the modified approach provides better classification accuracy compared with the classical approach.

Table 3.11: Accuracy values of the modified QDA model that used the unbalanced dataset and of the classical QDA that used the balanced and imputed data. (Simulation study when 40% of data are missing).

Classical quadratic discriminant analysis

prediction time	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
0	0.36	0.67	0.74	0.71	0.72	0.64	0.78
0, 3	0.36	0.74	0.79	0.77	0.79	0.71	0.82
0, 3, 6	0.37	0.75	0.80	0.78	0.81	0.73	0.83
0, 3, 6, 9	0.40	0.77	0.81	0.80	0.82	0.74	0.84

Modified quadratic discriminant analysis

prediction time	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
0	0.36	0.64	0.70	0.68	0.72	0.59	0.74
0, 3	0.37	0.72	0.77	0.75	0.81	0.68	0.81
0, 3, 6	0.37	0.76	0.80	0.79	0.85	0.72	0.84
0, 3, 6, 9	0.37	0.80	0.84	0.82	0.89	0.76	0.86

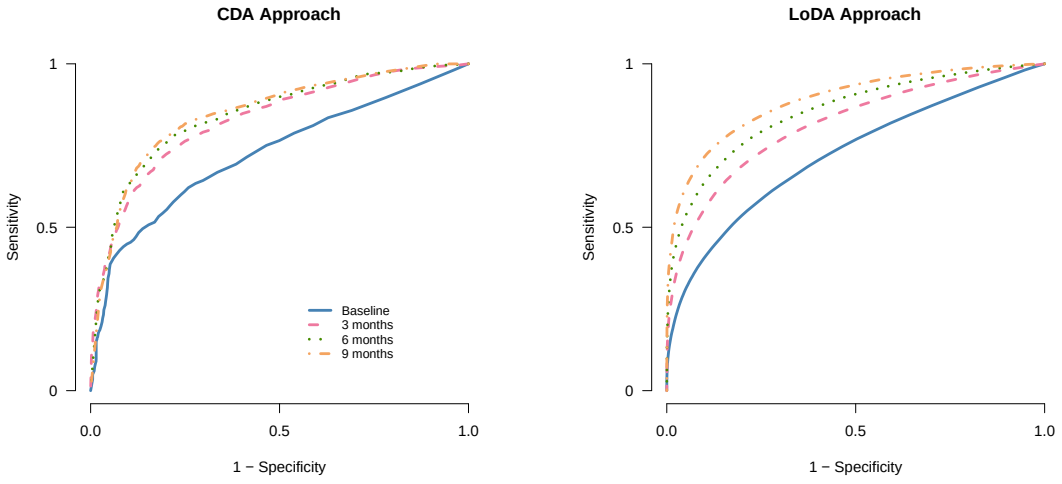

Figure 3.13: ROC curves of the modified discriminant analysis (right plot) and the classical discriminant analysis (left plot) at four time points (40% of data missing).

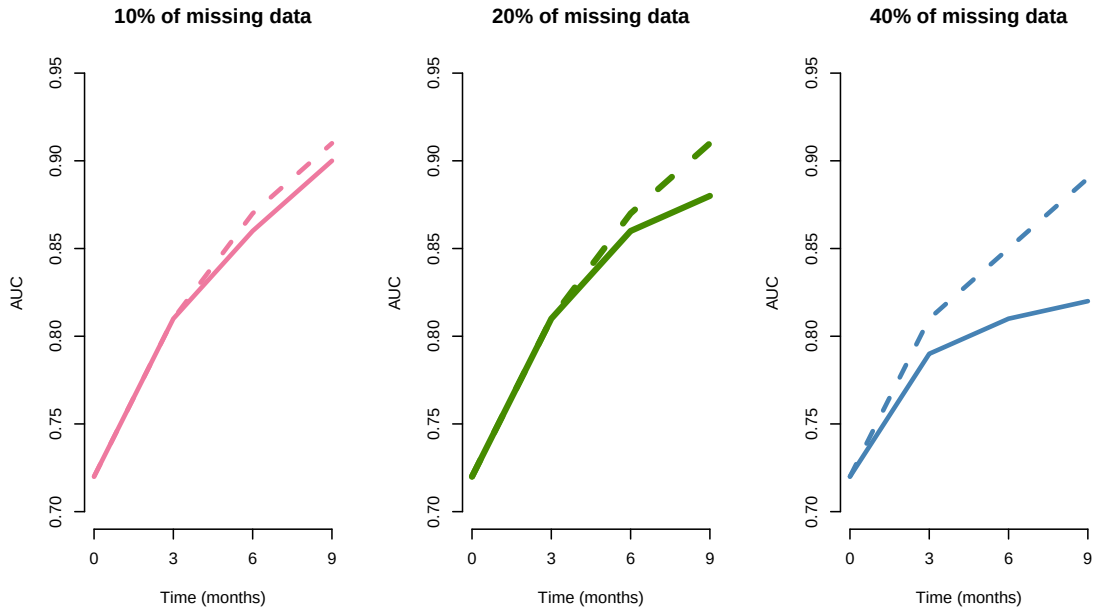

Figure 3.14: Area under ROC curve (AUC) for the modified (dash lines) and the classical (solid lines) approaches with 10%, 20% and 40% of data missing to predict failure of treatment over time.

3.6 Discussion

In this chapter, some of the approaches discussed in this thesis have been applied to the ophthalmic datasets. The classical discriminant analysis has been applied to a complete and balanced dataset, where each follow-up measurements is treated as a single variable. The covariance matrices and means for the classical discriminant analysis are estimated from the data.

The classical linear and quadratic discriminant analysis approaches have limitations. First, classical discriminant analysis requires balanced data with the same time points observed for all subjects, and it therefore cannot handle missing values. A second limitation of applying the classical discriminant analysis is that the LDA assumes the equality of the covariance matrices among two groups, while the QDA allows the covariance matrices differ across groups, and the choice of which one to use is not always clear. A third limitation is that estimation of the covariance matrices form a multivariate longitudinal setting might be be problematic if a large number of parameters need estimation (for example, when many time points are recorded, Tomasko et al. (1999)).

Furthermore, the modified discriminant analysis based on the mixed model has been applied to unbalanced longitudinal datasets to predict the patient's status at 12 months. This approach can deal with missing values and take the correlation between repeated measurements on the same subject into account.

Comparisons between classical QDA and the modified QDA have been presented in this chapter. If the sample size is small (in this case, 176 patients), the classical discriminant analysis gives better classification results at baseline compared with modified discriminant analysis. However, when the sample size is increased to 447 patients, the classical and modified approaches give similar results of classification accuracy using information just at baseline. However, when longitudinal data is used, the modified discriminant analysis offers better classification accuracy. With a small amount of data missing, more simple models (classical approaches) are just as accurate. However, when more data are missing, the more accurate mixed models outperform the classical

techniques. This suggests that the imputation of missing data is not as accurate as modelling the longitudinal trend using a linear mixed model when substantial amounts of observations are missing.

Using a mixed model that accounts for the longitudinal correlation within a dataset allows more accurate classification than simply treating each appointment time as a single variable. Moreover, the use of mixed models gives a more efficient use of the data since patients with missing visits can be counted without the need for imputation. The classification performance using the modified discriminant approach based on the mixed model increases the classification accuracy compared to the classical discriminant approach when using the ophthalmic dataset.

Chapter 4

Comparison of prediction approaches in longitudinal discriminant analysis

4.1 Introduction

In Chapter 2, three ways of using a patient's longitudinal data for the purposes of classification were described. The first, called marginal prediction, focuses on the average change of longitudinal profiles of markers over time, the conditional prediction is based on the growth over time of the patient-specific markers, while the random-effects approach is based on the random-effects distribution of each patient. In this chapter, I further explore the benefits of each approach using simulation studies. This work has been published in (Hughes et al., 2018a). I contributed to the analysis of the data and the simulations, interpretation of the results and helped to write of the manuscript.

This chapter is structured as follows. In Section 4.2, I compare of the three LoDA approaches on the PBC dataset. Section 4.3 presents two different scenarios and compares the three prediction approaches using simulation studies. Finally, this chapter concludes with a discussion.

4.2 Primary Biliary Cirrhosis Data

The work in this chapter uses data from a Mayo Primary Biliary Cirrhosis (PBC) dataset which includes patients with PBC (Dickson et al. (1989), Murtaugh et al. (1994)). Data on a large number of clinical, biochemical, serological and histological parameters were recorded for each of 312 patients, who met eligibility criteria for the randomised placebo controlled trial of the drug D-penicillamine, with a median of 6.3 years follow up. The Mayo PBC data has been used for longitudinal clustering of subjects into a predefined number of groups by Komárek et al. (2013) in which they used three markers (continuous logarithmic serum bilirubin, discrete platelet count and binary blood vessel malformations). This dataset is available in Appendix D of Fleming and Harrington (1991) and electronically at *http://lib.stat.cmu.edu/datasets/pbcseq*, and it is also included within the `mixAK` package in `R` (R Core (2017)). This dataset is used here to explore the three classification approaches for the multivariate longitudinal discriminant analysis with mixed types of markers: continuous, binary and discrete.

In this chapter, I considered the data of 253 patients who were observed for at least 2.5 years, and whose five years status was known. This work aims to predict patients who will not survive or required a liver transplant within five years. In total 202 of 253 patients were classified as known to survive without a transplant after five years (referred as Group 0), while 51 patients died or required a liver transplant at some time between 2.5 and five years (referred as Group 1). Four markers were used in this application, namely, albumin and logarithmic serum bilirubin as continuous markers, the platelet count as a discrete marker (Poisson variable) and a binary marker indicating blood vessel malformations. Figure 4.6 shows the observed longitudinal profiles of the markers.

For each approach and prognostic group, I fitted a multivariate generalised linear mixed model (MGLMM) to the longitudinal data. The GLMM included a random intercept and a random slope for each continuous and count marker, while a random intercept and a fixed effect for time were fitted for a binary marker since it is difficult to

estimate accurately random slopes for dichotomous outcomes with relatively few repeats per individual, see Komárek and Komárková (2014). Here I used a multivariate linear mixed model with one component for the random effects distribution (i.e., K = 1). The general structure of the MGLMM is, therefore:

$$\begin{aligned}
&E(Y_{i,1,j}|\mathbf{b}_{i,1})^g = b^g_{i,1,1} + b^g_{i,1,2}\, t^g_{i,1,j}, && \text{albumin}\\
&E(Y_{i,2,j}|\mathbf{b}_{i,2})^g = b^g_{i,2,1} + b^g_{i,2,2}\, t^g_{i,2,j}, && \log(\text{bilirubin})\\
&\log\{E(Y_{i,3,j}|\mathbf{b}_{i,3})^g\} = b^g_{i,3,1} + b^g_{i,3,2}\, t^g_{i,3,j}, && \text{platelet count}\\
&\text{logit}\{E(Y_{i,4,j}|b_{i,4},\alpha_4)^g\} = b^g_{i,4} + \alpha^g_4\, t^g_{i,4,j} && \text{blood vessel malformations}
\end{aligned} \tag{4.1}$$

where $i = 1, \ldots, N^g$, N^g is number of individuals in group g, g takes the value either 0 or 1 ($g = 0, 1$), $j = 1, \cdots, n_{i,r}$, r refers to the marker $r = 1, 2, 3, 4$, where the total number of markers is 4, $t_{i,r,j}$ is the follow-up time for each marker (which is reported in months). The MGLMM model contains a seven-dimensional vector of random effects (see Equation 4.1), where four of them are random effects intercepts, one for each marker ($b_{i,1,1}, b_{i,2,1}, b_{i,3,1}, b_{i,4}$) and the rest of them ($b_{i,1,2}, b_{i,2,2}, b_{i,3,2}$) are slopes of the random effects for the first three markers. The vector of random effects for ith patient for group g is $\mathbf{b}_i = (\mathbf{b}'_{i,1}, \ldots, \mathbf{b}'_{i,R})'$, where $\mathbf{b} \sim \mathcal{MVN}\big(\boldsymbol{\mu}^g, \mathbb{D}^g\big)$ follows a multivariate normal distribution with a vector of mean $\boldsymbol{\mu}^g$ and covariance matrix $\mathbb{D}^g$. Only blood vessel malformations ($Y_{i,4,j}$) does not have a random effects slope. All four longitudinal markers ($Y_{i,1,j}, Y_{i,2,j}, Y_{i,3,j}, Y_{i,4,j}$) are assumed to be independent given the random effects. The MGLMM includes only one fixed effect, the slope α_4 for binary marker (blood vessel malformations). The first two markers (albumin and log(bilirubin)) which are assumed to follow the Gaussian distribution, have residual errors ($\epsilon_{i,1,j}, \epsilon_{i,2,j}$). It is assumed that these errors are independent and follow normal distributions with mean 0 and variances σ_1^2 and σ_2^2, respectively. Also, these errors are assumed to be independent of the random effects $\mathbf{b}_i$. Section 2.3.4 in Chapter 2 describes the prior distribution of the mixture multivariate generalized linear mixed model. Further details can be found in Komárek and Komárková (2013).

The parameters of the MGLMM model were estimated via Markov chain Monte

Carlo (MCMC). For each MCMC, the results were based on 10,000 iterations of 1:10 thinned MCMC after a burn-in of 500 iterations.

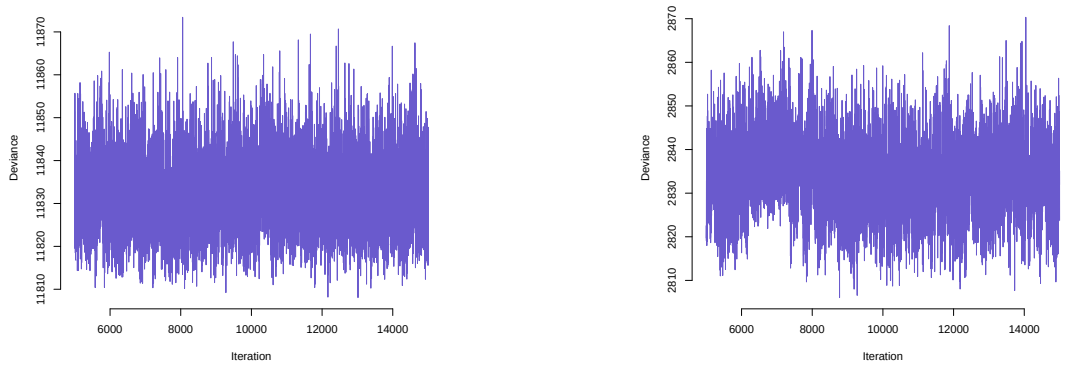

Figure 4.1: Trace plot of the MCMC chain of the model deviance $\mathbb{D}$. Right Figure refers to Group 1 and left Figure refers to Group 0.

Trace plots were used to assess the convergence of the MCMC procedure for each parameter in the model. Figures 4.2, 4.3, 4.4 and 4.5 show these trace plots for models fit to the PBC data and show good convergence of the MCMC procedure.

To evaluate the prediction results leave-one-out cross-validation was applied. The MGLMMs were fitted using the `GLMM_MCMC` function whilst the LoDA was performed using the `GLMM_longitDA2` from the `mixAK` package (Komárek and Komárková, 2014) in `R` (R Core (2017)). Accuracy measurements, mainly the sensitivity, specificity, probability of correct classification (PCC), positive predictive value (PPV) and negative predictive value (NPV) were calculated for each prediction approach by using the optimal cut-off value. The the area under the ROC curve (AUC) was measured for each the three prediction approaches for comparison purposes.

Results for the PBC data are shown in Table 4.1. The marginal approach gives the best prediction results compared to the other two approaches with 81% overall of PCC and the higher value of the AUC with 0.85.

Figure 4.7 shows ROCs for the three LoDA approaches. The ROC curve of the marginal approach is superior in this case. Furthermore, the random effects approach

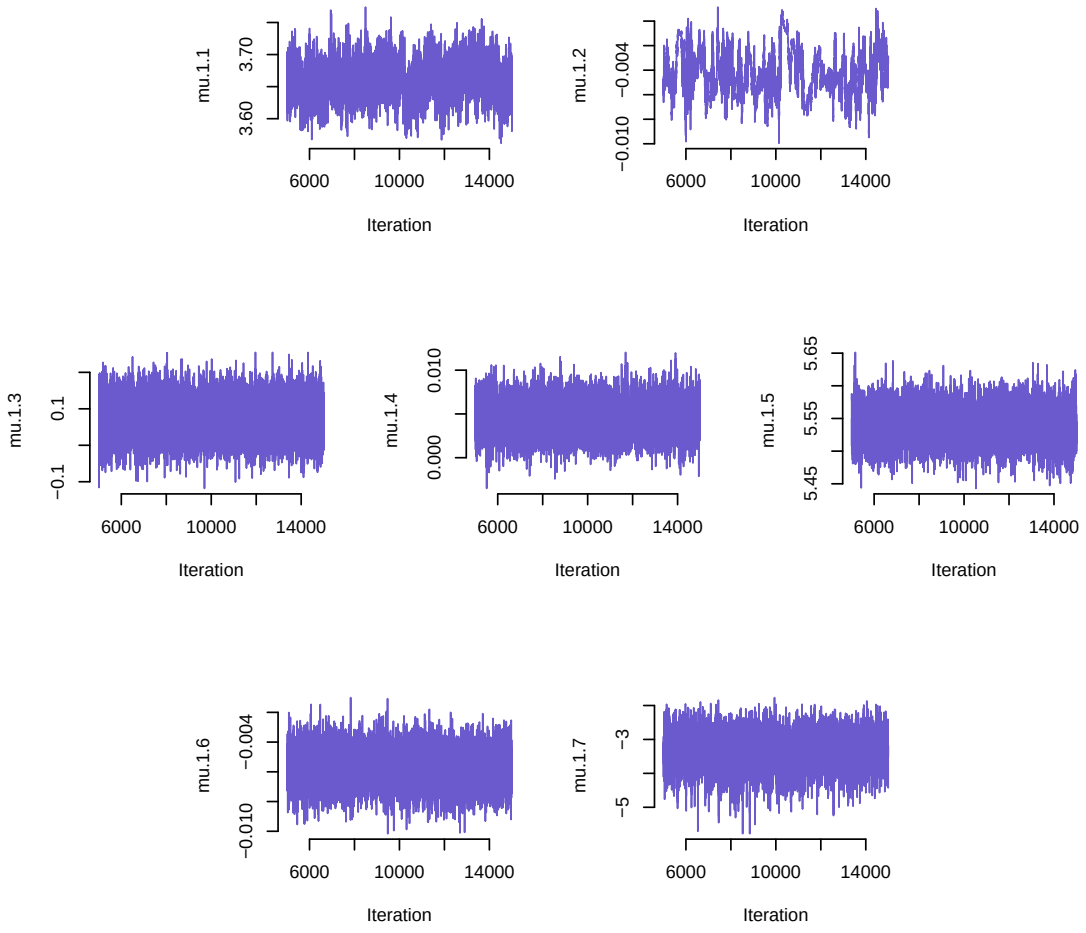

Figure 4.2: Trace plot of the MCMC chain of Group 0 of the model means μ.

also works well in terms of PCC and AUC, with values of 77% and 81% respectively while the conditional approach does not perform well, with 34% of patients classified incorrectly.

The overall mean profiles for three of the markers (albumin, log(bilirubin) and blood vessel malformation) show differences between the two groups (see Figure 4.6). In addition, the markers show variability between the two groups. These two aspects provide a clear explanation of the marginal and the random effects approach work well, since the marginal approach focuses on the average change of the mean of the markers over time, whilst the random effects prediction approach focuses on the patient-specific changes of markers in each group.

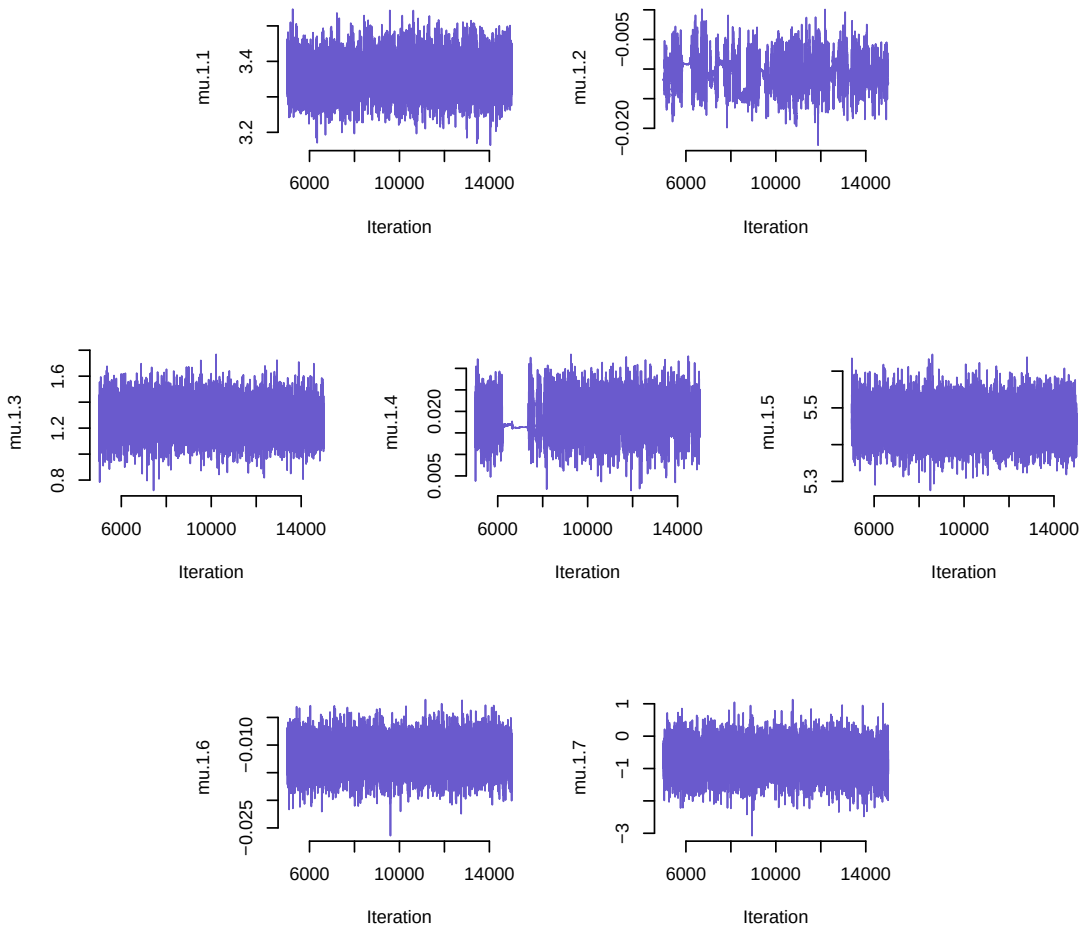

Figure 4.3: Trace plot of the MCMC chain of Group 1 of the model means μ.

A key feature of the random effects prediction approach is that it requires good estimates of the patients' random effects. An increasing number of visits per patient is expected to result in better estimates of the random effects for each individual. This point is further investigated using the full PBC dataset. In the previous application to the PBC data, the patients were followed up to 2.5 years with an average of approximately four visits per patient, while an average of 7.03 visits per patient was observed with the full PBC dataset.

I investigated the prediction of patients who will not survive or need a liver transplant within the five years using the full PBC dataset. As previously, the 253 patients were divided into two groups: 202 patients were known to survive without a transplant

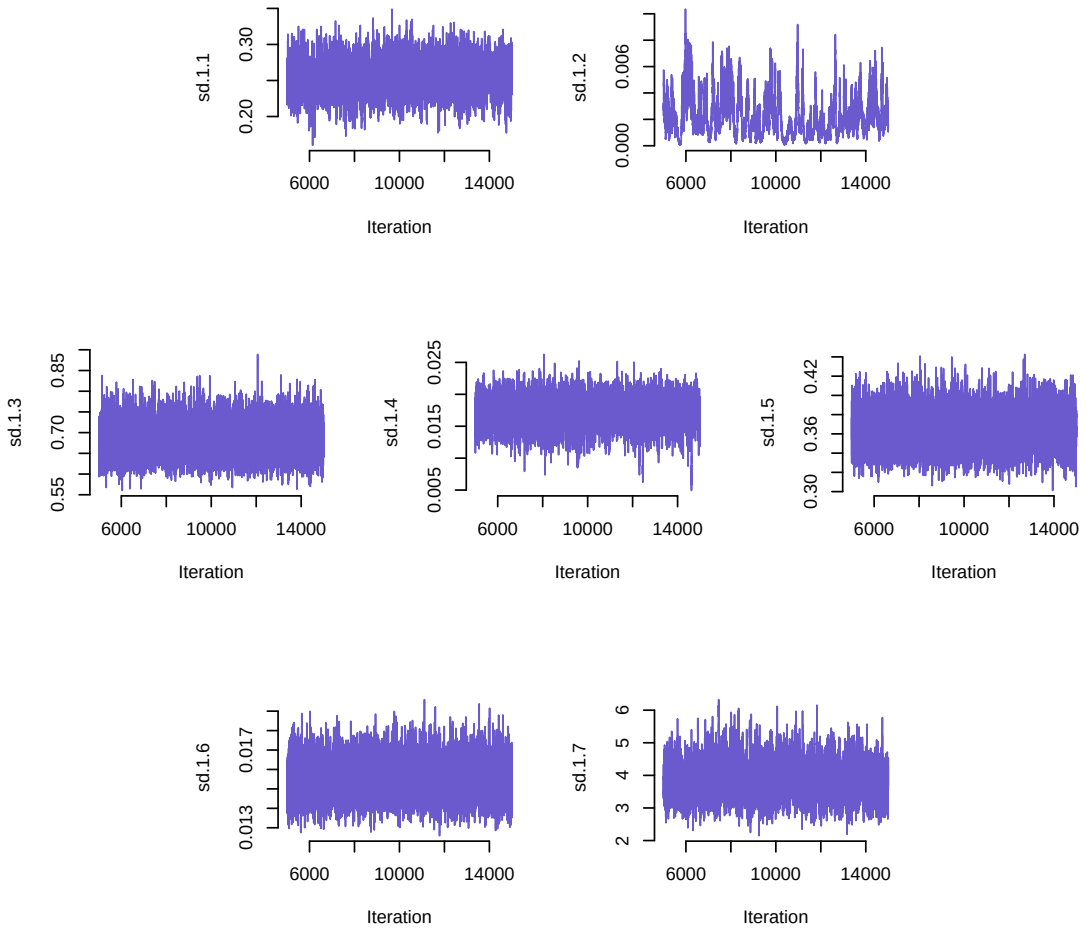

Figure 4.4: Trace plot of the MCMC chain of Group 0 of the model standard deviations derived from the covariance matrices $\mathbb{D}$.

after five years (referred as Group 0) and 51 patients were known not to survive or to need the transplant at some time between 2.5 and 5 years (referred as Group 1). The same markers that have been used in the previous analysis were investigated for this analysis: continuous albumin and logarithmic serum bilirubin, discrete platelet count and the dichotomous indication of blood vessel malformations. Equation 4.1 describes the MGLMM fitted for each prognostic group separately. Table 4.2 summaries the accuracy achieved by each of the three approaches when applied to the full PBC dataset.

The random effects approach shows the best predictions in terms of all measure-

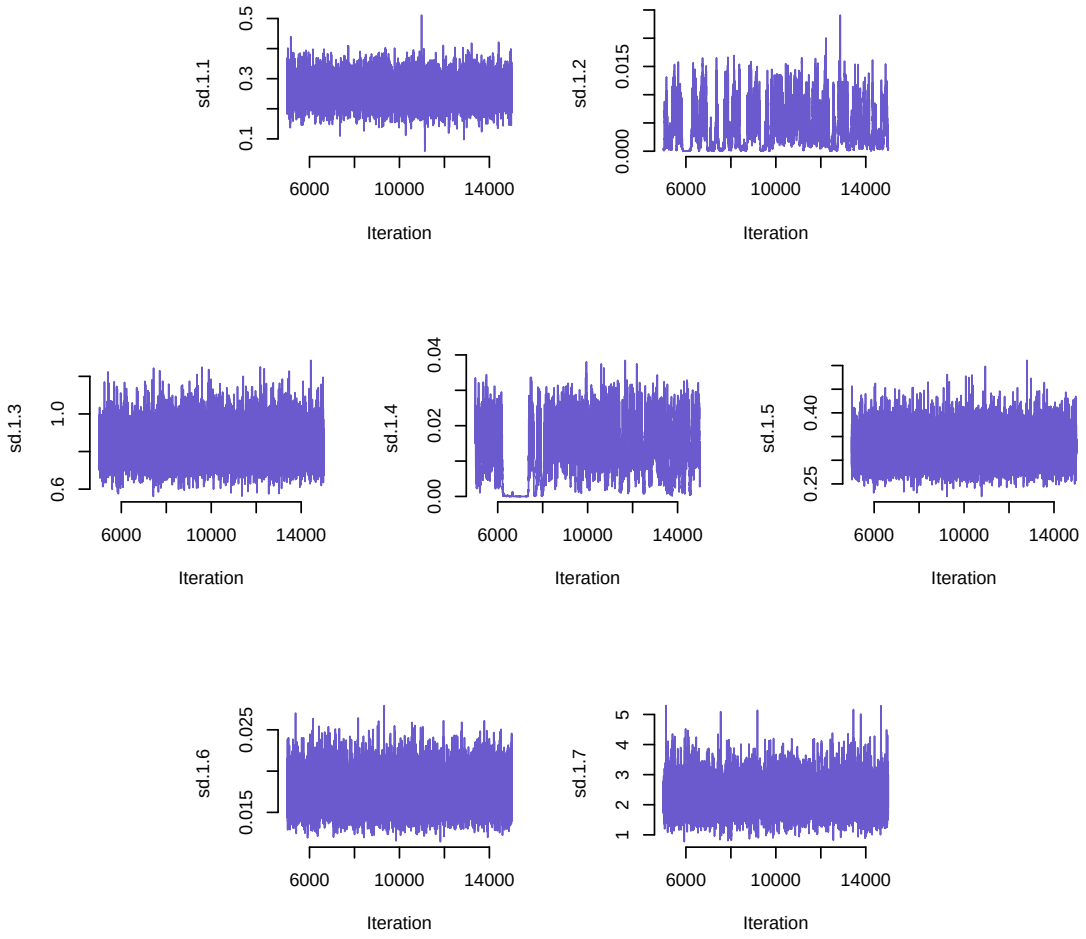

Figure 4.5: Trace plot of the MCMC chain of Group 1 of the model standard deviations derived from the covariance matrices $\mathbb{D}$.

ments. For example, 87% overall of patients correctly classified by the random effect prediction approach and gives the highest values of the AUC with 0.94. While the marginal and conditional approaches still are able to predict well in terms of the PCC with 83% and 70% of patients correctly identified respectively. This is not surprising since Komárek et al. (2010) shows in their paper that the random effects approach outperformed the marginal and the conditional approaches when applied to the Dutch Multicenter Primary Biliary Cirrhosis (PBC) data (similar variables were measured, but the cohorts were different) where the average number of visits per patient was 13 visits.

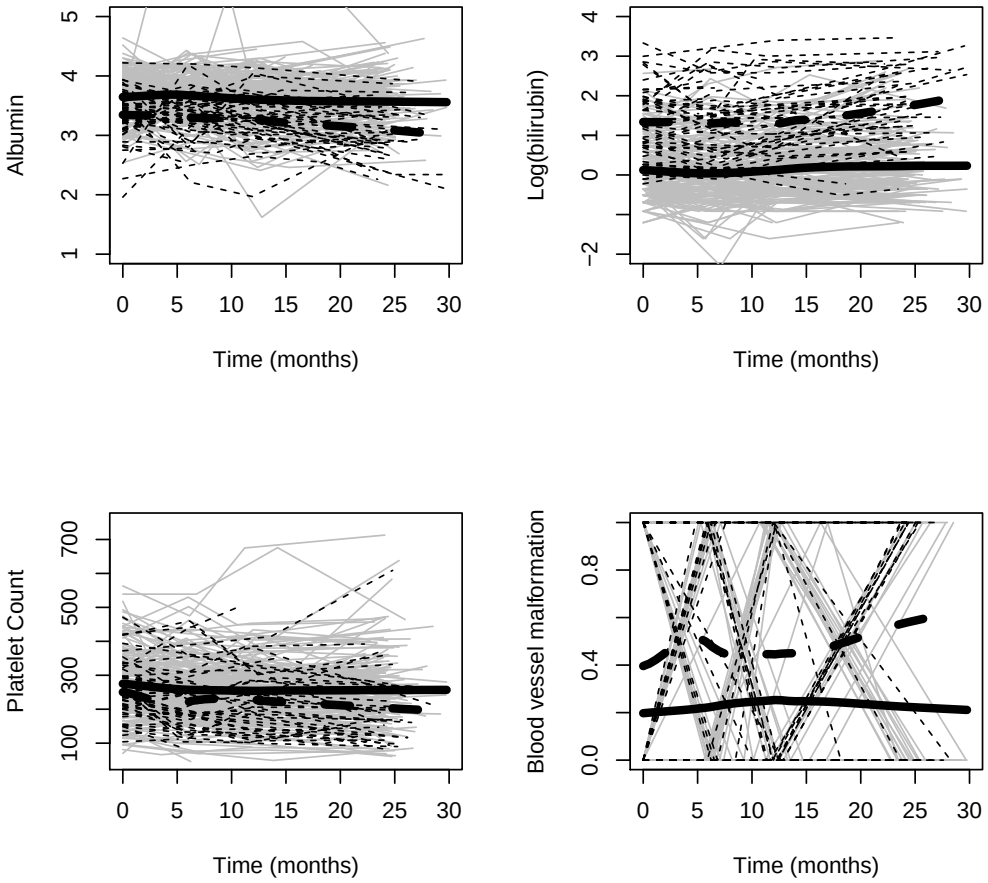

Figure 4.6: Longitudinal profiles of albumin (mg/dl), log(bilirubin) (log(mg/dl)), platelet counts and blood vessel malform (spiders) for patients who were known to be alive at 5 years (Group 0, solid lines) and who died between 2.5 and 5 years (Group 1, dashed lines). The thick lines show fitted mean of patients over time, calculated using loess.

The ROC curves of the three approaches are plotted in Figure 4.8 which shows the superiority of the random effects approach compared to the marginal and the conditional approaches.

4.3 Simulation Study

In recent years, there has been an increasing amount of literature on the use of patients' longitudinal data for prediction purposes. Three prediction approaches namely

Table 4.1: Prediction accuracy using leave-one-out cross validation for the random-effects, marginal and conditional approaches. PBC data collected during the first 2.5 years were used for the modelling and prediction. The average number of visits per patient was cohort 3.53, in this 2.5 years period.

	Random	Marginal	Conditional
Cutoff	0.98	0.21	0.12
Sensitivity	0.75	0.78	0.61
Specificity	0.78	0.81	0.67
PCC	0.77	0.81	0.66
AUC	0.81	0.85	0.63
PPV	0.46	0.51	0.32
NPV	0.92	0.94	0.94

Table 4.2: Prediction accuracy using leave-one-out cross validation based on the random-effects, marginal and conditional approaches. The average number of visits per patient was 7.03 visits in the full PBC dataset.

	Random	Marginal	Conditional
Cutoff	0.24	0.32	0.19
Sensitivity	0.88	0.78	0.75
Specificity	0.86	0.84	0.69
PCC	0.87	0.83	0.70
AUC	0.94	0.86	0.72
PPV	0.62	0.55	0.38
NPV	0.97	0.94	0.92

marginal, conditional and random-effects were first provided by Morrell et al. (2007) who compared these approaches based on some statistical measures. They focused on lead-time (time before patients were identified to the cancer group correctly) and the sensitivity (proportion of patients correctly identified to the cancer group) to select the best approach. With respect to these measures, the marginal approach gave the best results, while in the case of the overall correct classification (proportion of total patients that were correctly identified regardless of whether they were cancer or control cases) and the specificity (proportion of patients correctly identified to the healthy group), the random-effects approach showed the highest accuracy. A similar conclusion can be found in Morrell et al. (2011).

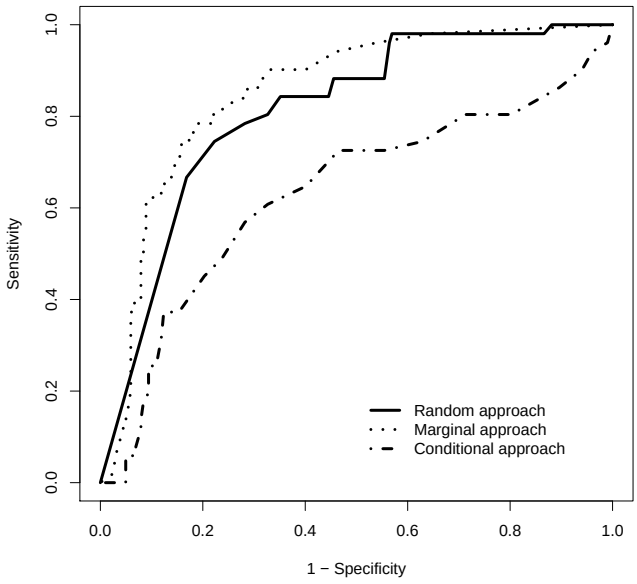

Figure 4.7: Receiver Operating Characteristic curves of the dynamic LoDA using the random effects (solid), marginal (dotted) and conditional (dot-dashed) prediction approaches. PBC data are collected during the first 2.5 years data are used for the modelling and prediction.

Komárek et al. (2010) also compared these three approaches using the PBC data and they showed the random-effects prediction was superior to other approaches. In addition, Komarek et al. (2009) pointed out that the random-effects prediction approach is the most promising approach to predict whether the patient will benefit from the treatment as early as possible when compared with the marginal and conditional approaches. A further investigation was conducted by Hughes et al. (2018b) who compared the three approaches to identify patients with epilepsy who will not achieve remission of seizures within five years, and they showed that the marginal and conditional approaches gave similar results, while the random-effects worked less well.

Since the PBC data analysis in Section 4.2 indicated that the predictions of the marginal and random effects provided the best classification accuracy, a simulation study was conducted to investigate further which prediction approaches give the best prediction accuracy based on different types of data. I created two simulated scenarios. For each scenario, data were simulated based on the PBC dataset (i.e., it includes

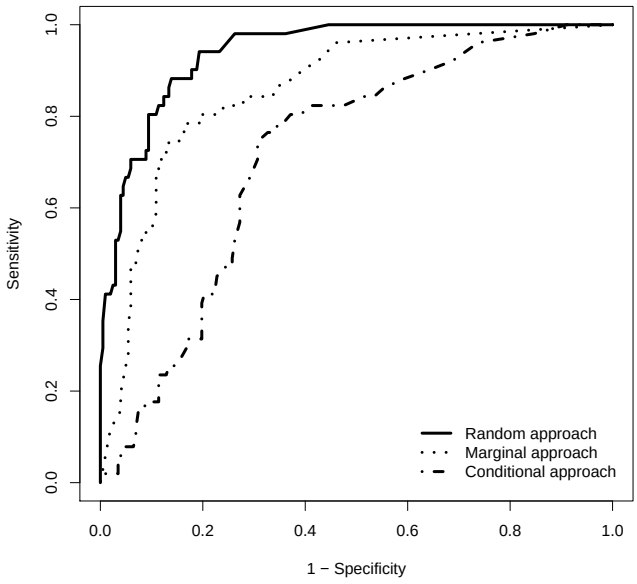

Figure 4.8: Receiver Operating Characteristic curves of the dynamic LoDA using the random effects (solid), marginal (dotted) and conditional (dot-dashed) prediction methods for the whole PBC data with average of 7.03 visits per patient.

data collected during the first 2.5 years). Each simulation consisted of 200 patients who survived after five years without requiring a liver transplant and 50 patients who did not survive or who needed the transplant at some time between 2.5 and 5 years. Following the simulation set up of Komárek et al. (2013), the observed process was sampled at four-time points: at baseline, then after approximately six months, one year and two years to mimic the patient's visits in the PBC dataset ($n_i = 4$). For each patient, the four visit times were generated as follows: the first visit time was set to 0 and uniform distributions in the intervals (170, 200), (350, 390) and (710, 770) days were used to generate the remaining visit times.

For the purposes of this simulation study, data were supposed to be balanced (i.e., no missing values, as I supposed all patients had the same number of longitudinal data). A total of 100 simulated datasets were generated using the MGLMMs described in Equations 4.1 where the number of mixture components was one (K = 1). The data were generated for each the four markers namely albumin, bilirubin, platelet count and blood vessel malform at each the four-time points. For each group, the MGLMM

Table 4.3: Parameter estimates for the PBC data and the modifications used for each simulation scenario. Blank entries occur when the parameter was not used in Scenario 2.

	Group 0			Group 1		
	PBC Data	Scenario 1	Scenario 2	PBC Data	Scenario 1	Scenario 2
Albumin						
E[Albumin:Intercept]	3.69	3.69	3.00	3.39	3.39	3.00
E[Albumin:slope]	-6.83×10^{-3}	-6.83×10^{-3}	0.00	-1.44×10^{-2}	-1.44×10^{-2}	0.00
SD[Albumin:Intercept]	2.73×10^{-1}	2.64×10^{-1}	6.50×10^{-2}	2.64×10^{-1}	2.64×10^{-1}	6.50×10^{-2}
Corr[Albumin:Intercept,Albumin:slope]	-8.60×10^{-2}	-6.46×10^{-2}	-6.46×10^{-2}	-6.46×10^{-2}	-6.46×10^{-2}	-6.46×10^{-2}
Corr[Albumin:Intercept,log(Bilirubin):Intercept]	-2.48×10^{-1}	-1.97×10^{-1}	-1.97×10^{-1}	-1.97×10^{-1}	-1.97×10^{-1}	-1.97×10^{-1}
Corr[Albumin:Intercept,log(Bilirubin):slope]	-1.10×10^{-1}	2.11×10^{-1}	2.11×10^{-1}	2.11×10^{-1}	2.11×10^{-1}	2.11×10^{-1}
Corr[Albumin:Intercept,Platelet:Intercept]	1.82×10^{-1}	1.91×10^{-1}		1.91×10^{-1}	1.91×10^{-1}	
Corr[Albumin:Intercept,Platelet:slope]	5.72×10^{-2}	1.09×10^{-1}		1.09×10^{-1}	1.09×10^{-1}	
Corr[Albumin:Intercept,Blood vessel malformation:Intercept]	-2.27×10^{-1}	-3.48×10^{-1}		-3.48×10^{-1}	-3.48×10^{-1}	
SD[Albumin:slope]	4.30×10^{-3}	7.76×10^{-3}	7.76×10^{-3}	7.76×10^{-3}	7.76×10^{-3}	7.76×10^{-3}
Corr[Albumin:slope,log(Bilirubin):Intercept]	-2.91×10^{-1}	1.57×10^{-3}	1.57×10^{-3}	1.57×10^{-3}	1.57×10^{-3}	1.57×10^{-3}
Corr[Albumin:slope,log(Bilirubin):slope]	-6.50×10^{-1}	-2.33×10^{-1}	-2.33×10^{-1}	-2.33×10^{-1}	-2.33×10^{-1}	-2.33×10^{-1}
Corr[Albumin:slope,Platelet:Intercept]	8.89×10^{-2}	-2.57×10^{-1}		-2.57×10^{-1}	-2.57×10^{-1}	
Corr[Albumin:slope,log(Bilirubin):slope]	2.96×10^{-1}	-2.60×10^{-1}		-2.60×10^{-1}	-2.60×10^{-1}	
Corr[Albumin:slope,Blood vessel malformation:Intercept]	-2.93×10^{-1}	2.27×10^{-1}		2.27×10^{-1}	2.27×10^{-1}	
SD[Albumin:residual]	3.18×10^{-1}	3.18×10^{-1}	3.18×10^{-1}	3.14×10^{-1}	3.14×10^{-1}	1.59×10^{-1}
log(Bilirubin)						
E[log(Bilirubin):Intercept]	2.13×10^{-2}	2.13×10^{-2}	1.00	1.23	1.23	1.00
E[log(Bilirubin):slope]	9.94×10^{-3}	9.94×10^{-3}	0.00	2.38×10^{-2}	2.38×10^{-2}	0.00
SD[log(Bilirubin):Intercept]	6.88×10^{-1}	8.45×10^{-1}	1.12×10^{-2}	8.45×10^{-1}	8.45×10^{-1}	1.12×10^{-2}
Corr[log(Bilirubin):Intercept,log(Bilirubin):slope]	2.32×10^{-1}	-1.75×10^{-1}	-1.75×10^{-1}	-1.75×10^{-1}	-1.75×10^{-1}	-1.75×10^{-1}
Corr[log(Bilirubin):Intercept,Platelet:Intercept]	-1.66×10^{-1}	2.47×10^{-1}		2.47×10^{-1}	2.47×10^{-1}	
Corr[log(Bilirubin):Intercept,Platelet:slope]	-2.04×10^{-1}	-1.87×10^{-1}		-1.87×10^{-1}	-1.87×10^{-1}	
Corr[log(Bilirubin):Intercept,Blood vessel malformation:Intercept]	3.42×10^{-1}	2.70×10^{-1}		2.70×10^{-1}	2.70×10^{-1}	
SD[log(Bilirubin):slope]	1.12×10^{-2}	1.49×10^{-2}	1.49×10^{-2}	1.49×10^{-2}	1.49×10^{-2}	1.49×10^{-2}
Corr[log(Bilirubin):slope,Platelet:Intercept]	1.44×10^{-2}	-1.69×10^{-1}		-1.69×10^{-1}	-1.69×10^{-1}	
Corr[log(Bilirubin):slope,Platelet:slope]	-2.40×10^{-1}	1.25×10^{-1}		1.25×10^{-1}	1.25×10^{-1}	
Corr[log(Bilirubin):slope,Blood vessel malformation:Intercept]	3.05×10^{-1}	8.13×10^{-3}		8.13×10^{-3}	8.13×10^{-3}	
SD[log(Bilirubin):residual]	3.38×10^{-1}	3.38×10^{-1}	3.38×10^{-1}	3.96×10^{-1}	3.96×10^{-1}	1.69×10^{-1}
Platelet Count						
E[Platelet:Intercept]	5.54	5.54		5.46	5.46	
E[Platelet:slope]	-4.29×10^{-3}	-4.29×10^{-3}		-1.14×10^{-2}	-1.14×10^{-2}	
SD[Platelet:Intercept]	3.73×10^{-1}	3.45×10^{-1}		3.45×10^{-1}	3.45×10^{-1}	
Corr[Platelet:Intercept,Platelet:slope]	-4.64×10^{-2}	6.14×10^{-2}		6.14×10^{-2}	6.14×10^{-2}	
Corr[Platelet:Intercept,Blood vessel malformation:Intercept]	-7.41×10^{-2}	-2.48×10^{-1}		-2.48×10^{-1}	-2.48×10^{-1}	
SD[Platelet:slope]	5.66×10^{-3}	1.51×10^{-2}		1.51×10^{-2}	1.51×10^{-2}	
Corr[Platelet:slope,Blood vessel malformation:Intercept]	-1.68×10^{-1}	-8.03×10^{-2}		-8.03×10^{-2}	-8.03×10^{-2}	
Blood Vessel Malformations						
E[Blood vessel malformation:Intercept]	-2.54	-2.54		-6.81×10^{-1}	-6.81×10^{-1}	
Blood vessel malformation:slope	1.46×10^{-2}	1.46×10^{-2}		4.81×10^{-2}	4.81×10^{-2}	
SD[Blood vessel malformation:Intercept]	3.00	1.88		1.88	1.88	

was based on 10,000 iterations of 1:10 thinned MCMC and burn-in of 500 iterations were considered. The predictions were assessed using the leave-one-out cross-validation approach. The sensitivity, specificity, PCC, PPV and NPV were measured for each simulated dataset using the optimal cutoff value. Also, the AUC was also measured for each simulation. These measurements were therefore used to compare the prediction approaches, based on the average of 100 simulated datasets. Table 4.3 presents the values of four markers that were used to simulate the two scenarios. My aim is to compare the three discriminant prediction approaches in these two different simulated scenarios.

4.3.1 Scenario 1

For this scenario, the random effects parameters and fixed effects parameters remained as the values of the PBC data, while the variance-covariance matrix of the random effects ($\mathbb{D}$) was considered to be the same in each group. In this situation, where the main differences between the two groups were in the mean profiles, I would expect that the marginal approach would give the best results. Figure 4.9 shows the observed longitudinal profiles of the markers for the scenario 1. The structure of the MGLMM for the scenario 1 is presented in Equation 4.1.

I simulated a dataset for scenario 1 consisting of four markers: albumin and bilirubin as continuous markers, platelet count as a discrete marker (Poisson) and blood vessel malformation as a binary marker. I assumed two groups for discrimination, 0 and 1. For Group 0, I considered 200 patients who survived after five years without requiring a liver transplant and 50 patients who did not survive or who needed the transplant at some time between 2.5 and 5 years. For each patient, the four visit times were generated as follows: the first visit time was set to 0 and uniform distributions in the intervals (170, 200), (350, 390) and (710, 770) days were used to generate the remaining visit times. The elements of mean vectors and variance-covariance matrices of the random effects considered for the four markers for each group are presented in the Table 4.3. At each time point I simulated values for each marker, by first generating random effects from a multivariate normal distribution with mean vector and covariance matrix given in Table 4.3 and then assuming a generalised linear mixed model with fixed effects parameters shown in Table 4.3. The R-code used for these simulations is given in Appendix B.

As can be seen in Table 4.4, the estimates for Scenario 1 (which was inspired by PBC data) are generally well-behaved based on the corresponding form bias and mean square error (MSE). The column labelled 'Posterior Mean' gives the average values for the parameters estimates over the 100 simulations with the 95% highest posterior density credible intervals (HPD). Table 4.4 reports poor coverage values of the random slope variances for the albumin and log(bilirubin) markers in both groups. These poor

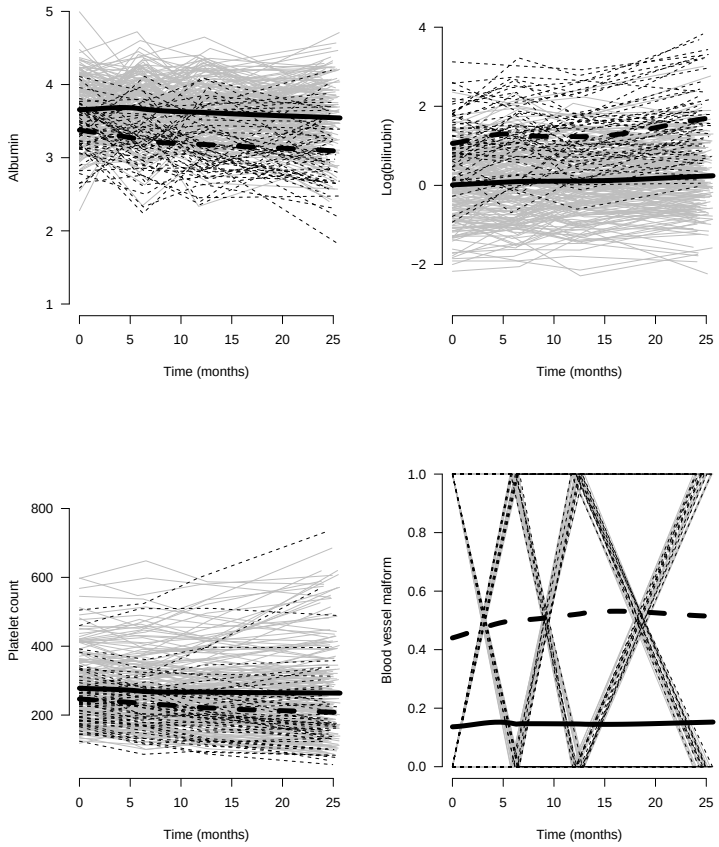

Figure 4.9: Simulation study Scenario 1: Longitudinal profiles of albumin, bilirubin, platelet count and blood vessel malform (spiders) for patients who were known to be alive at 5 years (Group 0, solid lines) and who died between 2.5 and 5 years (Group 1, dashed lines). The thick lines show fitted mean of patients over time, estimated using loess.

estimates for the two continuous markers might be due to the fact that the residual errors were larger than the true random slope variances (see Table 4.3) causing the incorrect estimate.

Table 4.4: Simulation study Scenario 1: Posterior Means, highly probable density (HPD) intervals, bias, standard deviation (SD), mean square error (MSE) and coverage for the fixed and random effects. These measurements were the average of 100 simulations.

	Group 0						Group 1					
	Posterior Mean	95% HPD Interval	SD	Bias	MSE	Coverage	Posterior Mean	95% HPD Interval	SD	Bias	MSE	Coverage
Albumin												
E[Albumin:Intercept]	3.69	(3.64,3.74)	4.03×10^{-3}	4.22×10^{-3}	6.35×10^{-4}	0.93	3.39	(3.29,3.49)	4.02×10^{-3}	-3.19×10^{-3}	2.08×10^{-3}	0.98
E[Albumin:slope]	-7.03×10^{-3}	$(-9.35,-4.68)\times10^{-3}$	3.47×10^{-4}	-2.02×10^{-4}	2.15×10^{-6}	0.88	-1.42×10^{-2}	$(-1.91,-0.93)\times10^{-2}$	2.76×10^{-4}	2.65×10^{-4}	6.85×10^{-6}	0.94
SD[Albumin:Intercept]	2.63×10^{-1}	$(2.22,3.06)\times10^{-1}$	1.81×10^{-3}	-6.12×10^{-4}	4.10×10^{-4}	0.98	2.58×10^{-1}	$(1.71,3.46)\times10^{-1}$	3.80×10^{-3}	-5.75×10^{-3}	2.77×10^{-3}	0.89
Corr[Albumin:Intercept,Albumin:slope]	2.09×10^{-2}	$(-5.25,5.89)\times10^{-1}$	2.45×10^{-2}	8.55×10^{-2}	2.28×10^{-2}	1.00	3.96×10^{-2}	$(-6.24,7.09)\times10^{-1}$	1.43×10^{-2}	1.04×10^{-1}	3.08×10^{-2}	0.99
Corr[Albumin:Intercept,log(Bilirubin):Intercept]	-1.75×10^{-1}	$(-3.49,0.02)\times10^{-1}$	6.79×10^{-3}	2.21×10^{-2}	6.34×10^{-3}	0.99	-1.61×10^{-1}	$(-4.97,1.81)\times10^{-1}$	1.14×10^{-2}	3.56×10^{-2}	2.87×10^{-2}	0.97
Corr[Albumin:Intercept,log(Bilirubin):slope]	1.73×10^{-1}	$(-1.67,5.06)\times10^{-1}$	1.35×10^{-2}	-3.87×10^{-2}	2.44×10^{-2}	0.96	5.84×10^{-2}	$(-5.94,6.94)\times10^{-1}$	1.38×10^{-2}	-1.53×10^{-1}	5.29×10^{-2}	0.98
Corr[Albumin:Intercept,Platelet:Intercept]	1.52×10^{-1}	$(-0.24,3.27)\times10^{-1}$	8.84×10^{-3}	-3.88×10^{-2}	9.83×10^{-3}	0.93	1.12×10^{-1}	$(-2.25,4.44)\times10^{-1}$	1.12×10^{-2}	-7.95×10^{-2}	3.13×10^{-2}	0.94
Corr[Albumin:Intercept,Platelet:slope]	6.59×10^{-2}	$(-1.14,2.46)\times10^{-1}$	8.38×10^{-3}	-4.35×10^{-2}	9.15×10^{-3}	0.93	4.21×10^{-2}	$(-3.00,3.86)\times10^{-1}$	1.17×10^{-2}	-6.73×10^{-2}	3.86×10^{-2}	0.91
Corr[Albumin:Intercept,Blood Vessel Malformations:Intercept]	-3.01×10^{-1}	$(-5.22,-0.76)\times10^{-1}$	9.16×10^{-3}	4.69×10^{-2}	1.36×10^{-2}	0.95	-2.58×10^{-1}	$(-6.39,1.37)\times10^{-1}$	1.31×10^{-2}	9.01×10^{-2}	4.24×10^{-2}	0.96
SD[Albumin:slope]	3.47×10^{-3}	$(0.65,7.13)\times10^{-3}$	6.27×10^{-4}	-4.30×10^{-3}	2.53×10^{-5}	0.45	3.98×10^{-3}	$(0.36,9.15)\times10^{-3}$	3.65×10^{-4}	-3.79×10^{-3}	2.46×10^{-5}	0.57
Corr[Albumin:slope,log(Bilirubin):Intercept]	-2.06×10^{-2}	$(-5.58,5.18)\times10^{-1}$	2.22×10^{-2}	-2.22×10^{-2}	1.29×10^{-2}	0.99	-1.31×10^{-2}	$(-6.54,6.34)\times10^{-1}$	1.23×10^{-2}	-1.47×10^{-2}	1.74×10^{-2}	0.99
Corr[Albumin:slope,log(Bilirubin):slope]	-2.65×10^{-2}	$(-6.18,5.74)\times10^{-1}$	2.55×10^{-2}	2.06×10^{-1}	6.09×10^{-2}	0.99	-1.03×10^{-2}	$(-7.62,7.46)\times10^{-1}$	1.14×10^{-2}	2.23×10^{-1}	5.82×10^{-2}	1.00
Corr[Albumin:slope,Platelet:Intercept]	-1.69×10^{-1}	$(-6.92,3.87)\times10^{-1}$	3.54×10^{-2}	8.78×10^{-2}	2.56×10^{-2}	1.00	-5.97×10^{-2}	$(-6.86,5.84)\times10^{-1}$	1.30×10^{-2}	1.97×10^{-1}	5.69×10^{-2}	0.99
Corr[Albumin:slope,Platelet:slope]	-1.54×10^{-1}	$(-6.54,3.73)\times10^{-1}$	3.06×10^{-2}	1.07×10^{-1}	3.29×10^{-2}	0.99	-8.70×10^{-2}	$(-6.94,5.46)\times10^{-1}$	1.39×10^{-2}	1.73×10^{-1}	5.08×10^{-2}	0.98
Corr[Albumin:slope,Blood Vessel Malformations:Intercept]	1.01×10^{-1}	$(-4.86,6.62)\times10^{-1}$	3.09×10^{-2}	-1.26×10^{-1}	3.14×10^{-2}	0.98	2.11×10^{-2}	$(-6.62,6.96)\times10^{-1}$	1.49×10^{-2}	-2.06×10^{-1}	6.47×10^{-2}	0.99
log(Bilirubin)												
E[log(Bilirubin):Intercept]	1.67×10^{-2}	$(-1.06,1.39)\times10^{-1}$	4.04×10^{-3}	-4.57×10^{-3}	3.71×10^{-3}	0.96	1.24	(0.99,1.49)	8.51×10^{-3}	9.60×10^{-3}	1.62×10^{-2}	0.96
E[log(Bilirubin):slope]	1.00×10^{-2}	$(0.69,1.32)\times10^{-2}$	1.27×10^{-4}	8.22×10^{-5}	3.41×10^{-6}	0.91	2.31×10^{-2}	$(1.65,2.99)\times10^{-2}$	3.44×10^{-4}	-6.91×10^{-4}	1.33×10^{-5}	0.94
SD[log(Bilirubin):Intercept]	8.42×10^{-1}	$(7.52,9.35)\times10^{-1}$	3.03×10^{-3}	-3.32×10^{-3}	2.37×10^{-3}	0.97	8.38×10^{-1}	(0.66,1.03)	6.08×10^{-3}	-7.28×10^{-3}	7.98×10^{-3}	0.96
Corr[log(Bilirubin):Intercept,log(Bilirubin):slope]	-1.33×10^{-1}	$(-4.25,1.78)\times10^{-1}$	1.11×10^{-2}	4.12×10^{-2}	2.26×10^{-2}	0.93	-4.34×10^{-2}	$(-6.45,5.95)\times10^{-1}$	1.31×10^{-2}	1.31×10^{-1}	4.00×10^{-2}	0.99
Corr[log(Bilirubin):Intercept,Platelet:Intercept]	2.35×10^{-1}	$(0.97,3.70)\times10^{-1}$	4.56×10^{-3}	-1.25×10^{-2}	4.23×10^{-3}	0.98	2.12×10^{-1}	$(-0.60,4.78)\times10^{-1}$	8.66×10^{-3}	-3.48×10^{-2}	1.67×10^{-2}	0.96
Corr[log(Bilirubin):Intercept,Platelet:slope]	-1.86×10^{-1}	$(-3.28,-0.42)\times10^{-1}$	4.66×10^{-3}	9.56×10^{-4}	5.38×10^{-3}	0.93	-1.49×10^{-1}	$(-4.27,1.34)\times10^{-1}$	9.22×10^{-3}	3.77×10^{-2}	2.06×10^{-2}	0.94
Corr[log(Bilirubin):Intercept,Blood Vessel Malformations:Intercept]	2.73×10^{-1}	$(0.82,4.60)\times10^{-1}$	6.34×10^{-3}	2.29×10^{-3}	8.31×10^{-3}	0.96	2.48×10^{-1}	$(-0.88,5.74)\times10^{-1}$	1.12×10^{-2}	-2.19×10^{-2}	2.70×10^{-2}	0.95
SD[log(Bilirubin):slope]	1.20×10^{-2}	$(0.64,1.72)\times10^{-2}$	4.52×10^{-4}	-2.93×10^{-3}	2.17×10^{-5}	0.76	7.03×10^{-3}	$(0.05,1.55)\times10^{-2}$	4.76×10^{-4}	-7.90×10^{-3}	8.64×10^{-5}	0.50
Corr[log(Bilirubin):slope,Platelet:Intercept]	-1.75×10^{-1}	$(-4.77,1.28)\times10^{-1}$	1.12×10^{-2}	-5.37×10^{-3}	1.87×10^{-2}	0.94	-6.26×10^{-2}	$(-6.68,5.62)\times10^{-1}$	1.30×10^{-2}	1.07×10^{-1}	3.67×10^{-2}	0.99
Corr[log(Bilirubin):slope,Platelet:slope]	1.10×10^{-1}	$(-1.88,4.07)\times10^{-1}$	1.01×10^{-2}	-1.48×10^{-2}	2.24×10^{-2}	0.94	6.57×10^{-2}	$(-5.44,6.61)\times10^{-1}$	1.26×10^{-2}	-5.91×10^{-2}	2.35×10^{-2}	1.00
Corr[log(Bilirubin):slope,Blood Vessel Malformations:Intercept]	1.92×10^{-2}	$(-3.55,3.97)\times10^{-1}$	1.35×10^{-2}	1.11×10^{-2}	3.06×10^{-2}	0.95	1.62×10^{-2}	$(-6.44,6.77)\times10^{-1}$	1.41×10^{-2}	8.09×10^{-3}	2.87×10^{-2}	0.99
Platelet Count												
E[Platelet:Intercept]	5.54	(5.49,5.59)	1.58×10^{-3}	2.20×10^{-3}	5.03×10^{-4}	0.96	5.46	(5.36,5.55)	3.12×10^{-3}	-4.86×10^{-3}	1.70×10^{-3}	0.98
E[Platelet:slope]	-4.29×10^{-3}	$(-6.45,-2.14)\times10^{-3}$	7.02×10^{-5}	4.70×10^{-6}	1.25×10^{-6}	0.94	-1.14×10^{-2}	$(-1.59,-0.70)\times10^{-2}$	1.42×10^{-4}	-3.02×10^{-5}	4.33×10^{-6}	0.95
SD[Platelet:Intercept]	3.49×10^{-1}	$(3.15,3.85)\times10^{-1}$	1.13×10^{-3}	4.27×10^{-3}	3.38×10^{-4}	0.96	3.49×10^{-1}	$(2.81,4.23)\times10^{-1}$	2.34×10^{-3}	4.34×10^{-3}	1.39×10^{-3}	0.94
Corr[Platelet:Intercept,Platelet:slope]	6.66×10^{-2}	$(-0.75,2.08)\times10^{-1}$	4.65×10^{-3}	5.20×10^{-3}	5.77×10^{-3}	0.93	6.77×10^{-2}	$(-2.11,3.45)\times10^{-1}$	9.28×10^{-3}	6.25×10^{-3}	1.59×10^{-2}	0.97
Corr[Platelet:Intercept,Blood vessel malformations:Intercept]	-2.57×10^{-1}	$(-4.40,-0.71)\times10^{-1}$	6.11×10^{-3}	-9.29×10^{-3}	9.27×10^{-3}	0.94	-2.23×10^{-1}	$(-5.42,1.05)\times10^{-1}$	1.04×10^{-2}	2.47×10^{-2}	2.60×10^{-2}	0.97
SD[Platelet:slope]	1.52×10^{-2}	$(1.37,1.69)\times10^{-2}$	5.20×10^{-5}	9.68×10^{-5}	6.37×10^{-7}	0.95	1.55×10^{-2}	$(1.23,1.90)\times10^{-2}$	1.09×10^{-4}	3.74×10^{-4}	2.92×10^{-6}	0.97
Corr[Platelet:slope,Blood vessel malformations:Intercept]	-7.64×10^{-2}	$(-2.72,1.20)\times10^{-1}$	6.45×10^{-3}	3.96×10^{-3}	9.90×10^{-3}	0.96	-7.61×10^{-2}	$(-4.16,2.68)\times10^{-1}$	1.13×10^{-2}	4.24×10^{-3}	2.39×10^{-2}	0.97
Blood Vessel Malformations												
E[Blood vessel malformations:Intercept]	-2.55	(-3.12,-2.00)	1.93×10^{-2}	-7.35×10^{-3}	1.19×10^{-1}	0.89	-6.74×10^{-1}	(-1.48,0.11)	2.57×10^{-2}	6.45×10^{-3}	1.82×10^{-1}	0.92
Blood vessel malformations:Slope	1.37×10^{-2}	$(-1.03,3.78)\times10^{-2}$	7.98×10^{-4}	-8.53×10^{-4}	1.84×10^{-4}	0.88	4.63×10^{-2}	$(0.67,8.66)\times10^{-2}$	1.35×10^{-3}	-1.77×10^{-3}	3.96×10^{-4}	0.95
SD[Blood vessel malformations:Intercept]	1.89	(1.39,2.40)	1.84×10^{-2}	4.76×10^{-3}	1.02×10^{-1}	0.89	1.82	(1.02,2.69)	3.42×10^{-2}	-5.73×10^{-2}	2.71×10^{-1}	0.91

Among the three approaches, the marginal approach produces the highest values of specificity, PCC, PPV and AUC, while the random effects approach provides the best sensitivity and NPV (Table 4.5). The random-effects prediction is able to correctly predict 93% of patients who will die or need a liver transplant (sensitivity). In addition, 98% of the patients who were predicted by the random-effects model as patients who will survive without a liver transplant did truly survive without a liver transplant. The positive predicted value showed that 57% of patients predicted to die or need transplant did truly die or needed a liver transplant, and the specificity showed that 71% of patients who will survive without a transplant were correctly classified. Regarding the overall accuracy, predictions of the random-effects and conditional approaches identified 25% and 27% of patients incorrectly, respectively. The marginal approach provided the best prediction accuracy, where 85% of patients who will not survive or need transplant were correctly classified, and 85% of patients who will survive without a liver transplant were correctly classified. The overall correct classification of the patients was 85%. Figure 4.10 represents receiver operating characteristic curves (ROC) for the marginal, random effects and conditional approaches, and again shows that the marginal approach gives the best prediction accuracy.

Table 4.5: Prediction accuracy for the simulated data from Scenario 1 based on leave-one-out cross validation for the random-effects, marginal and conditional approaches.

	Random	Marginal	Conditional
Cutoff	0.81	0.19	0.12
Sensitivity	0.93	0.85	0.70
Specificity	0.71	0.85	0.73
PCC	0.75	0.85	0.73
AUC	0.84	0.91	0.76
PPV	0.57	0.59	0.40
NPV	0.98	0.96	0.96

Figure 4.11 presents the histograms based on 100 simulated data of the sensitivity, specificity, PCC and AUC of each the random effects, marginal and conditional approaches. The wide variation in the accuracy measures for the random effects approach among each simulated dataset can be observed, where the values ranged from 0

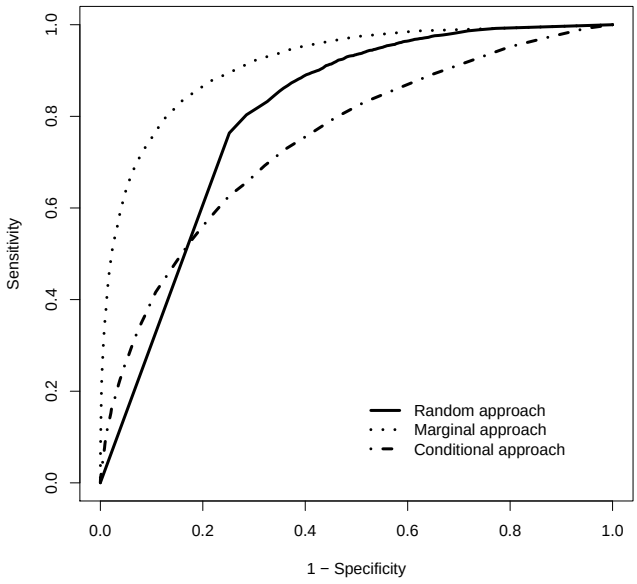

Figure 4.10: Receiver Operating Characteristic curves of the dynamic LoDA using the random effects (solid), marginal (dotted) and conditional (dot-dashed) prediction methods for Scenario 1.

to 1 using the optimal cut-off while the measurements of the marginal and conditional approaches are steady.

To summarise Scenario 1, the results indicate that the marginal approach performs well for prediction when the main differences between the groups are mainly due to differences between the mean profiles. Interestingly, this current scenario gave similar results to some references that also found that the best approach for prediction was the marginal approach. For example, Morrell et al. (2011) indicated that the marginal approach accurately identified patients who would develop prostate cancer almost ten years before they were clinically observed. Moreover, Table 5 and Figure 1 of Hughes et al. (2018b) showed that the marginal approach gave a good prediction for patients who would not achieve remission of seizures compared to the conditional and random effects approaches.

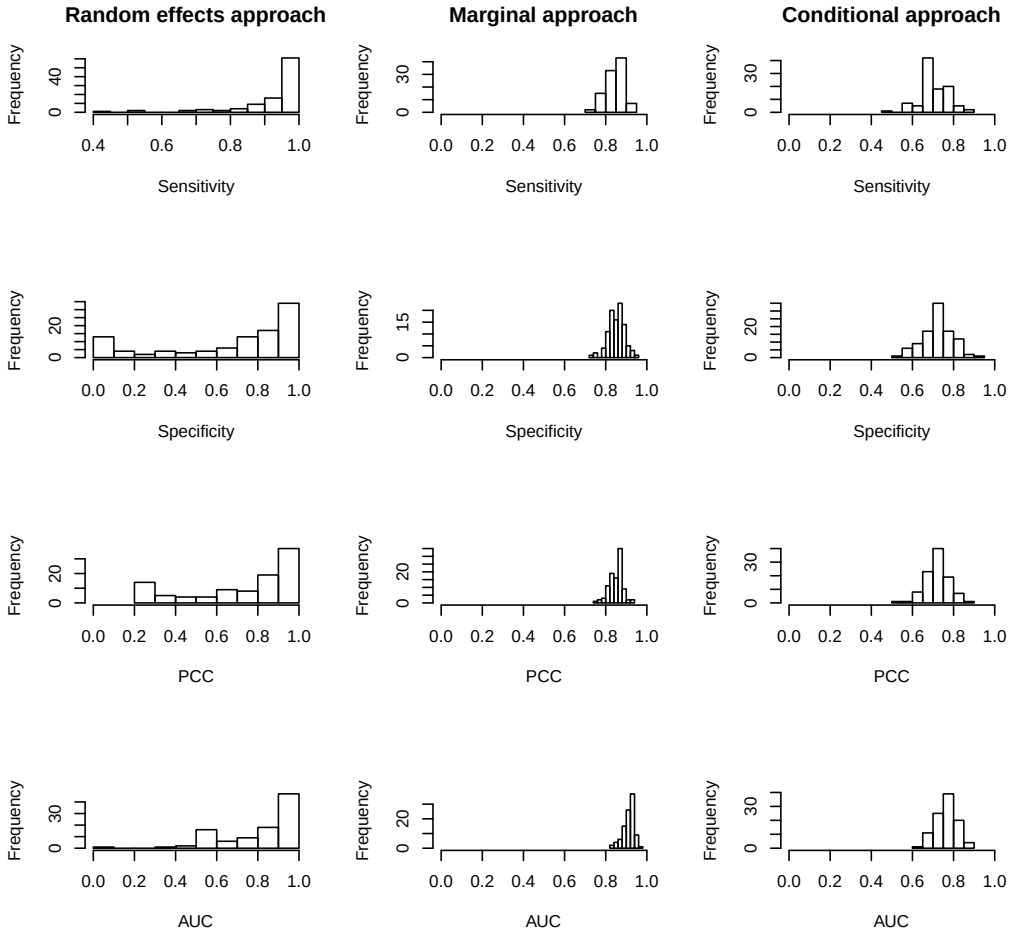

Figure 4.11: Histograms showing the sensitivity, specificity, PCC and AUC of each of the three approaches for each of the 100 simulated datasets under Scenario 1.

4.3.2 Scenario 2

The various circumstances where the marginal or random effects approaches work well are recognised in the literature, but in this scenario, I tried to explore a situation where the conditional approach would outperform the other approaches. However, finding a situation where the conditional approach gives the best prediction was not easy. All previous works on this comparison (included Morrell et al. (2007), Komárek et al. (2010), Morrell et al. (2011) and Hughes et al. (2018b)) showed that the marginal and random effects approaches provide the best prediction. To my knowledge, there is no study showing an example where the conditional approach outperformed the other approaches.

The idea behind this scenario comes from Morrell et al. (2011) who pointed out that if the variance of the residual error is large compared to the random effects variance, the conditional approach might outperform the marginal and random effects approaches. Furthermore, Komárek et al. (2010) mentioned that the marginal and conditional approaches take into account the residual error to estimate the posterior probabilities. Therefore in this scenario, only the continuous markers (albumin and bilirubin) were considered since only continuous markers have a residual error.

I simulated a dataset for Scenario 2 consisting of two continuous markers: albumin and bilirubin. I assumed two groups for discrimination, 0 and 1. For Group 0, I considered 200 patients who survived after five years without requiring a liver transplant and 50 patients who did not survive or who needed the transplant at some time between 2.5 and 5 years. For each patient, the four visit times were generated as follows: the first visit time was set to 0 and uniform distributions in the intervals (170, 200), (350, 390) and (710, 770) days were used to generate the remaining visit times. In describing the simulation Scenario 2, the means and variances of the random effects were kept to be the same in each group (Table 4.3). At each time point I simulated values for each marker, by first generating random effects from a multivariate normal distribution with mean vector and covariance matrix given in Table 4.3. While the residual errors differed between the two groups. The residual errors measurements for the two markers: albumin and bilirubin in Group 0 were 0.318, 0.338, while for Group 1 were 0.159, 0.169, respectively. The following Equation 4.2 presents the structure of the MGLMM for the Scenario 2.

$$\begin{aligned} Y^g_{i,1,j} &= b^g_{i,1,1} + b^g_{i,1,2}\, t^g_{i,1,j} + \epsilon^g_{i,1,j}, \qquad \text{albumin} \\ Y^g_{i,2,j} &= b^g_{i,2,1} + b^g_{i,2,2}\, t^g_{i,2,j} + \epsilon^g_{i,2,j}, \qquad \text{log(bilirubin)} \end{aligned} \tag{4.2}$$

The MGLMM for this scenario included two longitudinal markers $(Y_{i,1,j}, Y_{i,2,j})$ which were assumed to be independent given the random effects and to follow the Gaussian distribution. The random effects followed a 4-dimensional normal distribution where each marker had a random intercept and slope $(b_{i,1,1}, b_{i,1,2}, b_{i,2,1}, b_{i,2,2})$. The MGLMM also involved two residual errors $(\epsilon_{i,1,j}, \epsilon_{i,2,j})$. For each MCMC, the results were based

on 10,000 iterations of 1:10 thinned MCMC after a burn-in period of 500 iterations.

Table 4.6: Simulation study Scenario 2: Posterior Means, highly probable density (HPD) intervals, bias, standard deviation (SD), mean square error (MSE) and coverage for the fixed and random effects. These measurements were the average of 100 simulations.

	Group 0						**Group 1**					
	Posterior Mean	95% HPD Interval	SD	Bias	MSE	Coverage	Posterior Mean	95% HPD Interval	SD	Bias	MSE	Coverage
Albumin												
E[Albumin:Intercept]	3.00	(2.97,3.03)	2.80×10^{-3}	-2.04×10^{-3}	3.73×10^{-4}	0.89	3.00	(2.96,3.04)	1.89×10^{-3}	3.45×10^{-3}	4.00×10^{-4}	0.92
E[Albumin:slope]	-9.63×10^{-5}	$(-2.37,2.10)\times10^{-3}$	2.12×10^{-4}	-9.63×10^{-5}	1.49×10^{-6}	0.92	-3.08×10^{-4}	$(-3.01,2.38)\times10^{-3}$	1.31×10^{-4}	-3.08×10^{-4}	1.88×10^{-6}	0.92
SD[Albumin:Intercept]	4.19×10^{-2}	$(0.34,9.07)\times10^{-2}$	4.21×10^{-3}	-2.31×10^{-2}	1.06×10^{-3}	0.84	4.28×10^{-2}	$(0.68,8.68)\times10^{-2}$	2.10×10^{-3}	-2.22×10^{-2}	1.19×10^{-3}	0.76
Corr[Albumin:Intercept,Albumin:slope]	-7.48×10^{-3}	$(-8.53,8.51)\times10^{-1}$	2.02×10^{-2}	5.72×10^{-2}	1.26×10^{-2}	1.00	2.61×10^{-2}	$(-8.02,8.50)\times10^{-1}$	1.18×10^{-2}	9.07×10^{-2}	2.60×10^{-2}	1.00
Corr[Albumin:Intercept,log(Bilirubin):Intercept]	1.50×10^{-2}	$(-8.38,8.64)\times10^{-1}$	2.32×10^{-2}	2.12×10^{-1}	5.79×10^{-2}	1.00	3.21×10^{-3}	$(-8.46,8.64)\times10^{-1}$	1.10×10^{-2}	2.00×10^{-1}	5.28×10^{-2}	1.00
Corr[Albumin:Intercept,log(Bilirubin):slope]	3.21×10^{-2}	$(-7.43,7.95)\times10^{-1}$	2.46×10^{-2}	-1.79×10^{-1}	7.77×10^{-2}	0.97	1.01×10^{-1}	$(-6.29,7.96)\times10^{-1}$	1.57×10^{-2}	-1.10×10^{-1}	5.22×10^{-2}	0.98
SD[Albumin:slope]	2.13×10^{-3}	$(0.07,5.38)\times10^{-3}$	2.54×10^{-4}	-5.63×10^{-3}	3.38×10^{-5}	0.16	2.91×10^{-3}	$(0.56,5.79)\times10^{-3}$	1.32×10^{-4}	-4.85×10^{-3}	2.75×10^{-5}	0.24
Corr[Albumin:slope,log(Bilirubin):Intercept]	1.18×10^{-2}	$(-8.59,8.73)\times10^{-1}$	2.35×10^{-2}	1.02×10^{-2}	8.28×10^{-3}	1.00	-2.69×10^{-2}	$(-8.80,8.37)\times10^{-1}$	1.06×10^{-2}	-2.84×10^{-2}	8.65×10^{-3}	1.00
Corr[Albumin:slope,log(Bilirubin):slope]	-4.95×10^{-2}	$(-8.46,7.84)\times10^{-1}$	2.98×10^{-2}	1.83×10^{-1}	6.26×10^{-2}	0.99	-8.38×10^{-2}	$(-7.91,6.32)\times10^{-1}$	1.50×10^{-2}	1.49×10^{-1}	6.69×10^{-2}	0.98
log(**Bilirubin**)												
E[log(Bilirubin):Intercept]	1.00	(0.96,1.04)	3.05×10^{-3}	-3.92×10^{-4}	3.58×10^{-4}	0.96	1.00	(0.96,1.04)	1.74×10^{-3}	7.55×10^{-4}	3.71×10^{-4}	0.92
E[log(Bilirubin):slope]	8.35×10^{-5}	$(-3.20,3.37)\times10^{-3}$	1.95×10^{-4}	8.35×10^{-5}	2.18×10^{-6}	0.98	3.08×10^{-4}	$(-3.69,4.28)\times10^{-3}$	1.50×10^{-4}	3.08×10^{-4}	4.18×10^{-6}	0.93
SD[log(Bilirubin):Intercept]	3.35×10^{-2}	$(0.03,9.02)\times10^{-2}$	3.39×10^{-3}	2.23×10^{-2}	1.11×10^{-3}	0.99	1.79×10^{-2}	$(0.00,5.29)\times10^{-2}$	1.36×10^{-3}	6.65×10^{-3}	2.21×10^{-4}	0.98
Corr[log(Bilirubin):Intercept,log(Bilirubin):slope]	-1.11×10^{-2}	$(-8.24,8.29)\times10^{-1}$	2.61×10^{-2}	1.64×10^{-1}	5.96×10^{-2}	1.00	-1.66×10^{-2}	$(-8.36,8.30)\times10^{-1}$	1.48×10^{-2}	1.58×10^{-1}	4.98×10^{-2}	0.99
SD[log(Bilirubin):slope]	1.02×10^{-2}	$(0.59,1.43)\times10^{-2}$	2.67×10^{-4}	-4.74×10^{-3}	2.95×10^{-5}	0.38	1.10×10^{-2}	$(0.79,1.43)\times10^{-2}$	1.06×10^{-4}	-3.93×10^{-3}	1.82×10^{-5}	0.39

Results for the estimation of the random effects parameters are given in Table 4.6. In general, Scenario 2 produced good estimates with small bias and small MSE and low standard deviation. With regard to the coverages of the true parameters, Table 4.6 shown that the random slope variance for albumin and bilirubin are poorly estimated for both groups. This result may be explained by the fact that the residual errors are large compared to the true values of the random slope variances (Table 4.3) which could lead to inaccurate estimates of these slopes.

Table 4.7: Scenario 2 prediction accuracy based on leave-one-out cross validation for the random-effects, marginal and conditional approaches.

	Random	Marginal	Conditional
Cutoff	0.56	0.26	0.58
Sensitivity	0.78	0.92	0.92
Specificity	0.70	0.89	0.88
PCC	0.72	0.90	0.89
AUC	0.74	0.96	0.95
PPV	0.55	0.69	0.66
NPV	0.90	0.98	0.93

With respect to prediction, the conditional and marginal approaches outperform the random effects approach (Table 4.7). Only 11% and 10% of patients are classified wrongly using the conditional approach and marginal approach, respectively. Figure 4.12 displays ROC curves of the three approaches under scenario 2. The sensitivity and specificity that are used to plot the ROC curves are averaged across the 100 simulated datasets at each cut-off. The AUCs values (Table 4.7) are 0.96 and 0.95 for the marginal and conditional predictions, respectively. The random effects approach works less well than other approaches. I did not expect the marginal approach to be as good as the conditional approach. I expected that the conditional approach would outperform the other two approaches.

To explain this in more detail, Komárek et al. (2010) stated that in the case of continuous longitudinal markers, the residual error is taken into account when calculating group membership probabilities for both conditional and marginal approaches. For the conditional approach, the mean and the variance of the multivariate normal distribution are influenced by the residual variance, and the marginal approach only

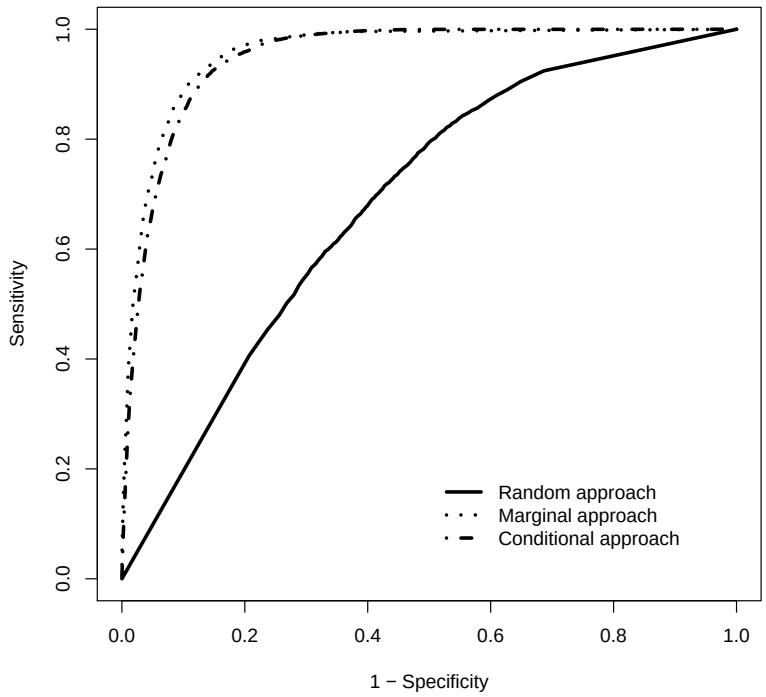

Figure 4.12: Receiver Operating Characteristic curves of the dynamic LoDA using the random effects (solid), marginal (dotted) and conditional (dot-dashed) prediction methods for Scenario 2.

the variance is affected while the normal distribution for the random effects approach does not handle the residual variance and relies on the individual random effects estimates. The random effects approach was unable to give accurate classification in this scenario since the estimation approach doesn't use the residual error variance, which both the marginal and conditional approaches use and so are able to provide more accurate classification.

Figure 4.13 shows the histograms based on 100 simulated data of the sensitivity, specificity, PCC and AUC of each the random effects, marginal and conditional approaches. The large disparity in the accuracy measures for the random effects approach among each simulated dataset can be observed, where the values ranged from 0 to 1 using the optimal cut-off while the measurements of the marginal and conditional approaches are stable (the histograms look symmetric).

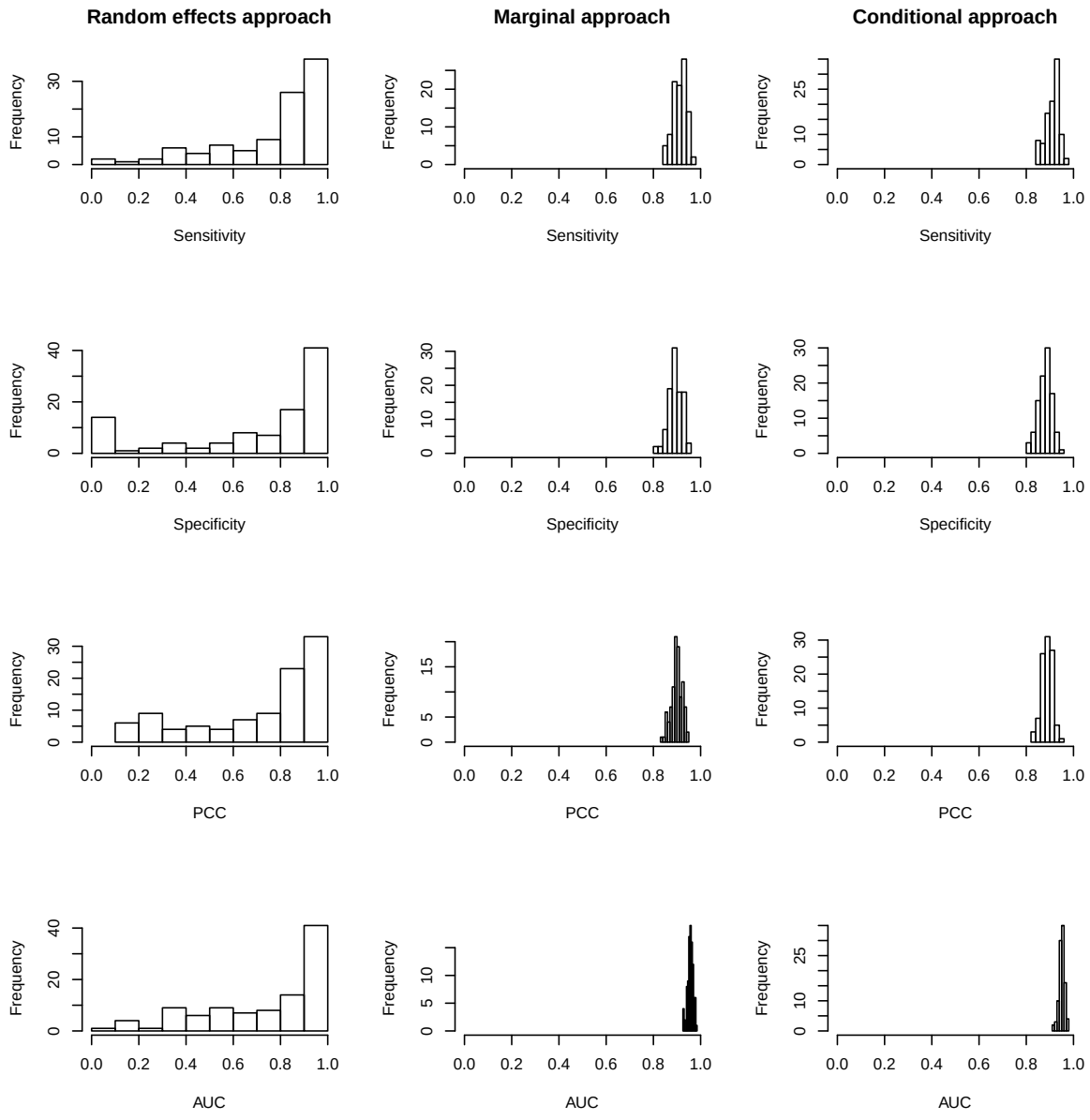

Figure 4.13: Histograms showing the sensitivity, specificity, PCC and AUC of each of the three approaches for each of the 100 simulated datasets under Scenario 2.

4.4 Discussion

In this chapter, analysis of the PBC data and simulation studies were conducted to explore in which situation each of the prediction approaches (namely: marginal, conditional and random effects) work well. Some comparisons between the three prediction approaches were published by Morrell et al. (2007, 2011), Komárek et al. (2010) and

Hughes et al. (2018b). These publications which used a real datasets concluded that the marginal and random effects approaches offer the best prediction accuracy. Here, I further investigated and compared these three approaches by using real and simulated data.

According to the results of the study in this chapter, the marginal approach provides the best prediction results in the case where the main differences between the prognostic groups are due to differences of the mean longitudinal profiles. However, if the number of observations increases per patients, the random effects approach is able to classify patients well.

If the variability about the mean profiles across the groups is noticeable, the conditional and the marginal approaches were expected to be less accurate than the random effect approach.

To identify a simulated scenario where the conditional approach outperforms the other approaches was not easy. I was unable to come up with a scenario where the conditional approach worked the best. My work shows that even in situations where the conditional approach was expected to do well, the marginal approach was just as good. To my knowledge, there is no publication which shows the conditional approach works better than the marginal or random effect approach.

In conclusion, this work conducted in this chapter suggests that before analysing a dataset, the longitudinal profiles of the data for each group should be plotted first. If the difference between the group mean profiles is clear then the marginal approach is expected to give the best prediction. If a variation in the level of variability across the group mean in each group is substantial, then the marginal and conditional approaches are not expected to give the best prediction but, the random effects approach is assumed to work well. However, the random effects prediction approach should be avoided if a substantial measurement error dominates the variability between patients leads to the random effects approach is unable to estimate the individual random effects correctly.

Chapter 5

Impact of misspecified random effects distribution

In the previous chapter, three prediction approaches for LoDA have been compared. These three approaches are based on a mixed model using a patient's longitudinal history to classify new patients according their future status. Mixed Models use random effects to model the correlation between repeated measurements on the same subject and assume a joint distribution of these random effects across all patients. The typical assumption is that the random effects follow a (potentially multivariate) normal distribution. In this chapter, I investigate whether the misspecification of the random effects distribution affects the classification performance. This work is currently under review as El Saeiti et al. (2019).

5.1 Introduction

Statistical models that use random effects terms in their modelling to analyse longitudinal data have become more common recently. A patient-specific random effect can take the correlation between measurements on the same patient into account. It is commonly assumed that random effects follow a normal distribution. However, Verbeke

and Lesaffre (1996) reported that checking the normality assumption for the random effects is not easy and misspecification can seriously affect their estimates.

In recent years, researchers have reported that misspecification of the distribution of the random effects can have an impact on the estimation of the parameters. Generalised linear mixed effects models are popularly adopted for the analysis of longitudinal data, when the responses are not assumed to be Gaussian. It is not yet clear whether the maximum likelihood estimates are robust to misspecification of the random effects distribution. Neuhaus et al. (1992) showed that estimation of the random effects intercept of mixed effects logistic model has a small bias if the random effects distribution is misspecified. Heagerty and Kurland (2001) investigated the impact of the misspecification of the random effects distribution on maximum likelihood estimates of the generalised linear mixed model. They showed that if the distribution of the random effects is far from the Gaussian distribution a substantial bias can occur. However, McCulloch and Neuhaus (2011*a*) pointed out that maximum likelihood estimates for the generalised linear mixed model are often robust to misspecification to the random effects distribution.

Agresti et al. (2004) carried out several investigations into the effect of the misspecification of a random effects distribution and gave possible solutions. They found that assuming a normal distribution affected model performance when the actual distribution is a two component normal mixture with high variance. Also, Litière et al. (2008) addressed the impact of a misspecified random effects distribution on generalised linear mixed models. They showed that if the variances of the random effects are large, the biases are significantly large, while the fixed effects have a small bias. They also pointed out that the estimates of the variance components are always heavily biased, even for small deviations of the distribution of random-effects.

Litière et al. (2007) discussed the impact of the misspecification of the random effects distribution on type I and type II errors in generalised linear mixed models. They found that the misspecification of the shape of the random effects distribution can severely affect the type I error and the power of a statistical test. They simulated a

number of scenarios where data were generated using four different 'true' random-effects distributions, and in each scenario, a normal distribution is assumed. Commenting on this paper, Neuhaus et al. (2011) argued that instead of fixing the distribution, four different distributions should be explored to identify the most appropriate and whether the normal distribution increases the bias. They showed that the type 2 error increases slightly when carrying out their simulation. Litière et al. (2011) argued that both types of simulations are useful to address the impact of misspecification of the random effects distribution.

With regard to prediction, McCulloch and Neuhaus (2011*b*) showed that the accuracy of prediction (as measured by mean square error) could be influenced slightly by mild to moderate violations of the model assumptions. Verbeke and Lesaffre (1997) concluded that misspecification of the random effects distribution has no effects on parameters that are estimated from the maximum likelihood method.

As I have summarised above, the impact of misspecification of the random effects distribution has been widely investigated. However, most of the researchers have focused on the accuracy of the parameters of the model. In this work, I explore whether the misspecification of the random effects distribution has an impact on the classification accuracy of a classifier. Parameters of the GLMM may be estimated with a small/large bias, and here I assess the effects of this bias on the ability to classify patients into their future status using LoDA approaches. In addition, I explore whether the sample size and number of repeated measurements can affect the classification performance. I have set four different distributions for random effects: a single normal distribution, a mixture of two Gaussian distribution, a mixture of three Gaussian distribution and T distribution for this investigation.

The purpose of longitudinal discriminant approaches is to use patient longitudinal history to predict which group a patient belongs to. There are three ways to do this, namely: marginal, conditional and random effects approaches, and each approach attempts to use the model information differently. This chapter focuses on marginal and random effects approaches which have shown good levels of accuracy (see Chapter

4 and also Hughes et al. (2018a), Komárek et al. (2010), Komarek et al. (2009) and Morrell et al. (2011)).

All previous studies investigating the effects of misspecification of the random effects distribution have only been carried out using a single longitudinal marker. In clinical research, multiple longitudinal variables may be collected, and these may be used in the discriminant analysis to provide better classification accuracy than variables collected at one unique time point. The approach for the discriminant analysis presented here allows for multiple markers, which has an effect on the model structure since increasing number of markers leads to an increasing number of random effects parameters.

The chapter has been organised in the following way. Section 5.2 shows an analysis of the PBC clinical data to investigate the effect that the selection of the random-effects distribution has on the level of accuracy. Section 5.3 contains the results of a simulation study investigating the effects of misspecification of random effects distributions. The chapter concludes with a brief summary.

5.2 PBC application

Data from the Mayo Clinic Primary Biliary Cirrhosis (PBC) dataset (Dickson et al. (1989)) have been used in the previous chapter and they are used for this study as well. The data contains a large number of longitudinal variables collected from 312 patients over a median of 6.3 years per patient. These data are used here to give an example of how a variety of random effects distributions might influence the classification accuracy.

The study consisted of 253 patients who were followed for at least 30 months (2.5 years), and whose 60 months (five years) status was known. The purpose of this work is to predict patients who will die or require a liver transplant within five years. A total of 202 of 253 patients were classified as known to be alive without a transplant after five years (referred as Group 0), while 51 patients died or needed a liver transplant at some time between 2.5 and 5 years after (referred as Group 1). There are three longitudinal

markers which had been used in this study namely, logarithmic serum bilirubin as a continuous marker, platelet count as a discrete marker (Poisson) and a binary marker indicating the presence of blood vessel malformations. Figure 5.1 shows the observed longitudinal profiles of all three markers for patients who were known to be alive at five years (Group 0) and who died after 2.5 years (Group 1).

The LoDA procedure is similar to the one described in Chapter 4. For each patient group, I fitted a multivariate generalised linear mixed model (MGLMM) to the longitudinal data for the marginal approach. For the continuous and count markers (log(bilirubin) and platelet counts), the GLMM included a random intercept and a random slope. The binary marker (blood vessel malformation) was modelled using a random intercept and a fixed effect term for time. Mixtures of multivariate normal distributions were assumed for the random effects distribution. The general structure of the MGLMM as follows (notation described in Section 2.3.2):

$$\begin{aligned}
&E(Y_{i,1,j}|\mathbf{b}_{i,1})^g = b^g_{i,1,1} + b^g_{i,1,2}\, t^g_{i,1,j}, && \text{log(bilirubin)} \\
&\log\{E(Y_{i,2,j}|\mathbf{b}_{i,2})^g\} = b^g_{i,2,1} + b^g_{i,2,2}\, t^g_{i,2,j}, && \text{platelet count} \\
&\text{logit}\{E(Y_{i,3,j}|b_{i,3},\alpha_3)^g\} = b^g_{i,3} + \alpha^g_3\, t^g_{i,3,j} && \text{blood vessel malformations}
\end{aligned} \tag{5.1}$$

where $i = 1, \ldots, N^g$, N^g indicates the number of patients in group g, where g can take the value either 0 or 1 ($g = 0, 1$), $j = 1, \cdots, n_{i,r}$, r indicates the number of markers $r = 1, 2, 3$, $t_{i,r,j}$ is the follow-up time for each marker (which is reported in months). The MGLMM contains a five dimensional vector of random effects $(b_{i,1,1}, b_{i,1,2}, b_{i,2,1}, b_{i,2,2}, b_{i,3})$ (see Equation 5.1), where $(b_{i,1,1}, b_{i,2,1}, b_{i,3})$ are random intercepts for each marker and $(b_{i,1,2}, b_{i,2,2})$ are random slopes for the first two markers. The vector of random effects is assumed to follow a joint distribution and in this chapter I consider four possible options for that distribution (specifically a single multivariate normal distribution, and then 2, 3 and 4 component mixtures). Blood vessel malformations ($Y_{i,3,j}$) has a fixed effects slope (α_3) only. The MGLMM includes one continuous marker (log(bilirubin)) which is assumed to follow the Gaussian distribution and has a

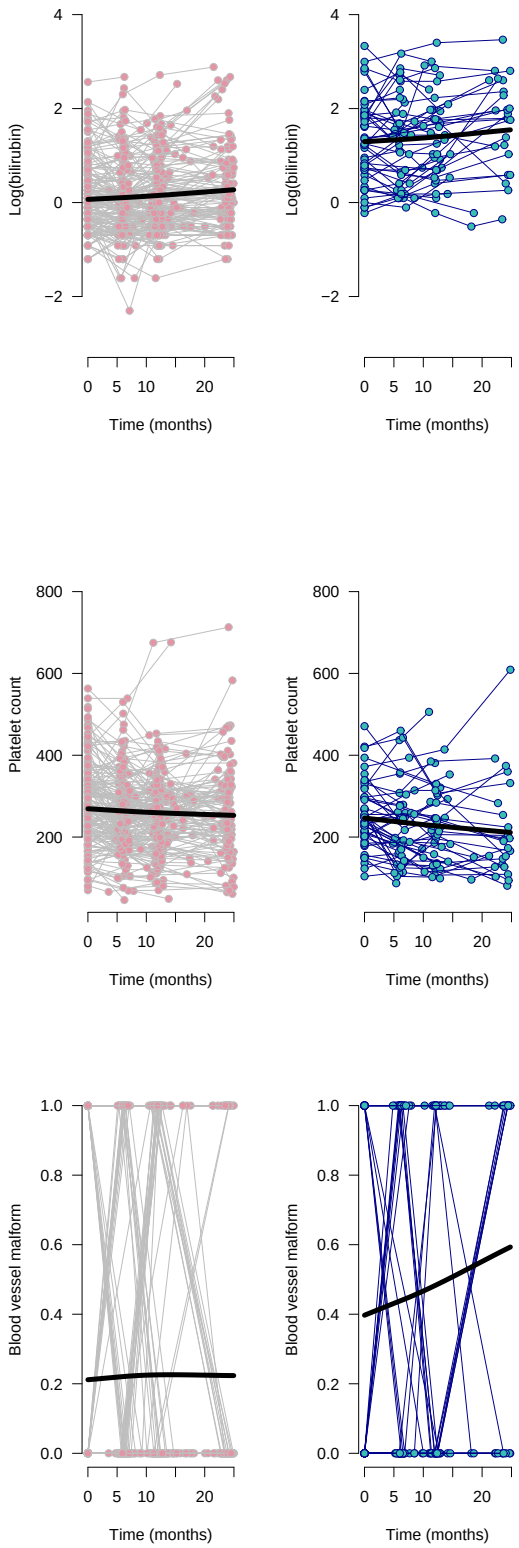

Figure 5.1: Profiles of three markers log(bilirubin), platelet counts and blood vessel malformation (spiders) for patients who were known to be alive at 5 years (Group 0, left panel) and who died after 2.5 years (Group 1, right panel). The thin lines show the profiles of individuals in the PBC data, and the thick lines show the overall mean, as estimated by loess.

residual error ($\epsilon_{i,1,j}$). It is assumed that this error is independent and follows a normal distribution with mean 0 and variance σ_1^2. Also, it is assumed that the error terms are independent of the random effects $\mathbf{b}_i$.

For each MCMC, the results are based on 10,000 iterations of 1:10 thinned samples after a burn-in of 5000 iterations. To assess the prediction results the leave-one-out cross-validation is applied for the PBC example. The MGLMMs are fit using the `GLMM_MCMC` function whilst the LoDA is performed using the `GLMM_longitDA2` from the `mixAK` package (Komárek and Komárková, 2014) in `R`. Sensitivity, specificity, probability of correct classification (PCC), positive predictive value (PPV) and negative predictive value (NPV) are computed for each model by using the optimal cutoff value. The AUC is measured for the marginal prediction approach to compare the four models (with different random effects specifications).

Table 5.1: Penalized Expected Deviance for models with different number of mixture components (K = 1, 2, 3, 4) in the random effects.

Group	$K = 1$	$K = 2$	$K = 3$	$K = 4$
Group 0	11112.80	**11021.41**	11046.15	11160.76
Group 1	**2987.29**	3469.04	4439.53	4547.39

Table 5.1 presents penalised expected deviance values (PED values, Plummer (2008)) for four different models in the two groups. PED can be used to select the most appropriate model among several considered models. Komárek and Komárková (2014) have suggested using PED for mixture models comparison. In Group 0, $K = 2$ shows the lowest values of PED which means that assuming a two-component mixture of multivariate normal distributions for the random effects gives the best fit to the data. However, in Group 1, $K = 1$ provides a better model fit with a smaller PED compared to the other models. It is important to note that Group 1 only contained 51 patients and so it is unlikely that all the parameters in a complex model would be well estimated in such a small sample. This naturally favours the simpler assumptions regarding the distribution of the random effects (e.g., $K = 1$).

The classification results of four possible prediction models for the PBC data are presented in Table 5.2. In this table, the reported accuracy measures are calculated

Table 5.2: Prediction accuracy from leave-one-out cross-validation of the marginal approach with K mixture components in the random effects distribution ($K = 1, 2, 3, 4$) using the PBC data.

Model	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
$K = 1$	0.20	0.78	0.82	0.81	0.86	0.53	0.94
$K = 2$	0.07	0.82	0.73	0.75	0.84	0.43	0.94
$K = 3$	0.01	0.65	0.81	0.78	0.74	0.46	0.90
$K = 4$	0.02	0.65	0.62	0.62	0.64	0.30	0.87

at the optimal cutoff point which is selected as the closest point to the top left corner of the ROC plot. A single multivariate normal distribution for the random effects distribution gives the best prediction accuracy compared to the other three models with 81% overall of patients correctly classified (PCC). Furthermore, this model is able to predict patients who will die or require a liver transplant with 78% accuracy (sensitivity), and 94% of patients predicted to survive without transplant truly did survive without a liver transplant for five years (NPV). 53% of patients predicted to die or need transplant truly did die or require a liver transplant (PPV) and 82% of patients who will survive without a transplant are correctly identified (specificity).

The ROC curve for the single multivariate normal mixture component model is the closest to the upper left corner (see Figure 5.2) which gives the higher value of the AUC with 0.86. However, the two mixture components ($K = 2$) model still works well in terms of AUC 0.84. The complex models that included more than 2-mixture components ($K = 3$, 4) do not perform well, with 0.74 and 0.64 AUCs, respectively. A comparison of the four models in the PBC data reveals that models with more than one mixture component ($K = 2$, 3, 4) in the random effects distribution provide worse classification accuracy.

In contrast to this result, Komárek et al. (2010) suggested that using a normal mixture in the random effects distribution improves the classification accuracy in their application. Komárek et al. (2010) showed using a similar dataset that when the random effects distribution is a two mixture component (K = 2), the AUC (area under the curve) is better than one mixture component (K = 1). In their analyses, they used three continuous markers, bilirubin, albumin and alkaline phosphotase. It is not yet

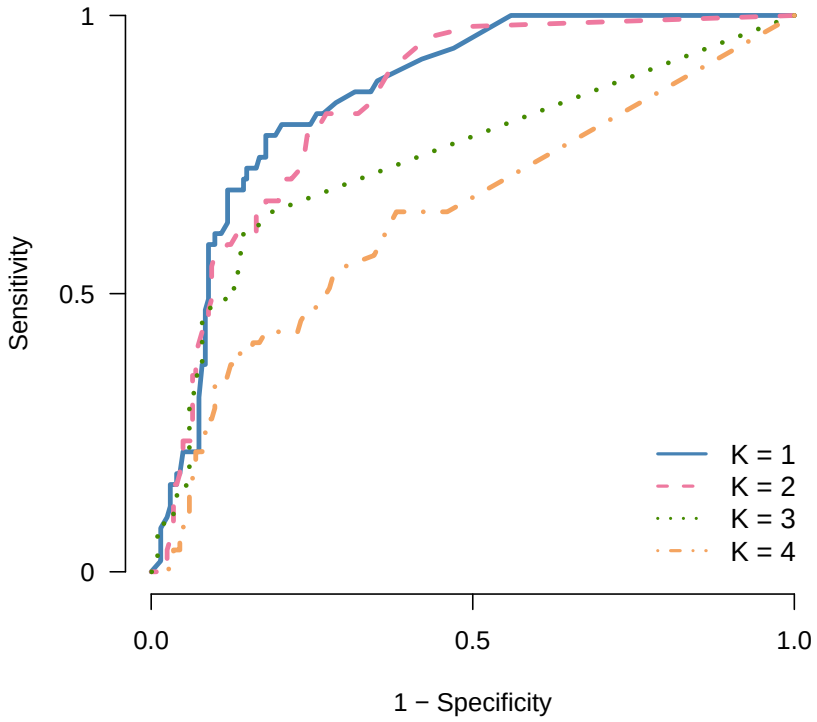

Figure 5.2: Receiver Operating Characteristic (ROC) curves for models with $K = 1$, 2, 3, 4 mixture components in the PBC data.

clear whether classification accuracy is made worse by a misspecified random effects distribution. As is pointed out in the introduction to this chapter, there is a degree of uncertainty around the effect of random effects misspecification on parameter estimates in GLMMs. In the next section, I explore how the misspecified random effects distribution might affect the classification accuracy using a simulation study.

5.3 Simulation

In Section 5.2, the PBC example is used to demonstrate how the selection of random effects distribution impacted the classification accuracy. I now use a simulation study to investigate how robust classification accuracy is to misspecification of the random effects distribution. I also investigate how the sample size and number of repeated

measurements influence the classification performance in the presence of potentially misspecified random effects.

5.3.1 Simulation Setup

The design of the simulation is based on the PBC dataset. I simulated a dataset consisting of three different markers: bilirubin as a continuous marker, platelet count as a discrete marker and blood vessel malformation as a binary marker. Two different sample sizes are considered in this simulation, 250 and 2,500 patients. To be consistent with the PBC data, for each sample size I assume that 80% of patients are alive after five years without requiring a transplant and 20% of patients who were alive at 2.5 years subsequently died or required transplant before five years. Following the simulation design of Komárek et al. (2013), the observed process is sampled at four-time points ($n_i = 4$) at baseline, then after approximately six months, one year and two years. I assumed two groups for discrimination, 0 and 1. For small sample size and Group 0, I considered 200 (2,000 for the large sample size) patients who survived after five years without requiring a liver transplant and 50 (500 for the large sample size) patients who did not survive or who needed the transplant at some time between 2.5 and 5 years. For each patient, the four visit times were generated as follows: the first visit time was set to 0 and uniform distributions in the intervals (170, 200), (350, 390) and (710, 770) days were used to generate the remaining visit times. The elements of mean vectors and variance-covariance matrices considered for the three markers for each group are presented in specific tables in each scenario. At each time point I simulated values for each marker, by first generating random effects from a multivariate normal distribution with mean vector and covariance matrix given in Table 5.3 and then assuming a generalised linear mixed model with fixed effects parameters shown in Table 5.3. Tables 5.4 and 5.5 show the values of three markers that are used to simulate two and three components for the random effects distribution with a small departure from normality, respectively, and to simulate two and three components for the random effects distribution with a large departure from normality are presented in Tables 5.6,

5.7.

The structure of the MGLMM for the simulation study is presented in Equation 5.1. For the purposes of this simulation study, data are supposed to be balanced (i.e., no missing values) across time points.

A number of simulation scenarios are investigated. First, the typical assumption of a single normal distribution for the random effects is chosen as the true distribution. Further, two different random effects distributions, two and three mixture components are included in the simulation study. In each of them, two scenarios are utilised. For the first scenario, the parameter estimates in this scenario are close to the PBC parameters. For the second scenario, the parameter estimates are chosen to show a large departure from normality. Finally, two additional scenarios consider a T-distribution with 3 and 5 degrees of freedom as the true random effects distribution. In total, for each of the two sample sizes, there are seven different simulation scenarios.

Single normal distribution

In this scenario, the true distribution of the random effects is a single normal distribution ($K = 1$). This scenario aims to explore how much using a wrong, but more flexible mixture distribution worsens the prediction accuracy compared to using the theoretically correct single normal distribution. Table 5.3 presents the values of three markers that are used to simulated the first scenario. Probability density functions of each random effects for each group are shown in Figure 5.3.

Two and three components of random effects distribution with a small departure from normality

The purpose of this scenario is to explore the influence of minor misspecification of the random effects distribution (i.e., a small departure from normality) has on the classification accuracy. Published studies on misspecified random effects suggest that

Table 5.3: Model parameters for the random effects under the assumption that the random effects jointly follow a single component multivariate normal distribution.

Parameters	Group 0	Group 1
log(bilirubin)		
E(intercept:log(bilirubin))	2.19×10^{-2}	1.23
E(slope:log(bilirubin))	9.82×10^{-3}	2.42×10^{-2}
SD(intercept:log(bilirubin))	6.88×10^{-1}	8.40×10^{-1}
cor(intercept:log(bilirubin),slope:log(bilirubin))	2.31×10^{-1}	-1.59×10^{-1}
cor(intercept:log(bilirubin),intercept:platelet)	-1.69×10^{-1}	2.52×10^{-1}
cor(intercept:log(bilirubin),slope:platelet)	-2.06×10^{-1}	-1.91×10^{-1}
cor(intercept:log(bilirubin),intercept:spiders)	3.47×10^{-1}	2.61×10^{1}
SD(slope:log(bilirubin))	1.13×10^{-2}	1.46×10^{-2}
cor(slope:log(bilirubin),intercept:platelet)	2.58×10^{-2}	-1.94×10^{-1}
cor(slope:log(bilirubin),slope:platelet)	-2.41×10^{-1}	9.92×10^{-2}
cor(slope:log(bilirubin),intercept:spiders)	3.04×10^{-1}	4.24×10^{-2}
Platelets		
E(intercept:platelet)	5.54	5.46
E(slope:platelet)	-4.29×10^{-3}	-1.14×10^{-2}
SD(intercept:platelet)	3.72×10^{-1}	3.44×10^{-1}
cor(intercept:platelet,slope:platelet)	-4.66×10^{-2}	6.57×10^{-2}
cor(intercept:platelet,intercept:spiders)	-8.00×10^{-2}	-2.45×10^{-1}
SD(slope:platelet)	5.64×10^{-3}	1.50×10^{-2}
cor(slope:platelet,intercept:spiders)	-1.68×10^{-1}	-7.40×10^{-2}
Spiders		
E(intercept:spiders)	-2.54	-6.80×10^{-1}
SD(intercept:spiders)	3.00	1.91
α (spiders)	1.42×10^{-2}	4.75×10^{-2}

when the departure from normality is small, there is not a significant effect on parameter estimates when choosing a single normal distribution. The next simulations are set out to assess whether this holds for classification accuracy as well. Figures 5.4 and 5.6 show the observed longitudinal profiles of the simulation with 2 and 3 mixture components in the random effects distribution of all the markers.

Tables 5.4 and 5.5 provide the values of the parameters used for the simulation scenarios when the random effects follow a multivariate normal distribution with 2-component and 3-component, respectively. It is not clear from a visual inspection that the distributions of the random effects came from two or three mixture components as the amount of deviation from normality is small (see Figures 5.5 and 5.7), reflecting the fact that this scenario is designed to show only a small departure from normality.

Table 5.4: Model parameters for the random effects under the assumption that the random effects jointly follow a 2-component multivariate normal distribution with a small departure from normality.

	Group 0		Group 1	
	1^{st}comp.	2^{nd}comp	1^{st}comp.	2^{nd}comp
Weight	5.4×10^{-1}	4.6×10^{-1}	2.3×10^{-1}	7.7×10^{-1}
log(bilirubin)				
E(intercept:log(bilirubin))	-3.1×10^{-1}	8.1×10^{-1}	2.1	8.2×10^{-1}
E(slope:log(bilirubin))	3.2×10^{-3}	1.8×10^{-2}	3.4×10^{-3}	2.9×10^{-3}
SD(intercept:log(bilirubin))	4.1×10^{-1}	7.7×10^{-1}	1.5	1.6
cov(intercept:log(bilirubin),slope:log(bilirubin))	-2.3×10^{-4}	-2.1×10^{-3}	4×10^{-5}	3.6×10^{-3}
cov(intercept:log(bilirubin),intercept:platelet)	-4.1×10^{-3}	-1.7×10^{-2}	-1.6×10^{-2}	9×10^{-2}
cov(intercept:log(bilirubin),slope:platelet)	-2.3×10^{-4}	-7.3×10^{-4}	-1.6×10^{-3}	1.2×10^{-2}
cov(intercept:log(bilirubin),intercept:spiders)	1.3×10^{-1}	3.8×10^{-1}	4.4×10^{-2}	-8.5×10^{-1}
SD(slope:log(bilirubin))	5.2×10^{-3}	1.2×10^{-2}	9.4×10^{-3}	1.5×10^{-2}
cov(slope:log(bilirubin),intercept:platelet)	2.4×10^{-5}	1.7×10^{-3}	-3.9×10^{-4}	4.4×10^{-4}
cov(slope:log(bilirubin),slope:platelet)	-7.6×10^{-6}	-1.8×10^{-6}	1.1×10^{-6}	1.5×10^{-5}
cov(slope:log(bilirubin),intercept:spiders)	4.2×10^{-3}	-1.9×10^{-3}	1.2×10^{-4}	4.9×10^{-3}
Platelets				
E(intercept:platelet)	5.8	5.2	5.3	5.5
E(slope:platelet)	-3.2×10^{-3}	-5.7×10^{-3}	-8.1×10^{-2}	7.5×10^{-3}
SD(intercept:platelet)	2.6×10^{-1}	4.6×10^{-1}	6.1×10^{-1}	5.1×10^{-1}
cov(intercept:platelet,slope:platelet)	-2.1×10^{-4}	-1.4×10^{-4}	1.5×10^{-3}	3×10^{-4}
cor(intercept:platelet,intercept:spiders)	6×10^{-2}	-1.2×10^{-2}	1×10^{-1}	-2.1×10^{-1}
SD(slope:platelet)	4×10^{-3}	7.8×10^{-3}	3×10^{-2}	2.6×10^{-2}
cov(slope:platelet,intercept:spiders)	-3.8×10^{-3}	7.8×10^{-4}	8.2×10^{-3}	-1.7×10^{-2}
Spiders				
E(intercept:spiders)	-5.2	-3×10^{-1}	-6.7	-1.1
SD(intercept:spiders)	3.4	2.3	3.3	4.1

Table 5.5: Model parameters for the random effects under the assumption that the random effects jointly follow a 3-component multivariate normal distribution with a small departure from normality.

	Group 0			Group 1		
	1^{st}comp.	2^{nd}comp.	3^{rd}comp.	1^{st}comp.	2^{nd}comp.	3^{rd}comp.
Weight	4×10^{-2}	2.7×10^{-1}	6.9×10^{-1}	2.1×10^{-2}	4.5×10^{-1}	5.3×10^{-1}
log(bilirubin)						
E(intercept:log(bilirubin))	-1.5×10^{-1}	-1.5×10^{-1}	6.4×10^{-1}	-5×10^{-1}	2.5	2.5
E(slope:log(bilirubin))	1.7×10^{-3}	4.5×10^{-2}	1.2×10^{-3}	2.7×10^{-2}	-5.1×10^{-2}	2.7×10^{-2}
SD(intercept:log(bilirubin))	9.5×10^{-1}	7.7×10^{-1}	8.3×10^{-1}	7.5×10^{-1}	2.2×10^{-1}	6.2×10^{-1}
cov(intercept:log(bilirubin),slope:log(bilirubin))	-8.1×10^{-4}	-3.6×10^{-3}	4.6×10^{-4}	-1.2×10^{-2}	-3.9×10^{-5}	-4.1×10^{-3}
cov(intercept:log(bilirubin),intercept:platelet)	-1.7×10^{-1}	4.9×10^{-2}	-5.7×10^{-4}	1×10^{-1}	3.6×10^{-3}	-6.3×10^{-3}
cov(intercept:log(bilirubin),slope:platelet)	-5×10^{-3}	-1.1×10^{-3}	-2.8×10^{-4}	-9.2×10^{-3}	-1.4×10^{-4}	-3.8×10^{-3}
cov(intercept:log(bilirubin),intercept:spiders)	1.4	5×10^{-1}	4×10^{-1}	-4.1×10^{-1}	1.6×10^{-2}	9.2×10^{-2}
SD(slope:log(bilirubin))	2.6×10^{-2}	1.1×10^{-2}	6.4×10^{-3}	2.6×10^{-2}	3.8×10^{-3}	1.2×10^{-2}
cov(slope:log(bilirubin),intercept:platelet)	-5.1×10^{-4}	2.8×10^{-3}	8.1×10^{-5}	-4.2×10^{-3}	-3.9×10^{-5}	3.4×10^{-4}
cov(slope:log(bilirubin),slope:platelet)	-7.9×10^{-4}	9.4×10^{-6}	-1.1×10^{-5}	4.3×10^{-4}	3.3×10^{-7}	9.2×10^{-5}
cov(slope:log(bilirubin),intercept:spiders)	-2.2×10^{-2}	3.2×10^{-4}	7.3×10^{-3}	1.8×10^{-2}	1.2×10^{-4}	-1.5×10^{-3}
Platelets						
E(intercept:platelet)	5.2	5.8	5.8	5.4	5.4	5.7
E(slope:platelet)	-7.1×10^{-3}	-3.9×10^{-2}	-7.6×10^{-3}	-1.5×10^{-2}	1.1×10^{-2}	-1×10^{-2}
SD(intercept:platelet)	4.3×10^{-1}	2.5×10^{-1}	2.4×10^{-1}	3.2×10^{-1}	8.9×10^{-2}	1.8×10^{-1}
cov(intercept:platelet,slope:platelet)	2.7×10^{-3}	-6.2×10^{-4}	-1.7×10^{-4}	-2.3×10^{-3}	7.9×10^{-6}	2.2×10^{-4}
cor(intercept:platelet,intercept:spiders)	-5.4×10^{-1}	3.2×10^{-2}	7.5×10^{-2}	-1.2×10^{-1}	-8×10^{-3}	-1.9×10^{-2}
SD(slope:platelet)	2.4×10^{-2}	3.1×10^{-2}	4.3×10^{-3}	2.2×10^{-2}	4.2×10^{-3}	1.2×10^{-2}
cov(slope:platelet,intercept:spiders)	1.2×10^{-1}	9.4×10^{-5}	-3.8×10^{-3}	1.5×10^{-2}	-2.3×10^{-4}	-2×10^{-3}
Spiders)						
E(intercept:spiders)	-2.5	-1.9	-4.6	-3.1	1.2	-2
SD(intercept:spiders)	4.7	2	3.1	1.4	4.7×10^{-1}	9.6×10^{-1}

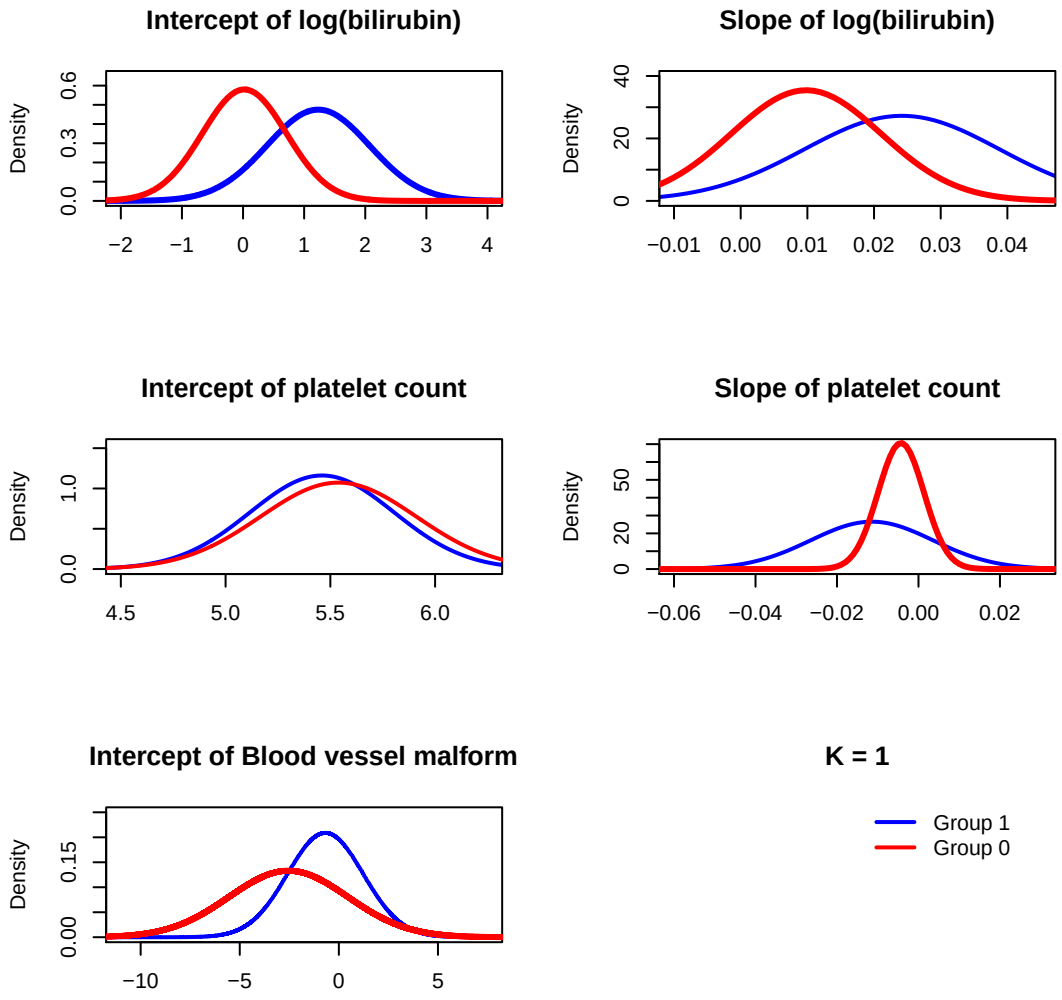

Figure 5.3: Density functions of five random effects distributions for the single normal assumption. Red and blue lines show the density functions for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1) respectively.

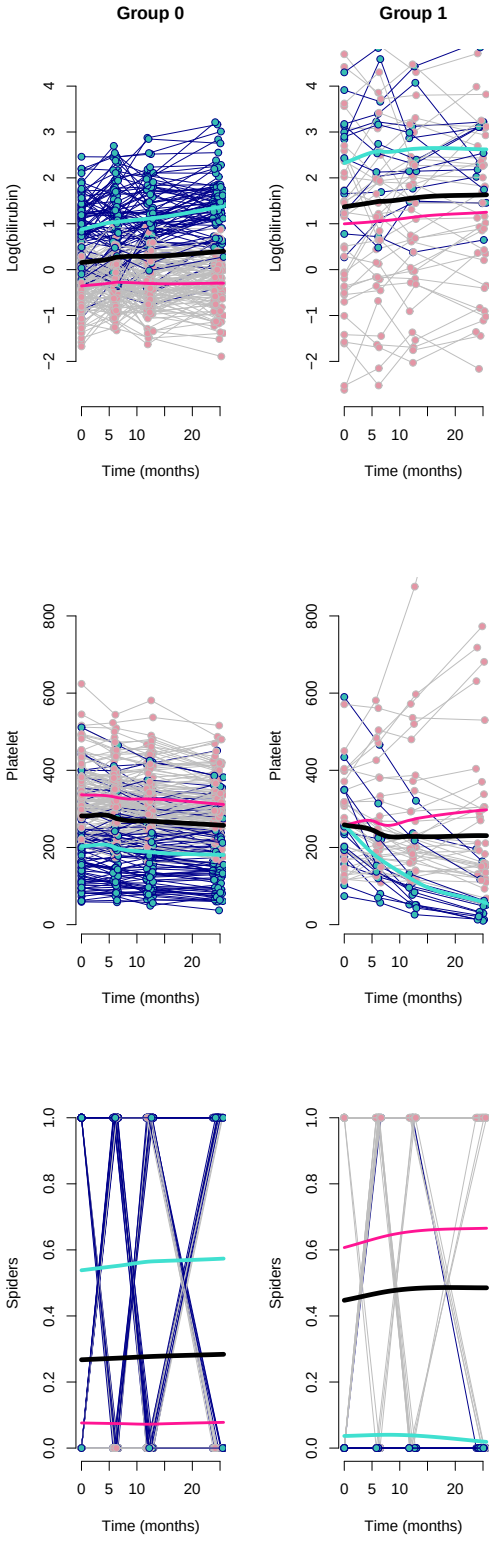

Figure 5.4: Profiles from randomly selected simulated dataset with 2 components normal distribution with small departure from normality. Profiles of three markers log(bilirubin), platelet counts and blood vessel malformation(spiders) for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1). The black line shows the overall mean, the pink and blue lines show the mean of two mixture components, estimated using loess.

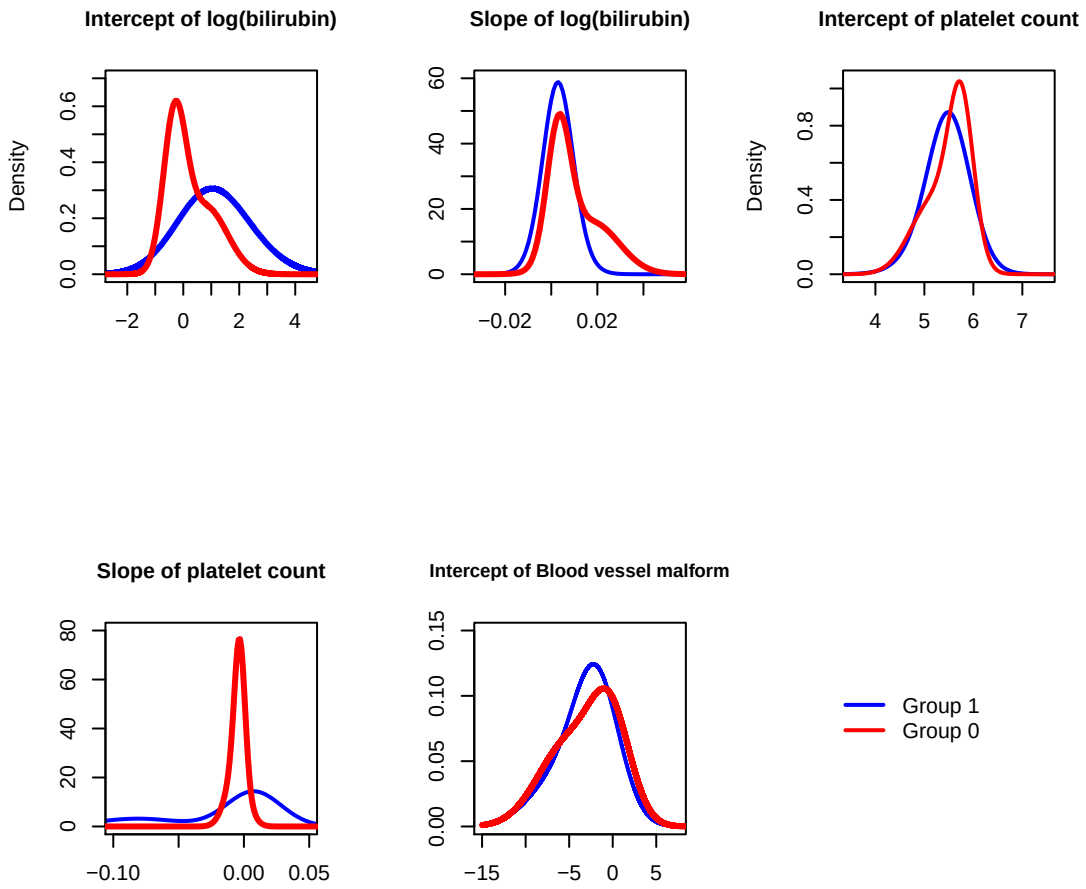

Figure 5.5: Density functions of the random effects for the 2-components assumption with small departure from normality for two groups. Red and blue lines show the density function for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1) respectively.

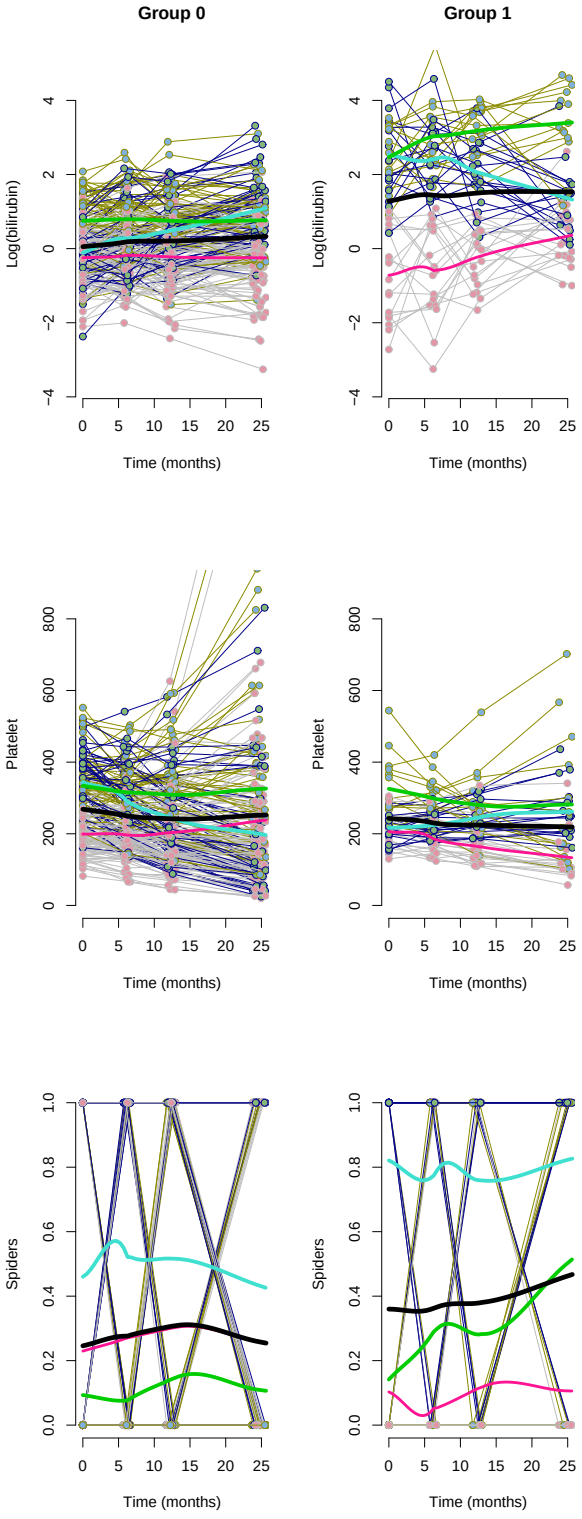

Figure 5.6: Profiles from randomly selected simulated data with 3 mixture components normal distribution with small departure from normality. Profiles of three markers log(bilirubin), platelet counts and blood vessel malformation(spiders) for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1). The black line shows the overall mean, while the pink, green and blue lines show the mean of three mixture components, estimated using loess.

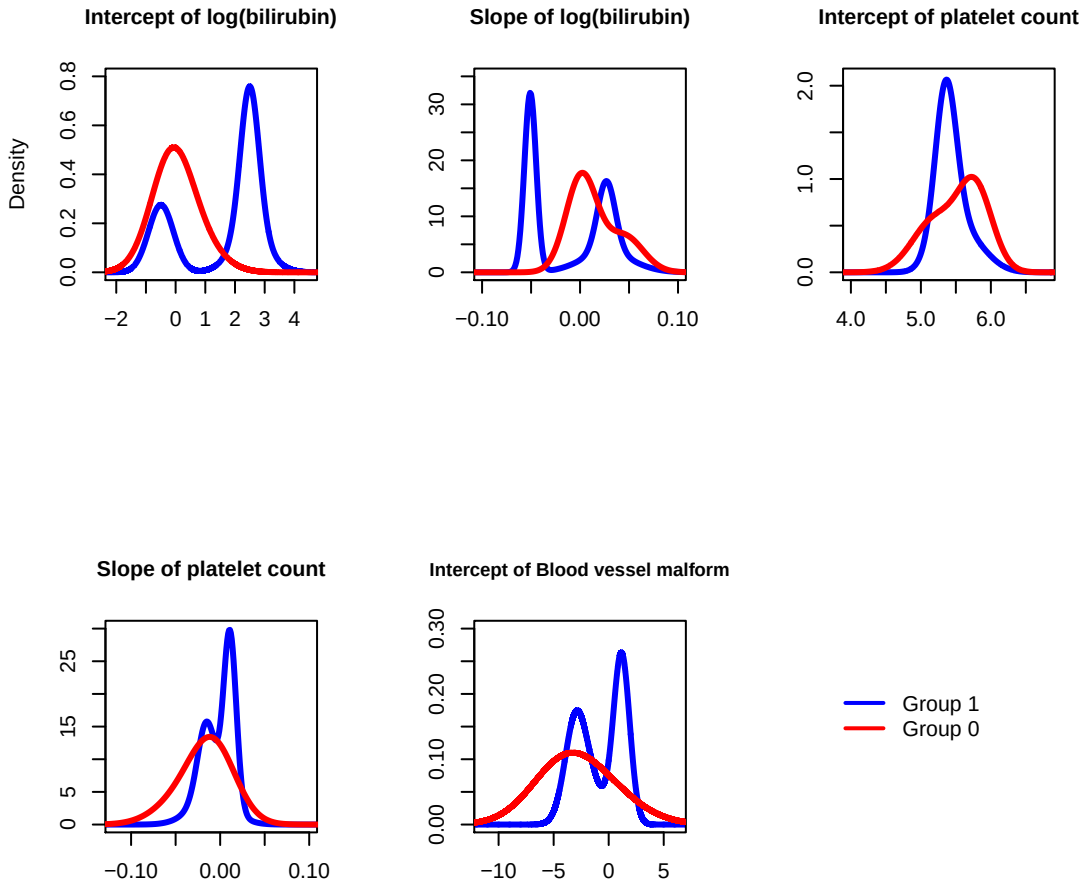

Figure 5.7: Probability density functions of five random-effects elements which follow a 3 mixture normal distribution with small departure from normality for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1).

Two and three components of random effects distribution ($K = 2$ and $K = 3$) with a large departure from normality

The primary purpose of this scenario is to investigate whether the model that assumed a single multivariate normal distribution can capture the 'true distribution' when there is a substantial departure from normality and see whether using more flexible models provides more accurate classifications. Basically, this simulation study assumes that the means for the two or three components have the same values for the two groups, but that the variances are different, such that for one group the data are narrowly spread around each mean and for the other group the data are widely spread. Figures 5.8 and 5.11 show the observed longitudinal profiles of the simulation with 2 and 3 mixture components in the random effects distribution of all the markers.

Figures 5.9 and 5.10 illustrate the scenarios and Tables 5.6 and 5.7 present the parameters used for the simulations when the random effects follow a multivariate normal distribution with 2-component and 3-component, respectively, and a large departure from normality.

Table 5.6: Model parameters for the random effects under the assumption that the random effects jointly follow a 2-component multivariate normal distribution with a large departure from normality.

	Group 0		Group 1	
	1^{st}comp.	2^{nd}comp	1^{st}comp.	2^{nd}comp
Weight	0.54	0.46	0.55	0.45
log(bilirubin)				
E(intercept:log(bilirubin))	-2×10^{-1}	1.9	-2×10^{-1}	2
E(slope:log(bilirubin))	3.2×10^{-3}	2.9×10^{-2}	2.9×10^{-2}	3.4×10^{-3}
SD(intercept:log(bilirubin))	2.2×10^{-1}	4.1×10^{-1}	1.1	1.1
cov(intercept:log(bilirubin),slope:log(bilirubin))	-6.6×10^{-5}	-6×10^{-4}	4×10^{-5}	3.6×10^{-3}
cov(intercept:log(bilirubin),intercept:platelet)	-1.2×10^{-3}	-4.9×10^{-3}	-1.6×10^{-2}	9×10^{-2}
cov(intercept:log(bilirubin),slope:platelet)	-6.7×10^{-5}	-2.1×10^{-4}	-1.6×10^{-3}	1.2×10^{-2}
cov(intercept:log(bilirubin),intercept:spiders)	3.8×10^{-2}	1.1×10^{-1}	4.4×10^{-2}	-8.5×10^{-1}
SD(slope:log(bilirubin))	2.8×10^{-3}	6.2×10^{-3}	6.7×10^{-3}	1.1×10^{-2}
cov(slope:log(bilirubin),intercept:platelet)	6.8×10^{-6}	4.8×10^{-4}	-3.9×10^{-4}	4.4×10^{-4}
cov(slope:log(bilirubin),slope:platelet)	-2.2×10^{-6}	-5.2×10^{-7}	1.1×10^{-6}	1.5×10^{-5}
cov(slope:log(bilirubin),intercept:spiders)	1.2×10^{-3}	-5.3×10^{-4}	1.2×10^{-4}	4.9×10^{-3}
Platelets				
E(intercept:platelet)	5.7	5	5.7	5
E(slope:platelet)	-2.5×10^{-3}	-5.6×10^{-2}	-2.5×10^{-3}	-5.6×10^{-2}
SD(intercept:platelet)	1.4×10^{-1}	2.5×10^{-1}	4.3×10^{-1}	3.6×10^{-1}
cov(intercept:platelet,slope:platelet)	-5.9×10^{-5}	-4.1×10^{-5}	1.5×10^{-3}	3×10^{-4}
cor(intercept:platelet,intercept:spiders)	1.7×10^{-2}	-3.3×10^{-3}	1×10^{-1}	-2.1×10^{-1}
SD(slope:platelet)	2.1×10^{-3}	4.2×10^{-3}	2.1×10^{-2}	1.9×10^{-2}
cov(slope:platelet,intercept:spiders)	-1.1×10^{-3}	2.2×10^{-4}	8.2×10^{-3}	-1.7×10^{-2}
Spiders				
E(intercept:spiders)	2.5	-7.1	2.5	-7.1
SD(intercept:spiders)	1.8	1.2	3.8	3.9

Table 5.7: Model parameters for the random effects under the assumption that the random effects jointly follow a 3-component multivariate normal distribution with a large departure from normality.

	Group 0			Group 1		
	1^{st}comp.	2^{nd}comp.	3^{rd}comp.	1^{st}comp.	2^{nd}comp.	3^{rd}comp.
Weight	3.4×10^{-1}	3.3×10^{-1}	3.3×10^{-1}	3.3×10^{-1}	3.3×10^{-1}	3.4×10^{-1}
log(bilirubin)						
E(intercept:log(bilirubin))	-1.1	9.8×10^{-1}	3.2	-1.1	9.8×10^{-1}	3.2
E(slope:log(bilirubin))	3.2×10^{-2}	1.5×10^{-2}	1.7×10^{-4}	3.2×10^{-2}	1.5×10^{-2}	1.7×10^{-4}
SD(intercept:log(bilirubin))	6.5×10^{-1}	6.5×10^{-1}	6.5×10^{-1}	7.6×10^{-1}	7.6×10^{-1}	7.6×10^{-1}
cov(intercept:log(bilirubin),slope:log(bilirubin))	-3.8×10^{-3}	-3.8×10^{-3}	-3.8×10^{-3}	-7.8×10^{-5}	-7.8×10^{-5}	-7.8×10^{-5}
cov(intercept:log(bilirubin),intercept:platelet)	-5.8×10^{-2}	-5.8×10^{-2}	-5.8×10^{-2}	2.7×10^{-3}	2.7×10^{-3}	2.7×10^{-3}
cov(intercept:log(bilirubin),slope:platelet)	1.4×10^{-2}	1.4×10^{-2}	1.4×10^{-2}	-3.5×10^{-4}	-3.5×10^{-4}	-3.5×10^{-4}
cov(intercept:log(bilirubin),intercept:spiders)	4.3×10^{-1}	4.3×10^{-1}	4.3×10^{-1}	2.2×10^{-4}	2.2×10^{-4}	2.2×10^{-4}
SD(slope:log(bilirubin))	2.7×10^{-2}	2.7×10^{-2}	2.7×10^{-2}	6×10^{-3}	6×10^{-3}	6×10^{-3}
cov(slope:log(bilirubin),intercept:platelet)	2×10^{-3}	2×10^{-3}	2×10^{-3}	2.3×10^{-4}	2.3×10^{-4}	2.3×10^{-4}
cov(slope:log(bilirubin),slope:platelet)	-4.7×10^{-4}	-4.7×10^{-4}	-4.7×10^{-4}	1.9×10^{-5}	1.9×10^{-5}	1.9×10^{-5}
cov(slope:log(bilirubin),intercept:spiders)	-4×10^{-3}	-4×10^{-3}	-4×10^{-3}	6.8×10^{-4}	6.8×10^{-4}	6.8×10^{-4}
Platelets						
E(intercept:platelet)	4.8	5.3	5.8	4.8	5.3	5.8
E(slope:platelet)	-1.9×10^{-2}	-4.3×10^{-2}	-1×10^{-3}	-1.9×10^{-2}	-4.3×10^{-2}	-1×10^{-3}
SD(intercept:platelet)	1×10^{-1}	1×10^{-1}	1×10^{-1}	1.8×10^{-1}	1.8×10^{-1}	1.8×10^{-1}
cov(intercept:platelet,slope:platelet)	-2.6×10^{-3}	-2.6×10^{-3}	-2.6×10^{-3}	4.3×10^{-4}	4.3×10^{-4}	4.3×10^{-4}
cor(intercept:platelet,intercept:spiders)	-5.9×10^{-2}	-5.9×10^{-2}	-5.9×10^{-2}	-1.1×10^{-2}	-1.1×10^{-2}	-1.1×10^{-2}
SD(slope:platelet)	2.5×10^{-2}	2.5×10^{-2}	2.5×10^{-2}	6.2×10^{-3}	6.2×10^{-3}	6.2×10^{-3}
cov(slope:platelet,intercept:spiders)	1.4×10^{-2}	1.4×10^{-2}	1.4×10^{-2}	2.4×10^{-5}	2.4×10^{-5}	2.4×10^{-5}
Spiders						
E(intercept:spiders)	0	-6	6	0	-6	6
SD(intercept:spiders)	6.6×10^{-1}	6.6×10^{-1}	6.6×10^{-1}	1.8	1.8	1.8

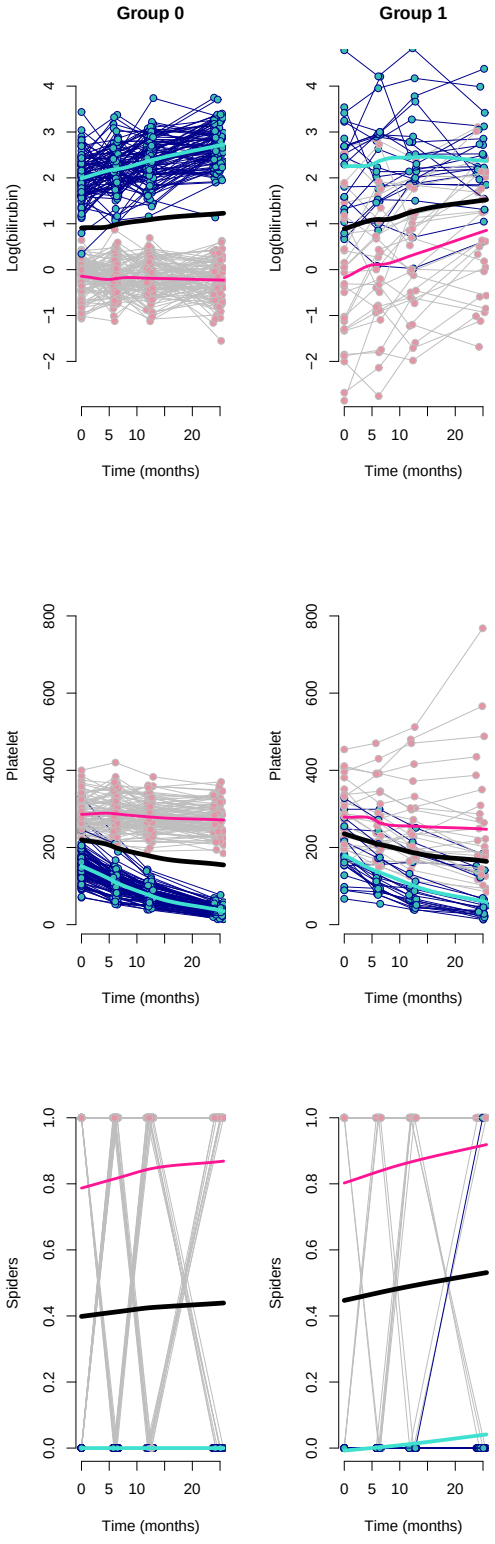

Figure 5.8: Simulated data from 2 mixture components normal distribution with a large departure from normality. Profiles of three markers log(bilirubin), platelet counts and blood vessel malformation(spiders) for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1). The black line shows the overall mean, while the pink and blue lines show the mean of two mixture components, estimated using loess.

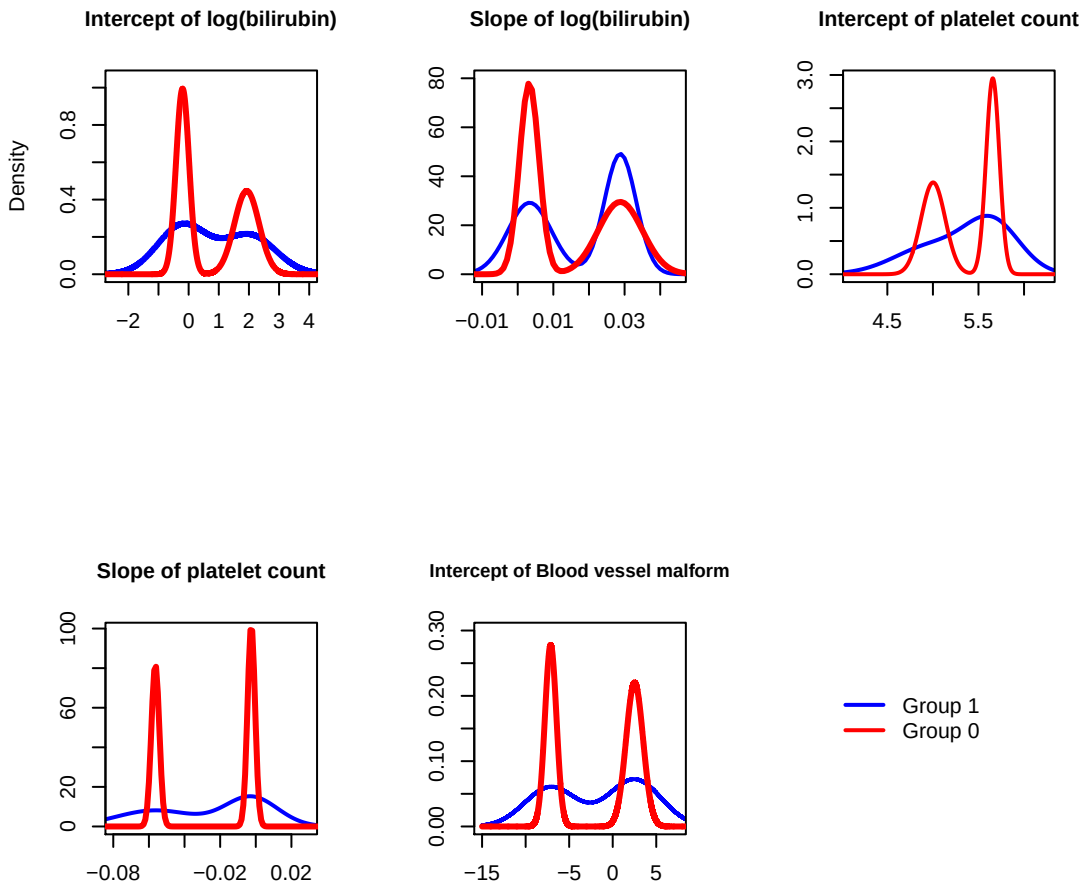

Figure 5.9: Density function for the 2-components normal assumption with large departure from normality. Red and blue lines show the density function for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1) respectively.

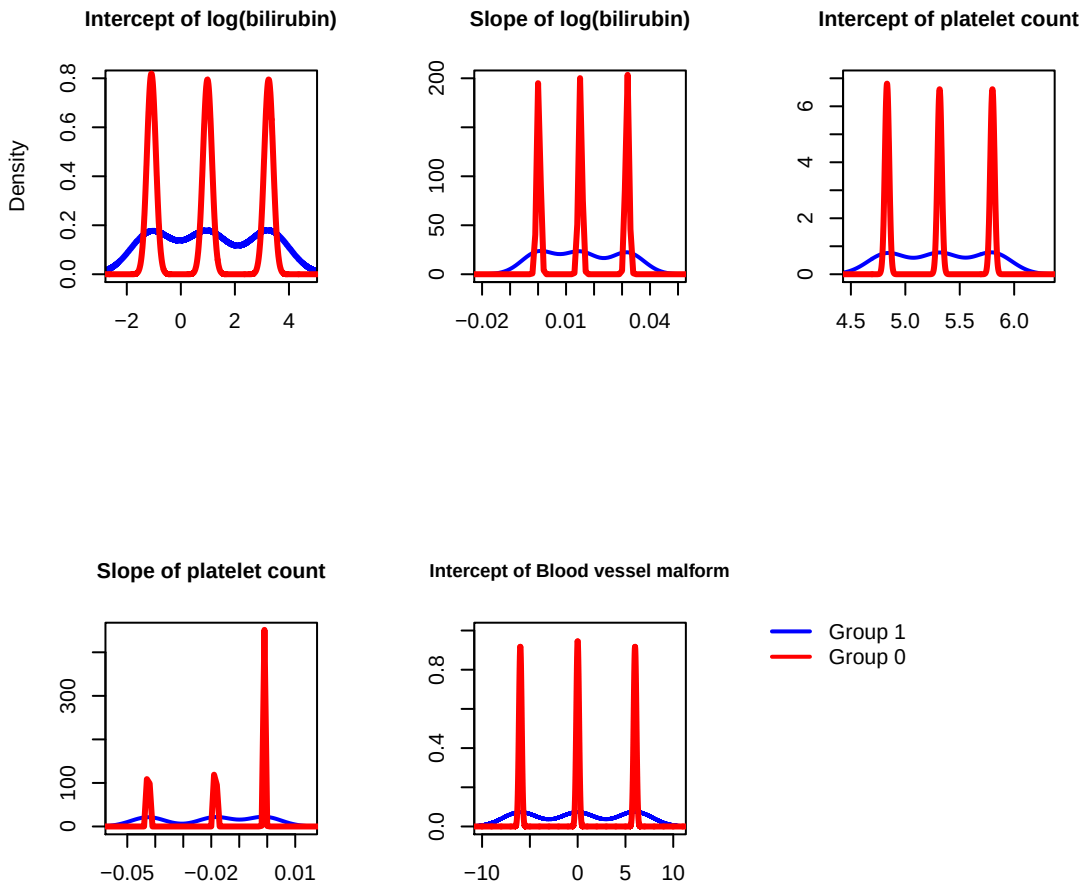

Figure 5.10: Density function for the 3-components normal assumption with a large departure from normality. Red and blue lines show the density function for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1) respectively.

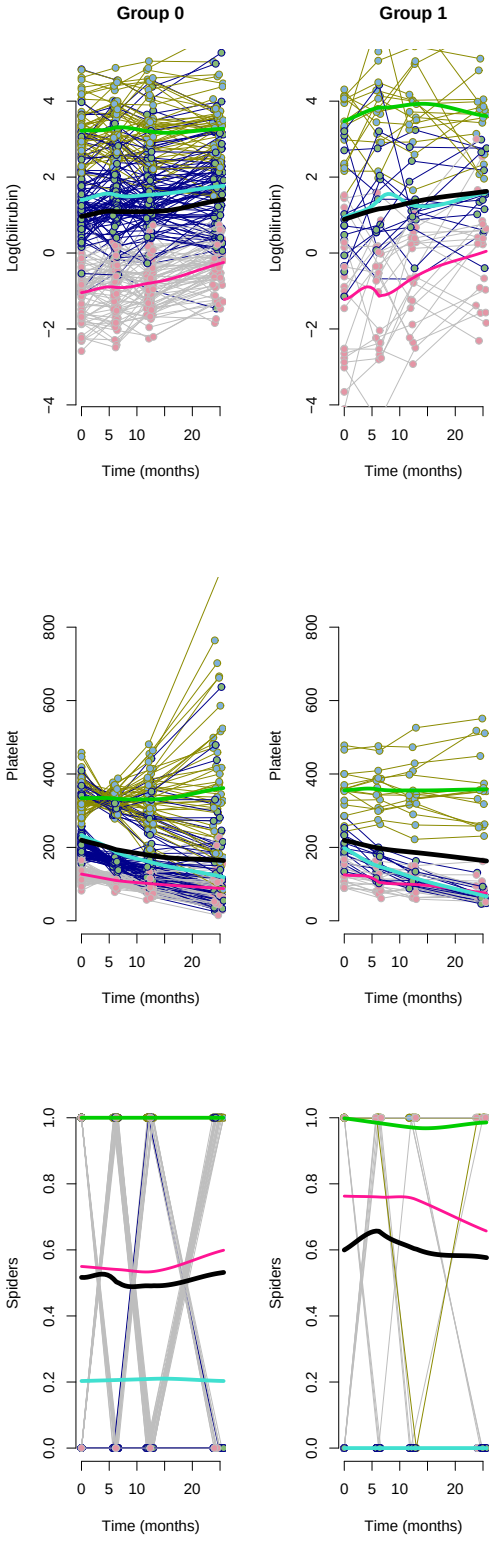

Figure 5.11: Simulated data from 3 mixture components normal distribution with a large departure from normality. Profiles of three markers log(bilirubin), platelet counts and blood vessel malformation(spiders) for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1). The black line shows the overall mean, while the pink, green and blue lines show the mean of three mixture components, estimated using loess.

T-distribution with 3 and 5 degrees of freedom for the random effects

In this case, the true distribution of the random effect is a T-distribution with degrees of freedom, 3 and 5, following the same design that is used for the 2 or 3 mixture components (large and small departures from normality). The density functions of the five random effects for degrees of freedom 3 and 5 in each group (Group 0 and Group 1) differ from each other as seen in Figure 5.12, where the dot lines refer to the T-distribution with 5 degrees of freedom, and the plain lines refer to the T-distribution with 3 degrees of freedom.

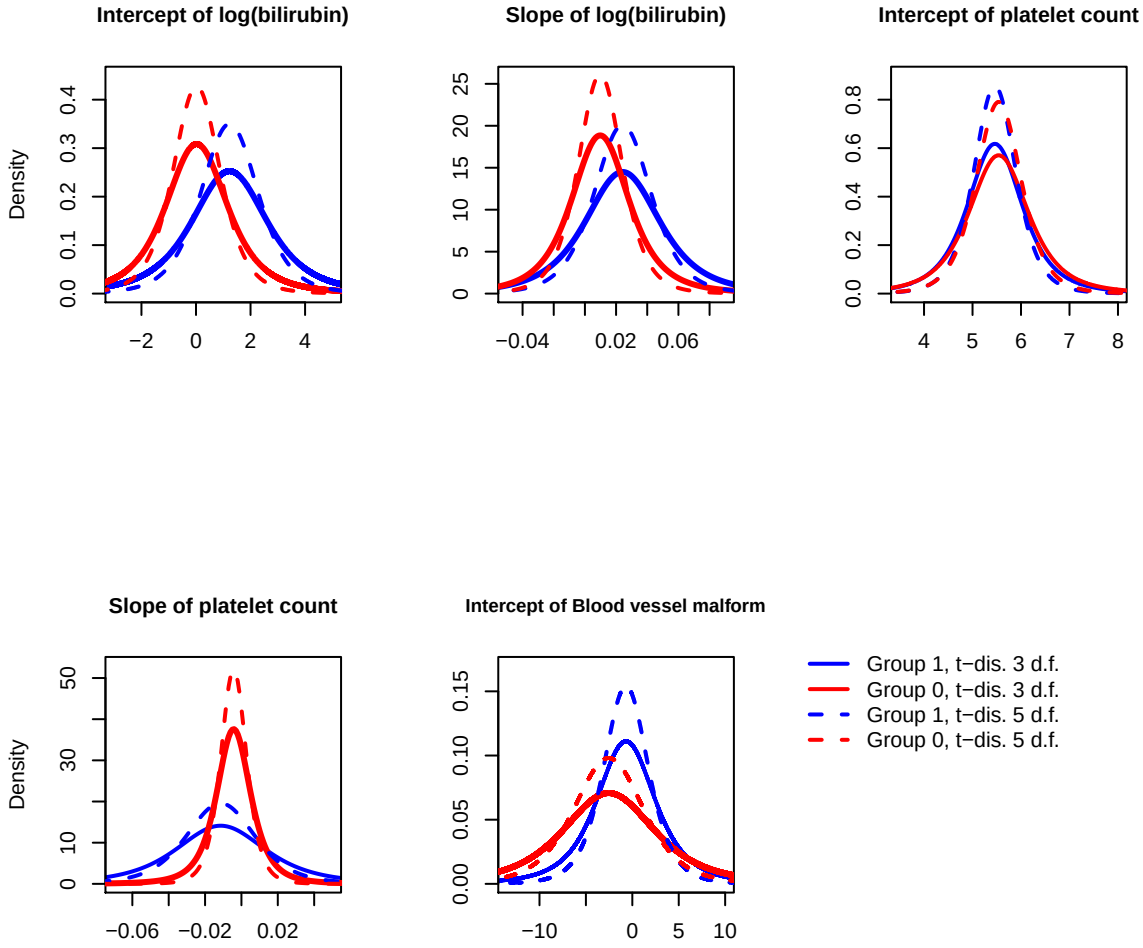

Figure 5.12: T-distribution density function for two groups with 3 (solid lines) and 5 (dashed lines) degrees of freedom. Red and blue lines show the density function for patients who were known to be alive at 5 years (Group 0) and who died after 2.5 years (Group 1) respectively.

Number of visits

It could be hypothesised that the more visits a patient has, then the more accurately the random effects will be estimated. To assess this hypothesis, I simulate nine clinic visits (approximately every three months) per patient for small sample size (i.e., 250

patients) and for the scenarios $K = 1$, $K = 2$ with large departure from normality and T-distribution with 3 degrees of freedom. Each patient has a visit every three months for two years, such that the first visit occurs at $t = 0$ and then at time points by using uniform distributions over the intervals (70, 110), (160, 200), (250, 290), (345, 385), (430, 470), (520, 560), (610, 650) and (710, 750) days to generate the rest of the visits.

Validation methods

Two validation methods are used to assess the classification performance of each model. In particular, leave one out cross-validation is used with small sample size $N = 250$ patients to assess the results. In large sample size $N = 2,500$ patients, I used 70% of data for training and 30% for testing. For each case, 100 datasets were simulated. These 100 simulated datasets were submitted to the cluster system at the University of Liverpool, and it took a month for the task (which involves 17 scenarios in total) to be completed. In each scenario, a MGLMM is used in each group with 15,000 iterations of 1:10 thinned MCMC and the burn-in is 5000 iterations.

5.3.2 Effect of the misspecification of the random effects on classification accuracy

This section describes and discusses the effect of the random effects misspecification on classification accuracy. Two methods of the LoDA are used in this investigation (the marginal and random effects approaches). First I describe how misspecification of the random effects distribution affects the classification accuracy of the marginal prediction approach. The second part moves on to discuss how factors such as the number of measurements per patient influence the random effects prediction approach in the presence of misspecified random effects.

Effect on marginal prediction

The results obtained from the marginal prediction when the true the random effects follow a single normal distribution (model coefficients can be seen in Table 5.3) are shown in Table 5.8.

Table 5.8: Results of the simulation study under the assumption of a single normal distribution for the random effects. Prediction accuracy of the marginal approach from leave-one-out cross validation for N = 250 patients, 70% training and 30% testing for N = 2,500.

Size	K	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
N = 250	K = 1	0.19	0.86	0.87	0.87	0.93	0.63	0.96
	K = 2	0.19	0.84	0.86	0.86	0.91	0.61	0.96
	K = 3	0.42	0.77	0.73	0.74	0.78	0.44	0.93
	K = 4	0.63	0.79	0.78	0.79	0.83	0.49	0.94
N = 2,500	K = 1	0.19	0.87	0.88	0.88	0.94	0.64	0.96
	K = 2	0.19	0.86	0.87	0.87	0.94	0.63	0.96
	K = 3	0.29	0.80	0.85	0.84	0.89	0.58	0.95
	K = 4	0.29	0.72	0.79	0.77	0.79	0.49	0.92

It can be seen from the data in Table 5.8 that the single normal distribution ($K = 1$) gives the best classification accuracy with 87% PCC and an AUC of 0.93. Despite this, the ROC curve (Figure 5.13, $N = 250$) of two mixture components ($K = 2$) is close to the single multivariate normal distribution. Models with more complex mixture components do not improve the prediction accuracy. It is clear from the boxplots that the variability is large for the measurements of sensitivity, specificity, PPC and AUC for the models with $K = 3$ and 4. It seems possible that these results are due to these models being more complicated and the small sample size being relatively small, particularly when one of the group sizes includes just 50 patients.

It is also shown however that increasing the sample size to 10 times (i.e., $N = 2,500$ patients) has a little improvement in terms of AUC as the true distribution of the random effects is a single normal distribution. The model with two mixture components is also able to predict well (see the Table 5.8). Although, this finding is expected, since the single normal distribution is the truth distribution, the two components of a normal

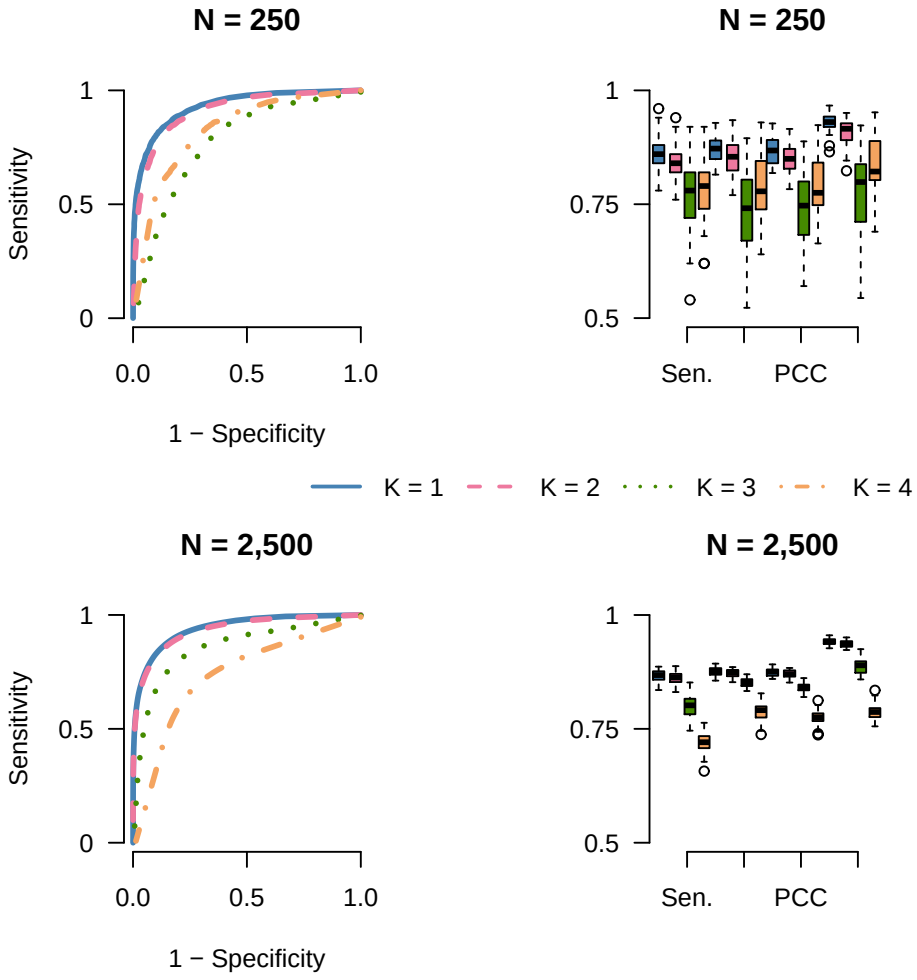

Figure 5.13: Receiver Operating Characteristic curves for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a single normal distribution. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets.

distribution ($K = 2$) performed as well as the single normal distribution, which shows that considering a more flexible distribution, although it is theoretically wrong, would not affect the classification accuracy.

The results of the second and third scenarios where the true random effects distribution follow a 2-component normal mixture distribution with a small and large departure from normality (models coefficients can be seen in Tables 5.4 and 5.6, respectively) can be compared in Table 5.9 and Figures 5.14 and 5.15. For small departure from normality and small sample size, using the true random effects distribution ($K = 2$) or a single

Table 5.9: Prediction accuracy for the marginal approach under the assumption that the random effects follow a 2 component normal mixture distribution.

Size	K	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
				small departure form normality				
250	K = 1	0.15	0.87	0.89	0.89	0.94	0.68	0.96
	K = 2	0.17	0.88	0.90	0.89	0.95	0.69	0.97
	K = 3	0.71	0.80	0.77	0.78	0.82	0.49	0.94
	K = 4	0.76	0.80	0.82	0.81	0.85	0.53	0.94
2,500	K = 1	0.12	0.94	0.97	0.96	0.98	0.88	0.99
	K = 2	0.17	0.96	0.97	0.97	0.99	0.90	0.99
	K = 3	0.30	0.92	0.93	0.93	0.94	0.82	0.98
	K = 4	0.46	0.88	0.90	0.90	0.92	0.73	0.97
				large departure form normality				
250	K = 1	0.18	0.83	0.87	0.86	0.91	0.62	0.95
	K = 2	0.17	0.89	0.92	0.92	0.95	0.75	0.97
	K = 3	0.40	0.82	0.81	0.82	0.85	0.55	0.95
	K = 4	0.63	0.77	0.72	0.73	0.77	0.42	0.93
2,500	K = 1	0.16	0.87	0.90	0.89	0.94	0.69	0.96
	K = 2	0.17	0.92	0.95	0.94	0.98	0.81	0.98
	K = 3	0.67	0.86	0.82	0.83	0.88	0.58	0.96
	K = 4	0.55	0.81	0.80	0.80	0.84	0.54	0.94

normal distribution ($K = 1$) both provide similar results in terms of PCCs and AUCs. Again, more complicated mixture models such as $K = 3$ and 4 do not work well as $K = 1$ and 2, probably due to the small sample size (50 patients). The large sample size ($N = 2,500$ patients) shows improvement in the accuracy measurements, and reduced variability in estimates of these measurements (see Figure 5.14). As expected, different values of K generate different models, and this results in different cut-off values. These cutoffs give the optimum performance (point closest to the top left corner of a ROC curve) for each scenario. There are differences between the cutoff values because each model assigns probabilities based on different numbers of parameters. The cut-off values for the different K models are the average based on 100 repeats of training and testing sets over 100 simulated datasets.

In the case of a large departure from normality, the single normal distribution is

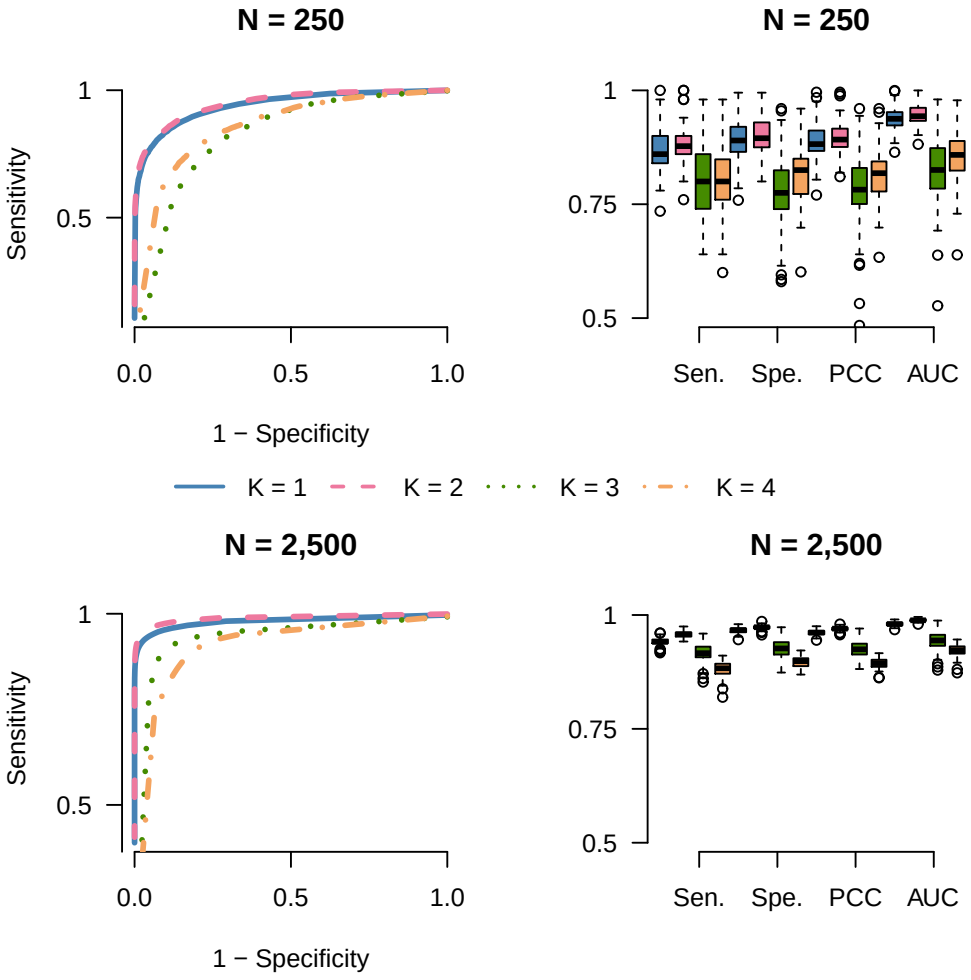

Figure 5.14: Receiver Operating Characteristic curves for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a 2-component mixture of normals with small departure from normality. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets.

unable to capture the true random effects distribution ($K = 2$) (see Table 5.9 and Figure 5.15), and so is unable to achieve as good a classification accuracy as the more flexible $K = 2$ model, which achieved the best accuracy. This result may be explained by the fact that the large departure from normality, where both groups have the same locations of 2 components, and the only difference is the variability between the groups, helps the model with 2 components to distinguish between the two groups.

The evidence presented in these two scenarios suggests that the single normal distri-

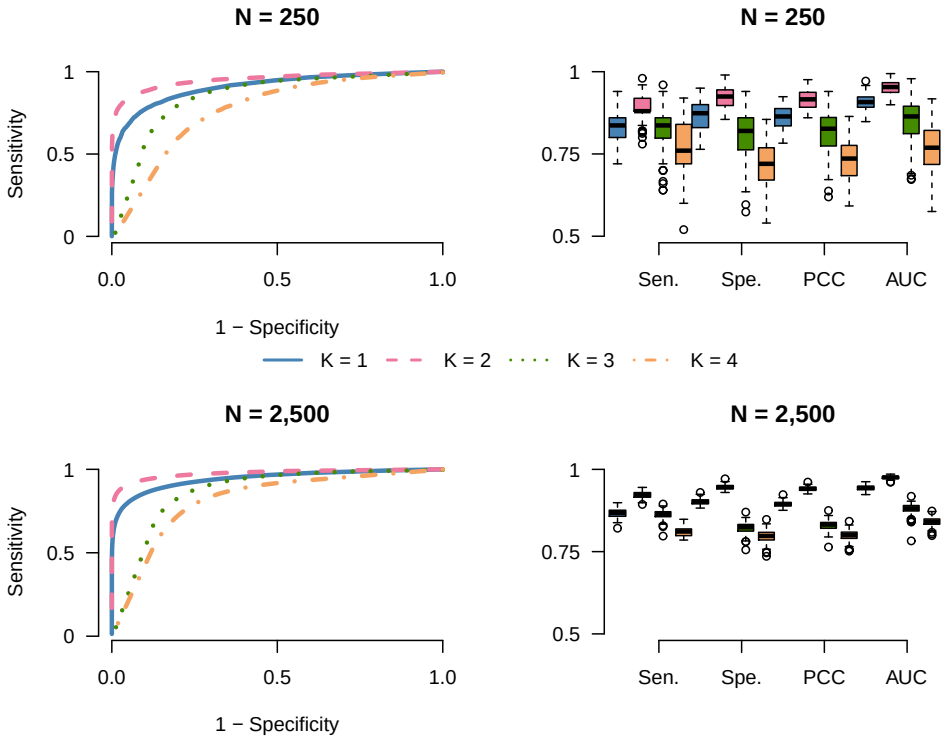

Figure 5.15: Receiver Operating Characteristic curves for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a 2-component mixture of normals with large departure from normality. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets.

bution works well in the case where there is a difference between two groups in locations and variability as well. However, in the case where the locations of the two groups are equal, and the difference in variability between the groups is large, the single normal distribution is unable to distinguish well between the two groups. The 2-component mixture model can capture this difference and use it to generate more accurate group membership probabilities.

The fourth and fifth scenario consider the case where the true distribution of the random effects is a 3-component mixture of normal distributions (models coefficients can be seen in Tables 5.5 and 5.7). Table 5.10 presents the classification accuracy for these scenarios. For the small sample size (250 patients), using a model with 3-components works less well in both scenarios (large and small departures from normality). A possi-

Table 5.10: Prediction accuracy for the marginal approach under the assumption that the random effects follow a 3 component normal mixture distribution.

Size	Model	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
				small departure form normality				
250	K = 1	0.21	0.91	0.94	0.93	0.97	0.79	0.98
	K = 2	0.19	0.93	0.95	0.94	0.98	0.82	0.98
	K = 3	0.21	0.88	0.86	0.87	0.90	0.64	0.97
	K = 4	0.22	0.87	0.89	0.89	0.92	0.68	0.97
2,500	K = 1	0.20	0.91	0.94	0.93	0.97	0.78	0.98
	K = 2	0.19	0.93	0.95	0.94	0.98	0.81	0.98
	K = 3	0.2	0.92	0.93	0.93	0.97	0.79	0.98
	K = 4	0.22	0.76	0.79	0.79	0.78	0.59	0.95
				large departure form normality				
250	K = 1	0.17	0.85	0.89	0.88	0.94	0.67	0.96
	K = 2	0.16	0.90	0.92	0.92	0.96	0.75	0.97
	K = 3	0.082	0.82	0.85	0.85	0.87	0.60	0.95
	K = 4	0.19	0.76	0.75	0.75	0.80	0.45	0.93
2,500	K = 1	0.17	0.86	0.89	0.88	0.95	0.66	0.96
	K = 2	0.17	0.91	0.93	0.92	0.97	0.76	0.98
	K = 3	0.16	0.93	0.94	0.94	0.98	0.81	0.98
	K = 4	0.14	0.77	0.83	0.82	0.85	0.58	0.93

ble explanation for this might be that the model with a small sample size of 50 patients is unable to estimate the parameter accurately since a large number of parameters are required for estimation. It is also worth pointing out in Table 5.10 that the model with 2-components ($K = 2$) provides the best classification accuracy compared with the single model ($K = 1$) and multiple components ($K = 3$ and 4).

In the case where the sample size is increased to 2,500 patients, using the 3-component model can estimate their parameters more accurately, and this is reflected by the fact that this model is comparable to $K = 1$ and 2 with a small departure from normality and it is the best model for the large departure case (see Table 5.10 and Figure 5.17). Therefore, using more flexible models can offer greater accuracy when the departure from normality is substantial. However, complicated models can only be beneficial when the sample size is sufficiently large, even if they capture the 'true'

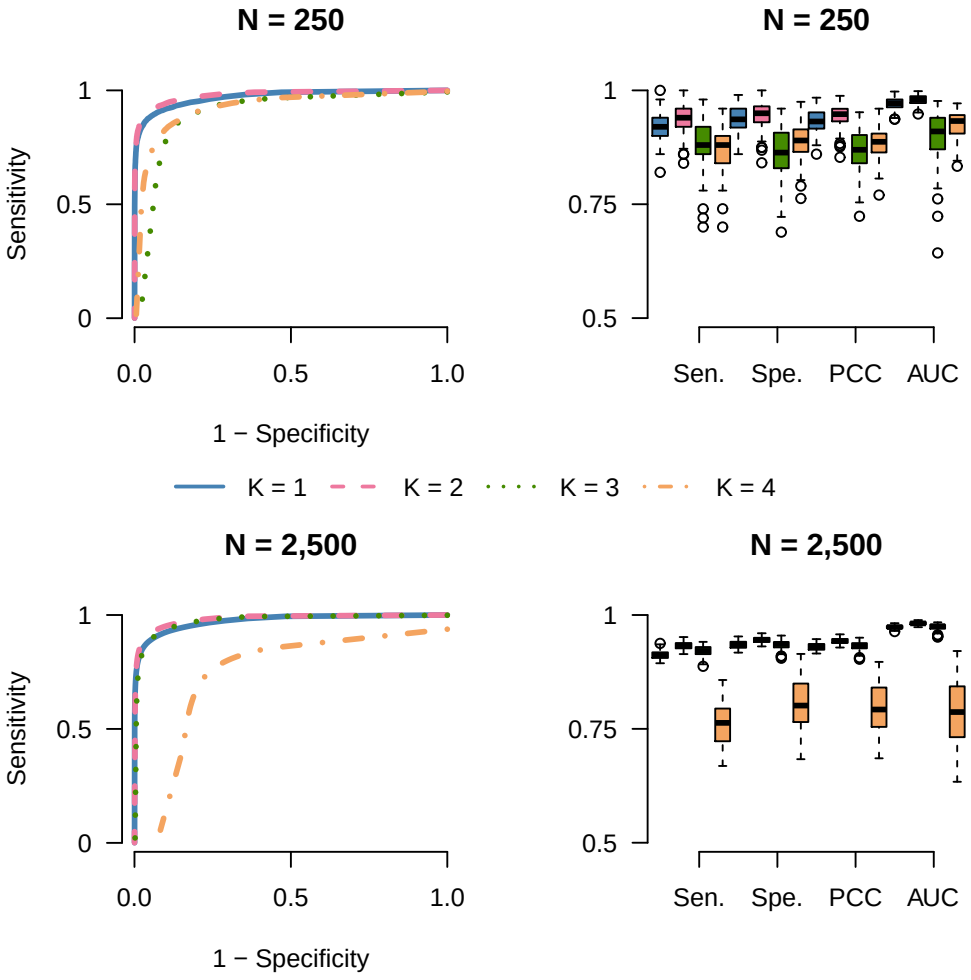

Figure 5.16: Receiver Operating Characteristic curves for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a 3-component mixture of normals with small departure from normality. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets.

distribution.

The remaining scenario of the simulation study is the T-distribution. It can be seen from the data in Table 5.11 that when the true distribution is the T-distribution with small degrees of freedom ($d.f. = 3$), assuming a mixture distribution provides good classification results compared with a single normal distribution ($K = 1$). As expected, for the small sample size (250 patients), the model with a 2-component mixture of normal distributions classifies patients correctly with AUC of 0.82. When the sample

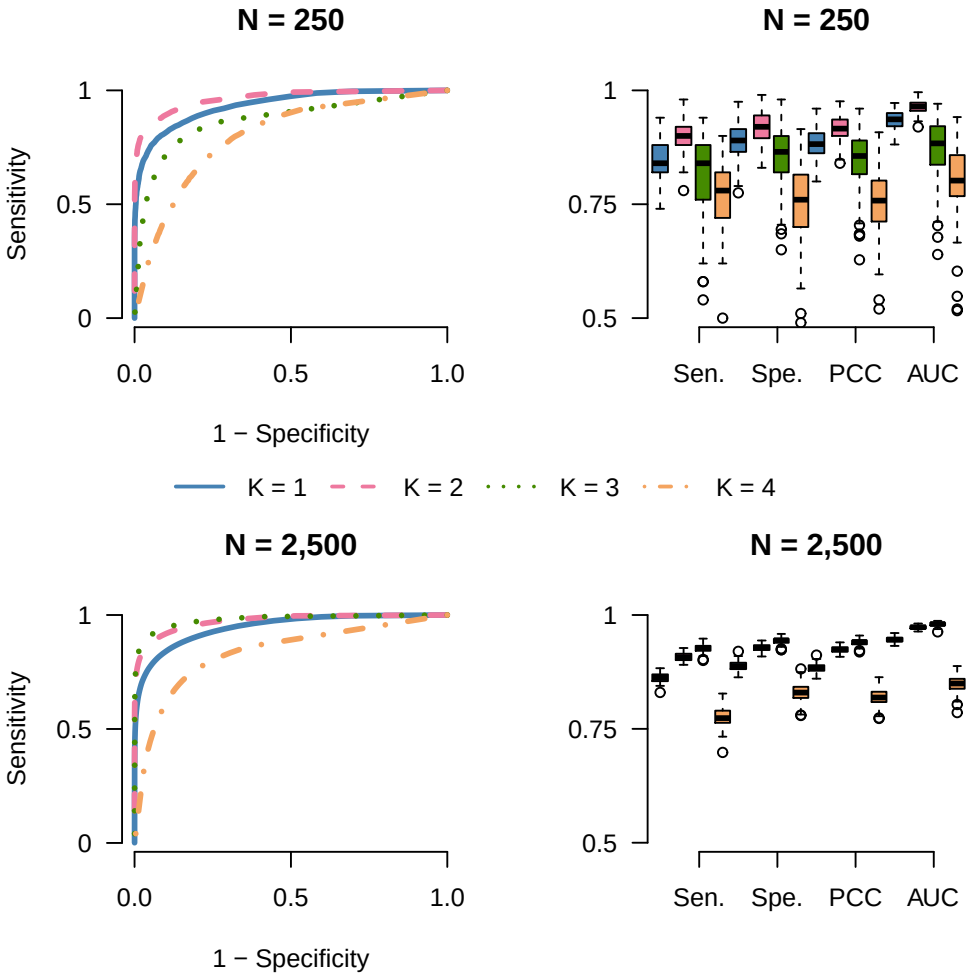

Figure 5.17: Receiver Operating Characteristic curves for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a 3-component mixture of normals with large departure from normality. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets.

size is increased to 2,500, the 3-component mixture model gave the best accuracy, reflecting the fact that more flexible models could capture the skewed nature of the T-distribution more accurately than a single normal distribution.

For small departure from normality (5 d.f.), the results for 250 patients show that the simple models such as $K = 1$ and $K = 2$ gave a good prediction, with the two components models most often being the best (see Table 5.11). Models with more flexible distributions are more accurate for prediction in the case the sample size is

Table 5.11: Prediction accuracy of the marginal approach under that assumption that the random effects follow a T-distribution with 3 and 5 degrees of freedom.

Size	Model	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
				3 degrees of freedom				
250	K = 1	0.14	0.73	0.74	0.74	0.77	0.46	0.92
	K = 2	0.16	0.76	0.77	0.77	0.81	0.46	0.93
	K = 3	0.17	0.73	0.75	0.74	0.79	0.44	0.92
	K = 4	0.23	0.70	0.70	0.70	0.74	0.39	0.90
2,500	K = 1	0.13	0.73	0.73	0.73	0.74	0.47	0.93
	K = 2	0.18	0.76	0.75	0.75	0.79	0.48	0.94
	K = 3	0.17	0.77	0.77	0.77	0.83	0.47	0.93
	K = 4	0.17	0.74	0.75	0.75	0.80	0.43	0.92
				5 degrees of freedom				
250	K = 1	0.16	0.81	0.81	0.81	0.86	0.53	0.95
	K = 2	0.15	0.81	0.82	0.82	0.87	0.53	0.95
	K = 3	0.22	0.78	0.78	0.78	0.83	0.48	0.93
	K = 4	0.24	0.77	0.76	0.76	0.81	0.46	0.93
2,500	K = 1	0.16	0.81	0.81	0.81	0.86	0.54	0.95
	K = 2	0.19	0.81	0.82	0.81	0.88	0.54	0.95
	K = 3	0.17	0.81	0.81	0.81	0.88	0.52	0.94
	K = 4	0.20	0.79	0.79	0.79	0.86	0.50	0.94

large (2,500 patients).

Together, these results show that when the departure from normality is small, assuming a single normal distribution can provide an accurate classification that is comparable with, or better than more flexible models. While in most cases, assuming a 2-components distribution works well, and there can be a benefit in assuming this extra flexible model.

However, this is not the whole story about the single normal distribution. There is a benefit of assuming a single distribution when the deviation from normality is substantial, and the location of mixture components is different between groups (this result is not stated here). This case allows a single distribution of the random effects to capture differences between the two groups, and gives a chance to classify patients correctly, despite the random effects misspecification.

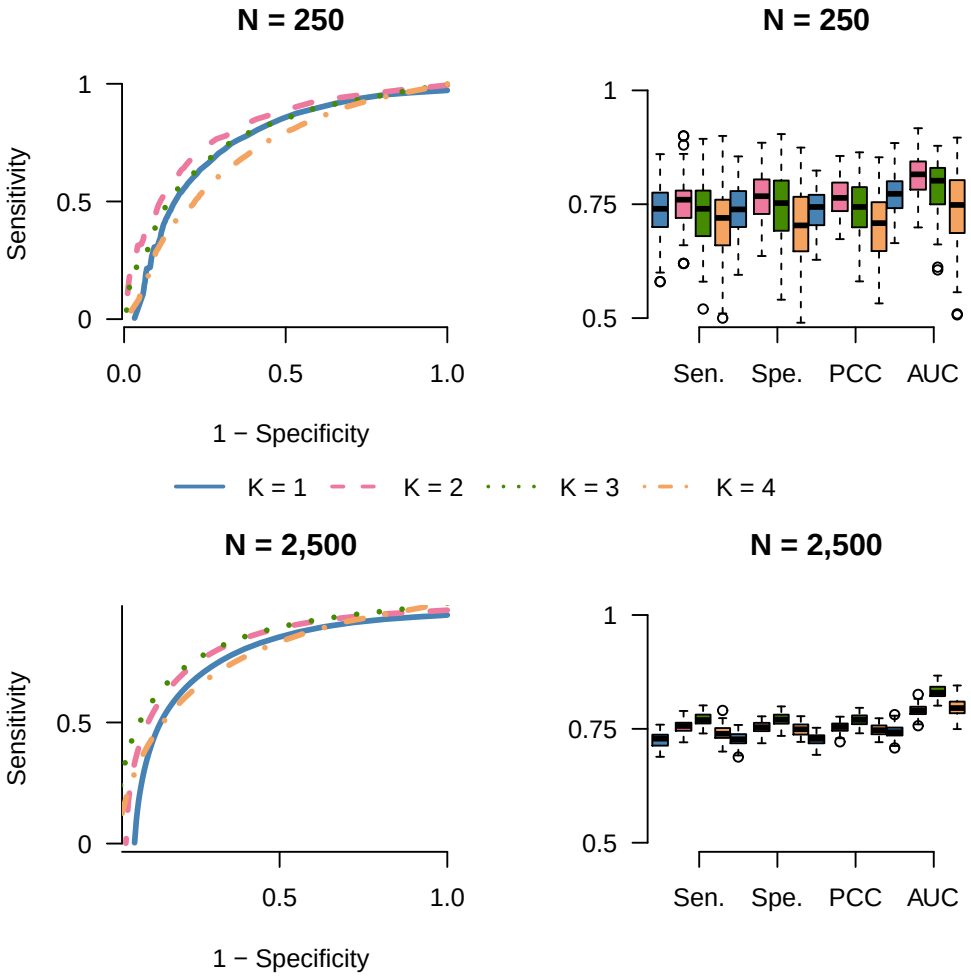

Figure 5.18: Receiver Operating Characteristic curves for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a T-distribution with 3 degrees of freedom. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets.

On the other hand, when the departure from normality increases, assuming a mixture distribution provides an improvement in classification accuracy, while assuming a single normal distribution performs less well, unless the difference is in the locations and levels of variations as well, in which case a single normal distribution can classify the patients well even though the model does not fit the data well. Some classification measurements such as AUC increase when a model with K = 2 is used, although this increase may not always be pronounced.

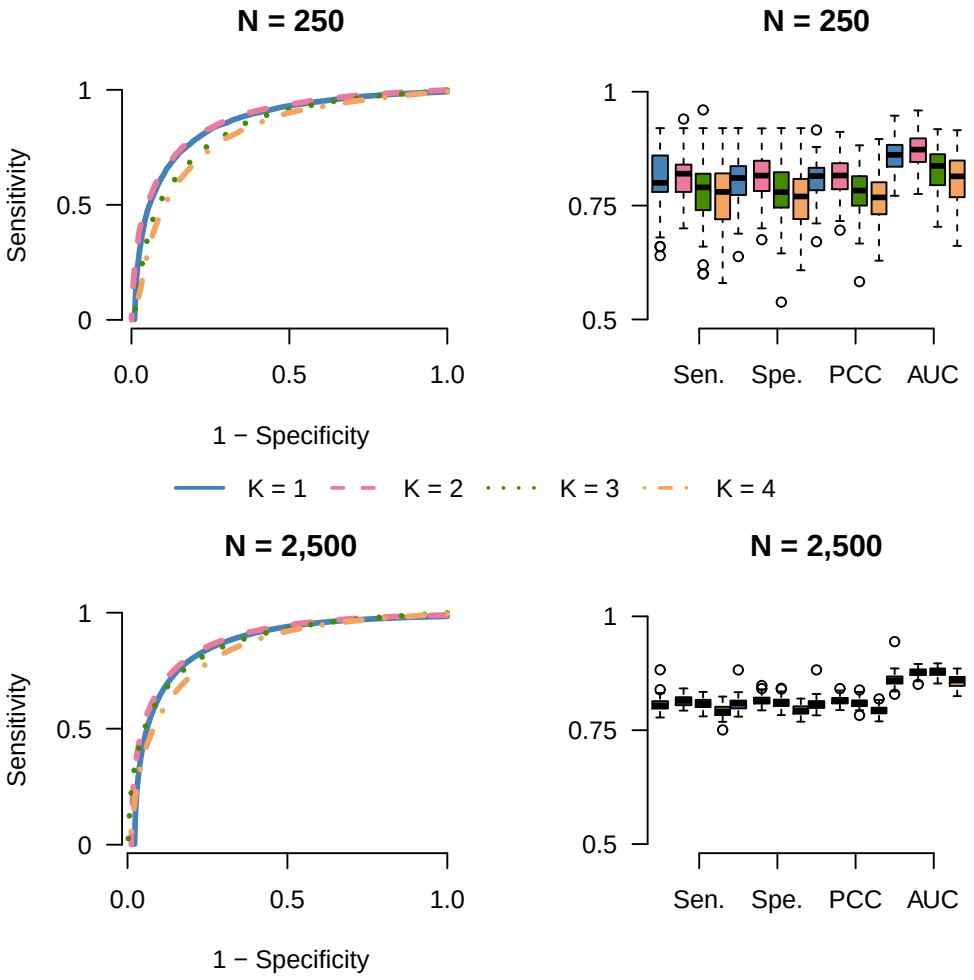

Figure 5.19: Receiver Operating Characteristic curves for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a T-distribution with 5 degrees of freedom. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets.

Effect on random effects prediction

Next, I turn to the random effects prediction approach. As indicated at the start of this section different studies have shown that the marginal and random effects prediction approaches to LoDA can each work well in different scenarios. For the PBC data considered in this thesis, the marginal approach gave the most accurate predictions. However, in this section, I explore factors that affect the accuracy of the random effects prediction approach in the presence of random effects misspecification. A key aspect of the random effect approach is the estimation of the patients' random effects. More clinic visits per patient may allow more accurate estimates of an individual patient's random effects which could in turn lead to more accurate predictions using the random effects approach.

Three simulation studies, a single normal distribution (K = 1), a 2-component mixture of normal distributions with large departure from normality (K = 2) and a T-distribution with 3 degrees of freedom, are considered to address this question. Four and nine visits per patients are simulated in this situation. The schedule of the 9 visits is designed so that each patient has approximately a visit every three months. The results obtained from the analysis of the random effects prediction approach can be compared in Tables 5.12, 5.13 and 5.14. Figures 5.20, 5.21 and 5.22 compare ROCs and box plots of the random effects approach when the number of visits per patients is 4 and 9 for the three scenarios.

The results of the scenario where the true distribution of the random effects is a single normal or 2-component mixture of normal distributions with large departure from normality, can be seen in Tables 5.12 and 5.13. The random effects prediction approach shows improvement in the classification accuracy when the number of visits per patients is 9 compares to 4 visits per patients. For example, in the case where the true distribution is a single normal, using more observations per patient helps to estimate the parameters of the random effects distribution more accurate (as shown in Table 5.12). From the data in Figure 5.20, there is much more variability in the specificity and AUC for the model K = 1 than for K = 2, while for sensitivity, it is the

Table 5.12: Prediction accuracy of the random-effect approach under the assumption that the random effects distribution follow a single normal distribution for N = 250 and 2,500 patients.

Size	visit	K	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
250	4	K = 1	0.86	0.84	0.68	0.71	0.78	0.47	0.94
		K = 2	0.74	0.68	0.64	0.65	0.64	0.35	0.88
		K = 3	0.30	0.49	0.58	0.56	0.50	0.25	0.81
		K = 4	0.097	0.37	0.70	0.63	0.51	0.25	0.81
	9	K = 1	0.73	0.85	0.75	0.77	0.83	0.55	0.95
		K = 2	0.77	0.80	0.76	0.77	0.80	0.49	0.93
		K = 3	0.43	0.52	0.63	0.61	0.55	0.28	0.84
		K = 4	0.21	0.40	0.71	0.65	0.53	0.28	0.83
2,500	4	K = 1	0.37	0.84	0.83	0.83	0.90	0.57	0.95
		K = 2	0.21	0.84	0.84	0.84	0.90	0.60	0.95
		K = 3	0.36	0.65	0.65	0.65	0.67	0.36	0.88
		K = 4	0.37	0.65	0.49	0.52	0.57	0.28	0.87

other way round. These factors may explain the apparent between the results in the Table 5.12 and Figure 5.20. In the case where the true distribution is K = 2, assuming K = 2 shows a noticeable improvement in the AUC of 0.09 between 4 and 9 visits per patients (see Table 5.13 and Figure 5.21). However, there is no significant improvement in using a mixture distribution rather than a single normal distribution K = 1.

In the next scenario, a T-distribution with 3 degrees of freedom for the random effects is considered. Table 5.14 and Figure 5.22 present the results of this scenario. Again, the more clinical visits a patient has, then the more accurately the random effects will be estimated. Furthermore, increasing the sample size to 2500 led to increased classification accuracy for all models, suggesting that more accurate estimates of the random effects could be made.

Table 5.13: Prediction accuracy of the random-effect approach under the assumption that the random effects distribution follow a 2-component normal distribution with large departure from normality for N = 250 and 2,500 patients.

Size	visit	K	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
N = 250	4	K = 1	0.45	0.66	0.74	0.73	0.67	0.46	0.89
		K = 2	0.15	0.46	0.82	0.74	0.62	0.48	0.86
		K = 3	0.074	0.56	0.73	0.70	0.65	0.37	0.87
		K = 4	0.065	0.50	0.70	0.66	0.60	0.33	0.85
	9	K = 1	0.24	0.75	0.79	0.78	0.79	0.53	0.92
		K = 2	0.29	0.61	0.82	0.77	0.71	0.50	0.89
		K = 3	0.22	0.60	0.75	0.72	0.67	0.41	0.88
		K = 4	0.19	0.55	0.66	0.64	0.59	0.32	0.85
N = 2,500	4	K = 1	0.011	0.71	0.87	0.84	0.83	0.61	0.92
		K = 2	0.070	0.69	0.93	0.88	0.84	0.76	0.93
		K = 3	0.026	0.43	0.75	0.69	0.60	0.68	0.85
		K = 4	0.083	0.30	0.86	0.75	0.58	0.55	0.83

Table 5.14: Prediction accuracy of the random-effect approach under the assumption that the random effects distribution follow a T-distribution with 3 degrees of freedom for N = 250 and 2,500 patients.

Size	visit	K	Cutoff	Sensitivity	Specificity	PCC	AUC	PPV	NPV
N = 250	4	K = 1	0.21	0.75	0.75	0.75	0.80	0.44	0.92
		K = 2	0.30	0.73	0.74	0.74	0.78	0.42	0.92
		K = 3	0.49	0.68	0.69	0.69	0.71	0.37	0.89
		K = 4	0.48	0.63	0.60	0.61	0.59	0.31	0.86
	9	K = 1	0.19	0.76	0.76	0.76	0.81	0.44	0.93
		K = 2	0.23	0.75	0.75	0.76	0.81	0.45	0.92
		K = 3	0.48	0.72	0.70	0.70	0.74	0.39	0.91
		K = 4	0.44	0.61	0.64	0.63	0.61	0.32	0.87
N = 2,500	4	K = 1	0.14	0.77	0.76	0.76	0.83	0.45	0.93
		K = 2	0.17	0.76	0.77	0.76	0.82	0.45	0.93
		K = 3	0.31	0.75	0.75	0.75	0.81	0.43	0.92
		K = 4	0.57	0.70	0.69	0.69	0.73	0.37	0.90

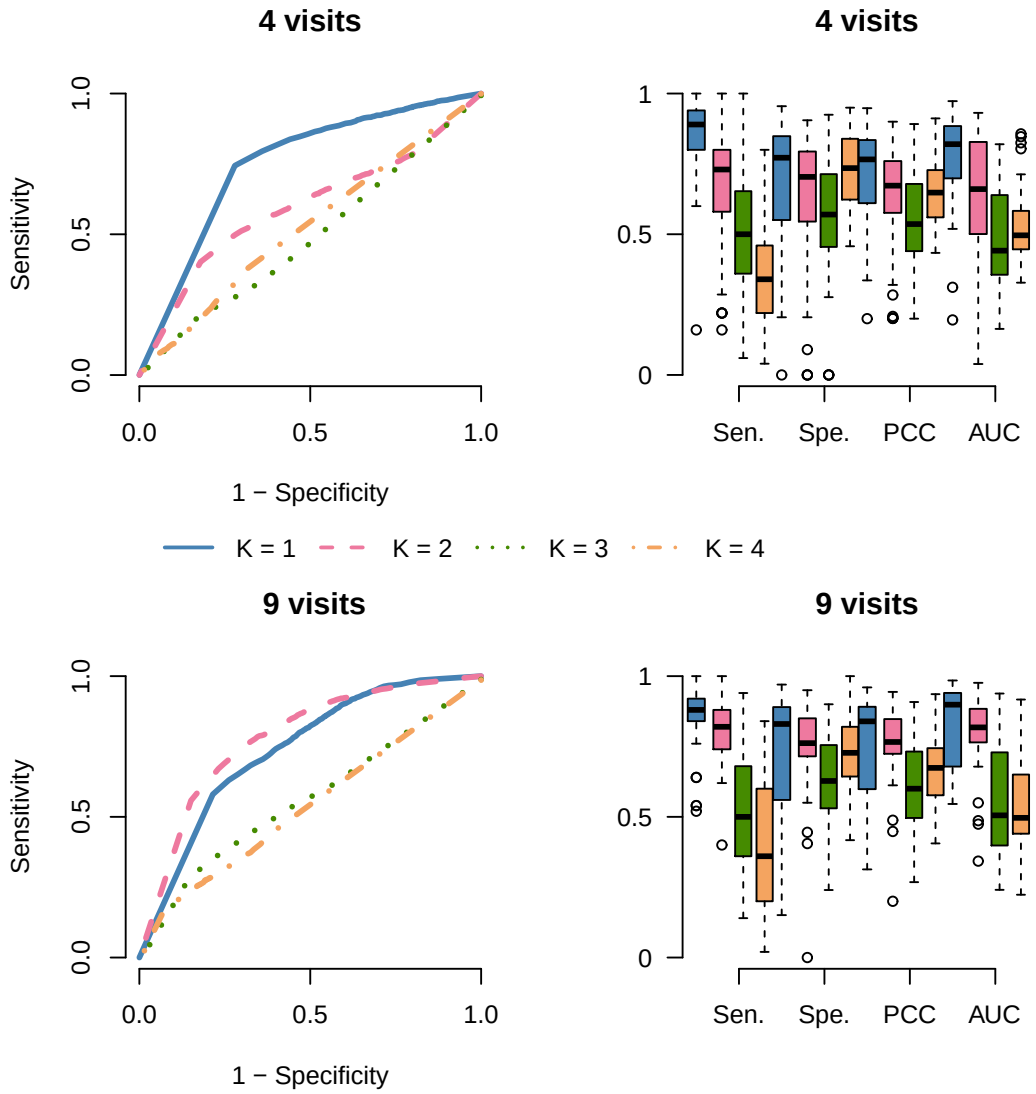

Figure 5.20: Receiver Operating Characteristic curves of the random effects approach for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a single normal. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets.

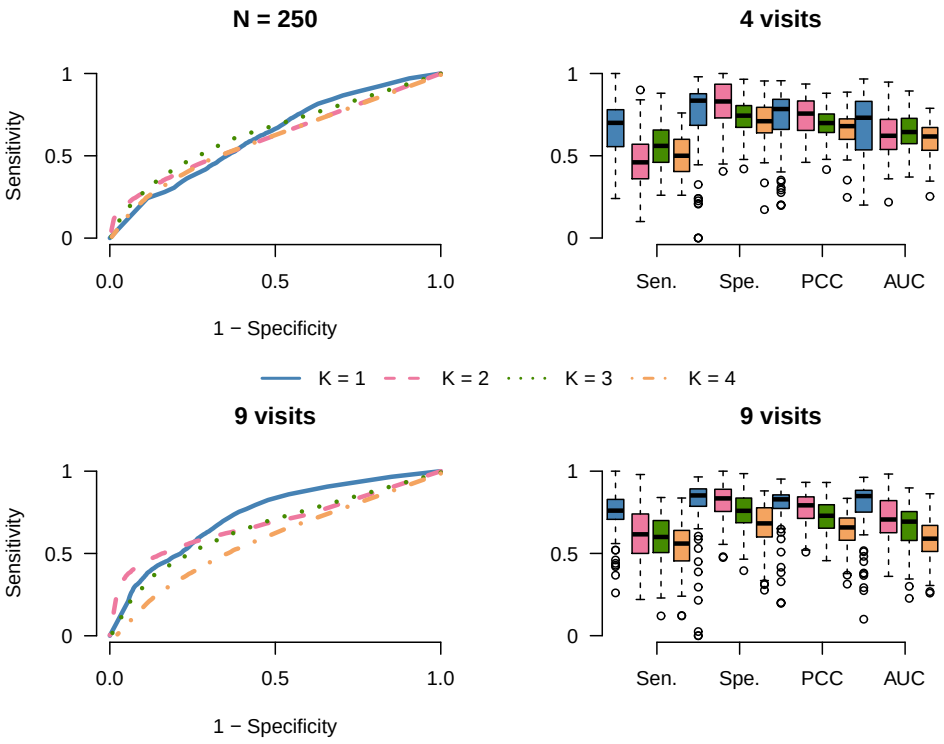

Figure 5.21: Receiver Operating Characteristic curves of the random effects approach for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a 2-components multivariate normal with a high degree of departure from normality. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets.

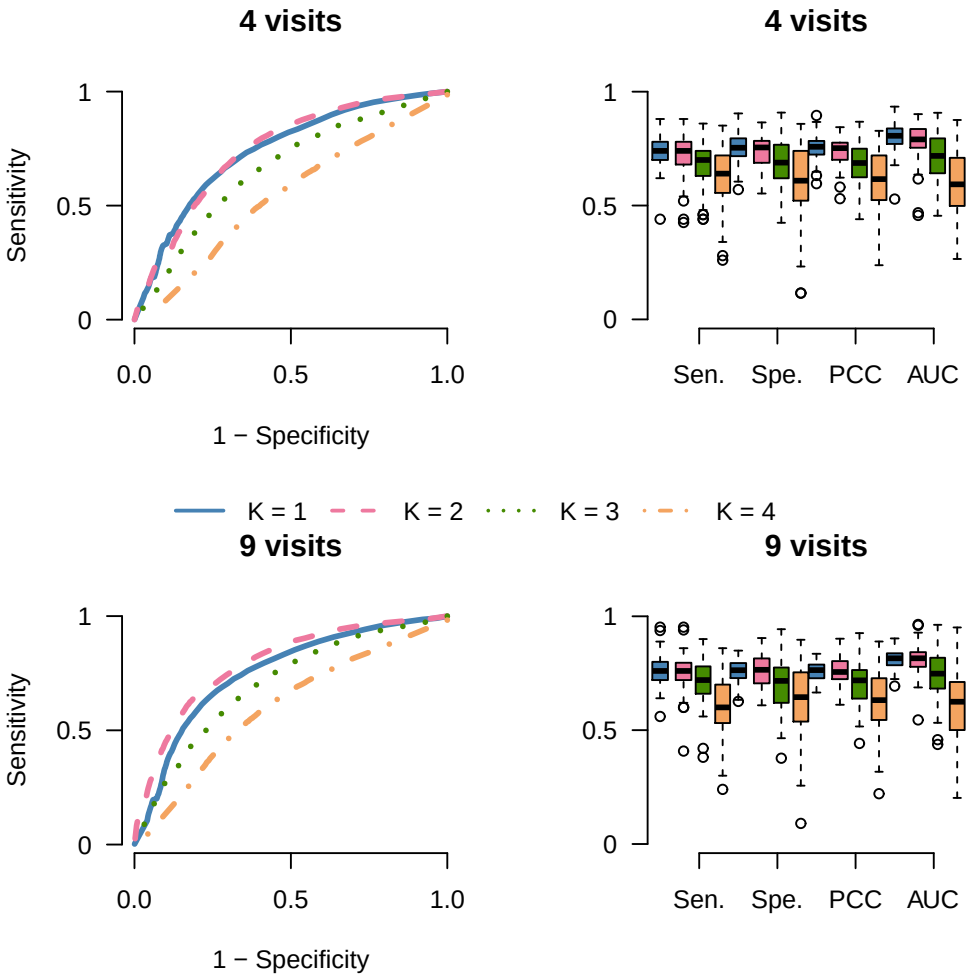

Figure 5.22: Receiver Operating Characteristic curves of the random effects approach for models with $K = 1, 2, 3, 4$ mixture components under the assumption that the true random effects distribution is a t-distribution with 3 degrees of freedom. The left panels show ROC curves for each model whilst the right panels show boxplots of the accuracy measures over 50 simulated datasets.

5.4 Summary

Together these simulation studies provide valuable insights into the effect of random effects misspecification on classification accuracy in longitudinal discriminant data. Two approaches of LoDA have been considered in this research, the marginal and the random effects approaches, since they have been shown to give the most accurate predictions in a number of settings (see Chapter 4, Hughes et al. (2018a), Komárek et al. (2010), Komarek et al. (2009) and Morrell et al. (2011)).

This research has shown that assuming a single normal distribution when the departure from normality is small, will not usually have much effect on the classification accuracy. When the departure from normality is substantial, the single normal distribution is unable to estimate the parameters accurately and the model performs less well. If the difference is in the locations and levels of variations as well, a single normal distribution can capture the differences between groups even though the estimates of the random effects are not estimated accurately.

On the other hand, in the case where the departure from normality is substantial, assuming a more flexible random effects distribution can allow more information to estimate the parameters correctly and that models can perform more accurate classification. This conclusion is in agreement with Komárek et al. (2010) who suggested that the classification accuracy will be improved when using a normal mixture in the random effects distribution.

It is also shown that increasing the sample size helps to improve classification accuracy, by allowing a more accurate estimation of the random effects. Similarly, increasing the number of observations per patient allows estimating the patient-specific intercepts and slopes more accurate. However if more flexible models, with mixtures of normal distributions, are assumed for the random effects, then researchers should take care to ensure that they have a suitable sample size, and a reasonable number of repeated measurements per patient to guarantee that the more complex models are accurately estimated.

Chapter 6

Conclusions and Further Work

6.1 Introduction

Discriminant analysis approaches are often used to classify patients into groups based on their risk of having a disease of interest. Linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) are the two principal traditional techniques of discriminant analysis. These techniques are typically designed to analyse cross-sectional data where data are collected from patients at one point in time. However, many clinical studies follow patients over prolonged periods of time which yields longitudinal data. In studies involving longitudinal data, discriminant analysis may be used to analyse the longitudinal information with the purpose of classifying the patients based on a predicted future outcome.

The increase in the amount of longitudinally collected clinical data has emphasised the need for longitudinal discriminant analysis (LoDA) approaches for classification purposes (Roy and Khattree (2005), Kim and Kong (2016*a*)). LoDA methods can be used to classify patients into prognostic groups based on mixed models that make use of the patient's longitudinal data. Some clinical studies that apply LoDA have considered only a single continuous marker (Brant et al. (2003), Kohlmann et al. (2009)). Since in many cases more than one longitudinal marker is measured, using information from

multiple markers for classification is often of interest. LoDA methods were extended to use multiple continuous markers (see for example Komárek et al. (2010)). A further extra extension of LoDA methods for multiple longitudinal markers considered different types of marker, such as continuous, discrete and binary (Fieuws et al. (2008), Hughes et al. (2017)). Hughes et al. (2018b) have recently developed a more flexible approach for LoDA (i.e., marginal, conditional and random effects approaches).

In this thesis, there were three main objectives as stated in Chapter 1. The first objective was to investigate the benefits of using LoDA rather than classical linear and quadratic discriminant analysis for clinical classification. The second objective was to explore the classification accuracy of three LoDA approaches (namely: marginal, conditional, random-effects predictions). The third objective was to assess the impact of the misspecification of the random-effects distribution on the classification accuracy.

In Chapter 2, the methodology associated with multivariate discriminant analysis using longitudinal data are described. In Chapter 3, a range of approaches for discriminant analysis was proposed and used to analyse a longitudinal ophthalmic dataset in order to predict treatment success or treatment failure in patients treated for age-related macular degeneration (AMD).

In Chapter 4, I further explored the accuracy of three ways of performing LoDA for the purpose of classification (i.e., marginal, conditional and random effects approaches) using the PBC dataset and simulation studies. Each of the three approaches has a different way to calculate a patient's posterior group membership probabilities. These three approaches were based on a mixed model using the patient's longitudinal history to predict new patients future disease status. An investigation of whether the misspecification of the random effects distribution has a minor or major effect on the classification performance was shown in Chapter 5.

A summary of the main findings and of the principal issues and suggestions which have arisen in this thesis is provided in this chapter.

6.2 Summary of the main findings

6.2.1 Classical discriminant analysis versus modified discriminant approach

LDA and QDA are two well-known approaches of discriminant analysis. LDA provides accurate classification results under the assumption of multivariate normality of the explanatory variables with common covariance matrix across groups. QDA is suggested when the assumption of covariance matrices being the same across groups is not met. It also requires larger sample sizes than LDA because QDA has to estimate an extra covariance matrix.

The key feature of longitudinal data is that repeated measurements of a marker within the same individual tend to be correlated. This correlation should be taken into account in the statistical analysis (Van Montfort et al. (2010)). Longitudinal clinical data are rarely measured at the same time point, and there may be missing data if a patient misses a clinical appointment. If I apply LDA or QDA to such data, this will require the exclusion of patients with missing values or imputed missing data. In addition, classical discriminant analysis (LDA or QDA) does not meet the main purpose of analysis of the longitudinal data in the sense that it deals with each time point as a single separate variable and does not model the correlation between repeated measurements.

A discriminant analysis approach based on a mixed-effects model can be used to overcome the limitations related to missing values or when patients do not arrive at the same time points.

In this thesis, the classical quadratic discriminant approach and a modified quadratic discriminant approach were compared using a longitudinal clinical dataset from ophthalmology to predict the patient's status at 12 months after treatment. Classical discriminant analysis can be applied to balanced longitudinal data and assumes that each follow-up measurement is a separate variable. The parameters (i.e., covariance

matrices and mean vectors) of the classical discriminant analysis were calculated from the multivariate data. While the modified discriminant analysis is based on the mixed-effects model and can be applied to unbalanced longitudinal datasets.

The main finding from this comparison was that utilising the discrimination methods that take into account the correlation between the repeated measurements within the same individuals does provide more accurate prediction when compared to the approach that uses each time point as a single variable. Furthermore, the use of mixed-effects models allows more effective use of the data since all the patients with missing visits (or visits that occurred at different time points across patients) could be included in the analysis, without the need to use imputation methods.

6.2.2 Longitudinal discriminant analysis (LoDA) approaches

Three approaches of longitudinal discriminant analysis (LoDA) (i.e., marginal, conditional and random effects approaches) have been proposed first by Morrell et al. (2007). More recently, approaches for longitudinal discriminant analysis (LoDA) have been further developed by considering a normal mixture for the random-effects distribution.

In chapter 4, I explored the advantages of each approach using the Primary Biliary Cirrhosis dataset (PBC dataset) and computer-simulated data. The marginal prediction focuses on the average change over time of the markers in each group. The conditional prediction is based on the patient-specific change of markers over time, without taking the error in the variability in the estimation of the patient's random effects into account. The conditional approach estimates conditional profiles of a new patient, given an estimate of their patient-specific deviations from the average longitudinal profile, then compares it with the overall mean longitudinal profiles of patient's with similar estimated random effects in each group. Finally, the random-effects prediction focuses on the patient-specific changes of markers in each group.

When the main differences between the prognostic groups were in the mean longitudinal profiles, the marginal approach was found to be the most accurate to clas-

sify patients the most accurately. This was expected since as mentioned above, the marginal approach focuses on the overall mean changes of the markers over time. However, when the number of repeated measurements increased per patient, the random effects approach was found to provide the best classification results; possibly because the additional measurements allowed a more accurate estimation of the random effects. When the patient-specific variability across the groups was noticeable, the random effects approach was expected to provide better prediction results than the marginal and conditional approaches. This result may be explained by the fact that the random effects approach focuses on the patient-specific variations of longitudinal profiles of markers over time, and so it can better capture this patient-specific deviation from the mean profile in each group.

For the simulation study, I considered two simulation scenarios to investigate which approach provides the best classification accuracy. In Scenario 1, the fixed effects parameters and the means of the random effects remained as in the PBC dataset in each group. However, the random-effects variance-covariance matrix was set to be the same in each group, in contrast with what was observed with the PBC dataset. This means that, in my simulation, the variations between the groups were due to differences in the mean profiles. For Scenario 1, I expected that the marginal prediction method was likely to outperform the other two classification methods. In the second scenario, I explored a situation where the conditional approach was expected to give the most accurate prediction. Morrell et al. (2011) stated that if the variance of residual error was large compared to the random effects variance, the random effects will be shrunken towards the mean for the group, making prediction into the most prevalent group more likely. The conditional approach utilises information about both the random effects and the residual errors, and so I hypothesised that this might lead to better predictions.

The findings of the first scenario study suggested that if the mean longitudinal profiles between the two groups capture the main differences, then the marginal method is able to classify patients most accurately. The findings observed in the second simulation study were not what I expected. I was unable to provide a scenario where the

conditional approach would outperform the marginal and random-effects approaches. The conditional approach appears to offer no improvement to these two approaches.

6.2.3 Impact of the misspecification of the random-effects distribution on the classification accuracy

Over the past decades, there has been a dramatic increase in statistical models that use random effects terms to analyse longitudinal data (Hansen et al. (2010), Kohlmann et al. (2009), Morrell et al. (2011)). Random-effects models take the correlation between measurements on the same patient into account. It is commonly assumed that the distribution of the random effects is a normal distribution.

The influence of the misspecification of the random effects distribution has been widely studied (e.g., Neuhaus et al. (1992), Verbeke and Lesaffre (1997)). However, most of the studies have focused on the accuracy of the parameters of the model. In chapter 5, I investigated whether the misspecification of the random-effects distribution has an effect on classification accuracy using the PBC dataset and simulated data. I used the PBC data to provide an example of how a mixture of random effects distributions influenced the classification accuracy, and used simulations to test the robustness of LoDA to different choices of random-effects distribution. In particular, I set four different distributions for random effects: multivariate normal distribution (K = 1), a mixture of two multivariate Gaussian distributions (K = 2), a mixture of three multivariate Gaussian distributions (K = 3) and the T distribution with 3 and 5 degrees of freedom for this investigation.

The results of this investigation showed that if the departure from normality is only small, assuming a single normal distribution will not affect the classification accuracy much. A possible explanation for this result may be that even though the random effects distribution is misspecified, it still captures sufficient information about the location and spread of the random effects to be useful for classification. On the other hand, in the situation where there is a substantial departure from normality, employing a more

flexible random effects distribution is recommended, because the more flexible models incorporate more information in the classification procedure.

A second finding was that the use of more flexible models requires a suitably large number of repeated measurements per patient and a reasonably large number of patients, in order to estimate the parameters of the random effects distribution accurately and to be able to improve classification accuracy.

6.3 Recommendations for practice

In this thesis, a range of approaches for discriminant analysis have been applied to several longitudinal datasets. The recommendations in relation to each study have been discussed in detail in the related chapters. This section summarises these recommendations.

My first recommendation is that LDA or QDA can be used to analyse longitudinal data for classification by fitting first a linear mixed-effects model. The estimated parameters (i.e., means and covariance matrices) can be used in LDA/QDA to classify patients into groups (a finding supported by others, e.g., Tomasko et al. (1999), Marshall and Barón (2000)). The mixed-effects model allows modelling explicitly the correlations between measurements from the same patients in the analysis. However, to decide which discriminant analysis approach (LDA/QDA) should be used, some statistical tests can be used to test the homogeneity of variance-covariance matrices (Box (1949)).

The second recommendation is that using LoDA approaches, such as marginal, conditional and random-effects approaches, (Morrell et al. (2007)), may be desirable where in some cases testing the homogeneity of variance-covariance matrices is unhelpful or unneeded. In addition, before analysing a data set, it is advisable to first plot the longitudinal profiles of their markers for patients in each group. If the differences in the group mean profiles are noticeable and if the variability between- and within-patients

are similar then the marginal approach is expected to give the best prediction. If, in addition, there is a difference in the level of variability between the group mean in each group, then the marginal and conditional approaches are not expected to give the best prediction. In such a case, the random effects approach is likely to provide an accurate classification, although care must be taken to ensure there are sufficient numbers of both patients and repeated measurements to be able to estimate the random effects accurately. However, if a considerable measurement error controls the variability between patients, then the random-effects approach should not be an option because the individual random effects estimates are incorrect. In such a case, the marginal approach would be a suitable option.

Finally, for small datasets, assuming a single multivariate distribution for the random-effects is acceptable. For large datasets, considering a more complex, flexible distribution for the random effects is expected to be a good option to guard against the misspecification of random effects.

6.4 Future work

The work presented in this thesis focussed on the longitudinal discriminant analysis (LoDA) approaches that use patients' longitudinal data to predict their future clinical status. In this section, I will demonstrate some future work related to the results and the analyses of LoDA approaches.

The impact that the sample size and the number of repeated measurements have on the classification accuracy of each of the three LoDA approaches is still an open question.

It would also be interesting to investigate whether a combined approach that uses both marginal and random effects predictions can improve the classification accuracy in terms of lead-time and sensitivity/specificity (Morrell et al. (2011)). One possible way of combining them could be by using Bayesian model averaging (Hoeting et al.

(1999)). For example, Rizopoulos et al. (2014) combined dynamic predictions from joint models for longitudinal and time-to-event data. Their idea is based on weighting the predictions from a number of models. The weights are determined using individual patient data allowing weights that are both patient specific and time-dependent.

The applications in this thesis focus on three different types of markers: continuous, binary and discrete. Another methodological development of longitudinal discriminant analysis would be to extend the approach to allow categorical longitudinal markers. This could be achieved through the use of multinomial logistic models within the mixed model framework. One of the difficulties is how to estimate the random effects parameters that are assumed to follow a mixture of normal distributions. Using the maximum likelihood method to estimate the parameters will not be easy. Two possible methods to obtain these estimates include the use of pairwise fitting methods joined with the EM algorithm (see e.g., Laffont et al. (2014)) or MCMC methods such as the Metropolis-Hastings algorithm along with numerical integration methods such as Gauss Quadrature could be another way to estimate the parameters. The challenge in this development would be the inclusion of mixture components in the random effects distribution, or even more flexible models.

Missing data are one of the most common problems in longitudinal studies. Longitudinal studies may show missing values since marker measurements may be incomplete for some subjects. It would be of interest to investigate how missing data affects the classification accuracy, in a more complete extension of the initial simulation work developed in Chapter 3. There are three forms of missing data: missing completely at random (MCAR), missing at random (MAR) and not missing at random (NMAR). There are three principal approaches to handle missing data: (1) Imputation, where missing values are replaced, (2) omission where missing values are dropped and (3) analysis by applying methods that can deal with the missing values. The last observation carried forward is one of the imputation method I have used to impute the missing values. However, there are other different approaches to deal with missing data such as listwise and pairwise deletion; mean imputation; regression imputation; stochastic

imputation; and multiple imputations (see e.g.,Engels and Diehr (2003)).

An additional area of interest is when the study involves a large number of markers compared to the number of subjects, and which is often referred as high-dimensional data. One of the main questions that researchers face is how to find the best marker or set of markers that can discriminate between groups. Within the discriminant analysis framework, variable selection methods could be considered as a first step to reduce the dimensionality in the dataset before applying discriminant analysis. Alternatively, variable selection algorithms could be embedded within the longitudinal discriminant analysis (Nkiet (2012)).

6.5 Conclusions

The demand for the longitudinal multivariate methods is increasing due to a large amount of longitudinal data currently available in healthcare databases. Especially, there has been an increased interest in LoDA approaches that use a patient's clinical history for prediction.

This study has shown that using LoDA approaches to analyse longitudinal data for classification purpose can provide accurate classification results. In addition, considering more complex, flexible distributions for the random effects can provide a robust classification even if the random effects are misspecified.

Bibliography

Agresti, A., Caffo, B. and Ohman-Strickland, P. (2004), 'Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies', *Computational Statistics & Data Analysis* **47**(3), 639–653.

Alexakos, C. (1966), 'Predictive efficiency of two multivariate statistical techniques in comparison with clinical predictions.', *Journal of Educational Psychology* **57**(5), 297.

Arribas-Gil, A., De la Cruz, R., Lebarbier, E. and Meza, C. (2015), 'Classification of longitudinal data through a semiparametric mixed-effects model based on lasso-type estimators', *Biometrics* **71**(2), 333–343.

Bandyopadhyay, S., Ganguli, B. and Chatterjee, A. (2011), 'A review of multivariate longitudinal data analysis', *Statistical methods in medical research* **20**(4), 299–330.

Bartlett, M. S. (1937), 'Properties of sufficiency and statistical tests', *Proc. R. Soc. Lond. A* **160**(901), 268–282.

Bellmann, C., Unnebrink, K., Rubin, G. S., Miller, D. and Holz, F. G. (2003), 'Visual acuity and contrast sensitivity in patients with neovascular age-related macular degeneration', *Graefe's archive for clinical and experimental ophthalmology* **241**(12), 968–974.

Bouveyron, C., Girard, S. and Schmid, C. (2007), 'High-dimensional discriminant analysis', *Communications in Statistics—Theory and Methods* **36**(14), 2607–2623.

Box, G. E. (1949), 'A general distribution theory for a class of likelihood criteria', *Biometrika* **36**(3/4), 317–346.

Brant, L. J., Sheng, S. L., Morrell, C. H., Verbeke, G. N., Lesaffre, E. and Carter, H. B. (2003), 'Screening for prostate cancer by using random-effects models', *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **166**(1), 51–62.

Brant, L. J., Sheng, S. L., Morrell, C. H. and Zonderman, A. B. (2005), 'Data from a longitudinal study provided measurements of cognition to screen for alzheimer's disease', *Journal of clinical epidemiology* **58**(7), 701–707.

Brown, M. B. and Forsythe, A. B. (1974), 'Robust tests for the equality of variances', *Journal of the American Statistical Association* **69**(346), 364–367.

Brown, P. J., Kenward, M. G. and Bassett, E. E. (2001), 'Bayesian discrimination with longitudinal data', *Biostatistics* **2**(4), 417–432.

Bruckers, L., Molenberghs, G., Drinkenburg, P. and Geys, H. (2016), 'A clustering algorithm for multivariate longitudinal data', *Journal of biopharmaceutical statistics* **26**(4), 725–741.

Cochran, W. G., Bliss, C. I. et al. (1948), 'Discriminant functions with covariance', *The Annals of Mathematical Statistics* **19**(2), 151–176.

Coster, L. d., Leentjens, A. F., Lodder, J. and Verhey, F. R. (2005), 'The sensitivity of somatic symptoms in post-stroke depression: a discriminant analytic approach', *International journal of geriatric psychiatry* **20**(4), 358–362.

De La Cruz-Mesia, R. and Quintana, F. A. (2006), 'A model-based approach to bayesian classification with applications to predicting pregnancy outcomes from longitudinal β-hcg profiles', *Biostatistics* **8**(2), 228–238.

De la Cruz, R., Fuentes, C., Meza, C., Lee, D.-J. and Arribas-Gil, A. (2017), 'Predicting pregnancy outcomes using longitudinal information: a penalized splines mixed-effects model approach', *Statistics in medicine* **36**(13), 2120–2134.

De la Cruz, R., Fuentes, C., Meza, C. and Núñez-Antón, V. (2018), 'Error-rate estimation in discriminant analysis of non-linear longitudinal data: A comparison of resampling methods', *Statistical methods in medical research* **27**(4), 1153–1167.

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D. and Langworthy, A. (1989), 'Prognosis in primary biliary cirrhosis: model for decision making', *Hepatology* **10**(1), 1–7.

Diggle, P., Diggle, P. J., Heagerty, P., Heagerty, P. J., Liang, K.-Y., Zeger, S. et al. (2002), *Analysis of longitudinal data*, Oxford University Press.

El Saeiti, R., García-Fiñana, M. and Hughes, D. M. (2019), The effect of random-effects

misspecification on classification accuracy. submitted.

Engels, J. M. and Diehr, P. (2003), 'Imputation of missing longitudinal data: a comparison of methods', *Journal of clinical epidemiology* **56**(10), 968–976.

Fawcett, T. (2006), 'An introduction to roc analysis', *Pattern recognition letters* **27**(8), 861–874.

Ferris III, F. L., Wilkinson, C., Bird, A., Chakravarthy, U., Chew, E., Csaky, K., Sadda, S. R., for Macular Research Classification Committee, B. I. et al. (2013), 'Clinical classification of age-related macular degeneration', *Ophthalmology* **120**(4), 844–851.

Fieuws, S. and Verbeke, G. (2006), 'Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles', *Biometrics* **62**(2), 424–431.

Fieuws, S., Verbeke, G., Maes, B. and Vanrenterghem, Y. (2008), 'Predicting renal graft failure using multivariate longitudinal profiles', *Biostatistics* **9**(3), 419–431.

Fisher, R. A. (1936), 'The use of multiple measurements in taxonomic problems', *Annals of eugenics* **7**(2), 179–188.

Fleming, T. R. and Harrington, D. P. (1991), *Counting processes and survival analysis*, Vol. 169, John Wiley & Sons.

Flury, B. W. and Schmid, M. J. (1992), 'Quadratic discriminant functions with constraints on the covariance matrices: Some asymptotic results', *Journal of multivariate analysis* **40**(2), 244–261.

Fox, J., Friendly, M. and Monette, G. (2018), *heplots: Visualizing Tests in Multivariate Linear Models*. R package version 1.3-5.

Freeman, E. A. and Moisen, G. G. (2008), 'A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa', *Ecological Modelling* **217**(1-2), 48–58.

Gamerman, D. (1997), 'Sampling from the posterior distribution in generalized linear mixed models', *Statistics and Computing* **7**(1), 57–68.

García-Fiñana, M., Hughes, D. M., Cheyne, C. P., Broadbent, D. M., Wang, A., Komárek, A., Stratton, I. M., Mobayen-Rahni, M., Alshukri, A., Vora, J. P. et al. (2019), 'Personalized risk-based screening for diabetic retinopathy: A multivariate approach versus the use of stratification rules', *Diabetes, Obesity and Metabolism*

21(3), 560–568.

García-Finana, M., Murjaneh, S., Mahmood, S. and Harding, S. (2010), 'Baseline clinical measures and early response predict success in verteporfin photodynamic therapy for neovascular age-related macular degeneration', *Eye* **24**(7), 1213.

Gastwirth, J. L., Gel, Y. R. and Miao, W. (2009), 'The impact of levene's test of equality of variances on statistical theory and practice', *Statistical Science* pp. 343–360.

Geisser, S. and Greenhouse, S. W. (1958), 'An extension of box's results on the use of the f distribution in multivariate analysis', *The Annals of Mathematical Statistics* **29**(3), 885–891.

Goodenough, D. J., Rossmann, K. and Lusted, L. B. (1974), 'Radiographic applications of receiver operating characteristic (roc) curves', *Radiology* **110**(1), 89–95.

Hair, J. F., Anderson Rolph, E., Tatham Ronald, L. and Black William, C. (1994), *Multivariate data analysis with readings*, Macmillan Publishing Company.

Hajian-Tilaki, K. (2013), 'Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation', *Caspian journal of internal medicine* **4**(2), 627.

Hand, D. J. (2009), 'Measuring classifier performance: a coherent alternative to the area under the roc curve', *Machine learning* **77**(1), 103–123.

Hand, D. J. (2012), 'Assessing the performance of classification methods', *International Statistical Review* **80**(3), 400–414.

Hand, D. J. (2017), *Practical longitudinal data analysis*, Routledge.

Hand, D. J. and Till, R. J. (2001), 'A simple generalisation of the area under the roc curve for multiple class classification problems', *Machine learning* **45**(2), 171–186.

Hanley, J. A. and McNeil, B. J. (1982), 'The meaning and use of the area under a receiver operating characteristic (roc) curve.', *Radiology* **143**(1), 29–36.

Hansen, B. E., Komárek, A., Buster, E. H., Steyerberg, E. W., Janssen, H. L. and Lesaffre, E. (2010), 'Dynamic prediction of response to hbv-treatment using multivariate longitudinal profiles', *Statistical Models of Treatment Effects in Chronic Hepatitis B and C* p. 79.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The elements of statistical learning: data mining, inference, and prediction*, New York, NY: Springer.

Heagerty, P. J. and Kurland, B. F. (2001), 'Misspecified maximum likelihood estimates and generalised linear mixed models', *Biometrika* **88**(4), 973–985.

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999), 'Bayesian model averaging: a tutorial', *Statistical science* pp. 382–401.

Huberty, C. J. and Curry, A. R. (1978), 'Linear versus quadratic multivariate classification', *Multivariate behavioral research* **13**(2), 237–245.

Huberty, C. J. and Olejnik, S. (2006), *Applied MANOVA and discriminant analysis*, Vol. 498, John Wiley & Sons.

Huberty, C. and Wisenbaker, J. (1992), 'Discriminant analysis: Potential improvements in typical practice', *Advances in social science methodology* **2**, 169–208.

Hughes, D. M., El Saeiti, R. and García-Fiñana, M. (2018a), 'A comparison of group prediction approaches in longitudinal discriminant analysis', *Biometrical Journal* **60**(2), 307–322.

Hughes, D. M., Komárek, A., Bonnett, L. J., Czanner, G. and García-Fiñana, M. (2017), 'Dynamic classification using credible intervals in longitudinal discriminant analysis', *Statistics in medicine* **36**(24), 3858–3874.

Hughes, D. M., Komárek, A., Czanner, G. and Garcia-Fiñana, M. (2018b), 'Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types', *Statistical methods in medical research* **27**(7), 2060–2080.

Jain, S. and Jain, R. (1994), 'Discriminant analysis and its application to medical data', *Biometrical journal* **36**(2), 147–151.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An introduction to statistical learning*, Vol. 112, Springer.

Joy, O. M. and Tollefson, J. O. (1975), 'On the financial applications of discriminant analysis', *Journal of Financial and Quantitative Analysis* **10**(5), 723–739.

Kim, Y. and Kong, L. (2016*a*), 'Classification using longitudinal trajectory of biomarker in the presence of detection limits', *Statistical methods in medical research* **25**(1), 458–471.

Kim, Y. and Kong, L. (2016*b*), 'Improving classification accuracy by combining longitudinal biomarker measurements subject to detection limits', *Statistics in Biophar-*

maceutical Research **8**(2), 171–178.

Kohlmann, M., Held, L. and Grunert, V. P. (2009), 'Classification of therapy resistance based on longitudinal biomarker profiles', *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **51**(4), 610–626.

Komárek, A., Hansen, B. E., Kuiper, E. M., van Buuren, H. R. and Lesaffre, E. (2010), 'Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution', *Statistics in medicine* **29**(30), 3267–3283.

Komarek, A., Hansen, B., Jansen, H. and Lesaffre, E. (2009), 'Prediction of binary response using multivariate longitudinal profiles: Study on chronic hepatitis b patients'.

Komárek, A. and Komárková, L. (2013), 'Supplement to "clustering for multivariate continuous and discrete longitudinal data."'.

Komárek, A. and Komárková, L. (2014), 'Capabilities of r package mixak for clustering based on multivariate continuous and discrete longitudinal data', *Journal of Statistical Software* **59**(1), 1–38.

Komárek, A., Komárková, L. et al. (2013), 'Clustering for multivariate continuous and discrete longitudinal data', *The Annals of Applied Statistics* **7**(1), 177–200.

Krzanowski, W. (2000), *Principles of multivariate analysis*, Vol. 23, OUP Oxford.

Kumar, R. and Indrayan, A. (2011), 'Receiver operating characteristic (roc) curve for medical researchers', *Indian pediatrics* **48**(4), 277–287.

Laffont, C. M., Vandemeulebroecke, M. and Concordet, D. (2014), 'Multivariate analysis of longitudinal ordinal data with mixed effects models, with application to clinical outcomes in osteoarthritis', *Journal of the American Statistical Association* **109**(507), 955–966.

Laird, N. M., Ware, J. H. et al. (1982), 'Random-effects models for longitudinal data', *Biometrics* **38**(4), 963–974.

Levene, H. (1960), 'Robust tests for equality of variances.'', *Contributions to probability and statistics: essay in honour of Harold Hotelling'.(Eds I Olkin, SG Ghurye, W Hoeffding, WG Madow, HB Mann)(Stanford University Press: London)* .

Levesque, L., Ducharme, F., Zarit, S. H., Lachance, L. and Giroux, F. (2008), 'Pre-

dicting longitudinal patterns of psychological distress in older husband caregivers: further analysis of existing data', *Aging and Mental Health* **12**(3), 333–342.

Li, H. and Gatsonis, C. (2019), 'Combining biomarker trajectories to improve diagnostic accuracy in prospective cohort studies with verification bias', *Statistics in medicine* **38**(11), 1968–1990.

Litière, S., Alonso, A. and Molenberghs, G. (2007), 'Type i and type ii error under random-effects misspecification in generalized linear mixed models', *Biometrics* **63**(4), 1038–1044.

Litière, S., Alonso, A. and Molenberghs, G. (2008), 'The impact of a misspecified random-effects distribution on the estimation and the performance of inferential procedures in generalized linear mixed models', *Statistics in medicine* **27**(16), 3125–3144.

Litière, S., Alonso, A. and Molenberghs, G. (2011), 'Rejoinder to "a note on type ii error under random effects misspecification in generalized linear mixed models"', *Biometrics* **67**(2), 656–660.

Lix, L. and Sajobi, T. (2010), 'Discriminant analysis for repeated measures data: a review', *Frontiers in psychology* **1**, 146.

Lukasiewicz, E., Gorfine, M., Neumann, A. U. and Freedman, L. S. (2011), 'Combining longitudinal discriminant analysis and partial area under the roc curve to predict non-response to treatment for hepatitis c virus', *Statistical methods in medical research* **20**(3), 275–289.

Lusted, L. B. (1971), 'Decision-making studies in patient management', *New England Journal of Medicine* **284**(8), 416–424.

Mardia, K., Kent, J. and Bibby, J. (1979), *Multivariate analysis*, Academic Press, London.

Marks, S. and Dunn, O. J. (1974), 'Discriminant functions when covariance matrices are unequal', *Journal of the American Statistical Association* **69**(346), 555–559.

Marshall, G. and Barón, A. E. (2000), 'Linear discriminant models for unbalanced longitudinal data', *Statistics in medicine* **19**(15), 1969–1981.

Marshall, G., De la Cruz-Mesía, R., Quintana, F. A. and Barón, A. E. (2009), 'Discriminant analysis for longitudinal data with multiple continuous responses and possibly

missing data', *Biometrics* **65**(1), 69–80.

McCulloch, C. E. and Neuhaus, J. M. (2011*a*), 'Misspecifying the shape of a random effects distribution: why getting it wrong may not matter', *Statistical science* pp. 388–402.

McCulloch, C. E. and Neuhaus, J. M. (2011*b*), 'Prediction of random effects in linear and generalized linear models under model misspecification', *Biometrics* **67**(1), 270–279.

McLachlan, G. (2004), *Discriminant analysis and statistical pattern recognition*, Vol. 544, John Wiley & Sons.

Meshbane, A. and Morris, J. D. (1995), 'A method for selecting between linear and quadratic classification models in discriminant analysis', *The Journal of experimental education* **63**(3), 263–273.

Metz, C. E. (1978), Basic principles of roc analysis, *in* 'Seminars in nuclear medicine', Vol. 8, Elsevier, pp. 283–298.

Michaelis, J. (1973), Simulation experiments with multiple group linear and quadratic discriminant analysis, *in* 'Discriminant analysis and applications', Elsevier, pp. 225–238.

Michie, D., Spiegelhalter, D. J. and Taylor, C. C. (1994), 'Machine learning, neural and statistical classification', *Ellis Horwood Series in Artificial Intelligence, New York, NY: Ellis Horwood,— c1994, edited by Michie, Donald; Spiegelhalter, David J.; Taylor, Charles C.* .

Minassian, D. C., Reidy, A., Lightstone, A. and Desai, P. (2011), 'Modelling the prevalence of age-related macular degeneration (2010–2020) in the uk: expected impact of anti-vascular endothelial growth factor (vegf) therapy', *British Journal of Ophthalmology* pp. bjo–2010.

Morrell, C. H., Brant, L. J. and Sheng, S. (2007), 'Comparing approaches for predicting prostate cancer from longitudinal data', *Proceedings of the American Statistical Association, Alexandria, American Statistica Association* .

Morrell, C. H., Brant, L. J., Sheng, S. and Metter, E. J. (2005), 'Using multivariate mixed-effects models to predict prostate cancer', *Proceedings of the American Sta-*

tistical Association, Biometrics Section Alexandria, American Statistical Association pp. 332–337.

Morrell, C. H., Brant, L. J., Sheng, S. and Metter, E. J. (2012), 'Screening for prostate cancer using multivariate mixed-effects models', *Journal of applied statistics* **39**(6), 1151–1175.

Morrell, C. H., Sheng, S. L. and Brant, L. J. (2011), 'A comparative study of approaches for predicting prostate cancer from longitudinal data', *Communications in Statistics-Simulation and Computation* **40**(9), 1494–1513.

Murtaugh, P. A., Dickson, E. R., Van Dam, G. M., Malinchoc, M., Grambsch, P. M., Langworthy, A. L. and Gips, C. H. (1994), 'Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits', *Hepatology* **20**(1), 126–134.

Neuhaus, J. M., Hauck, W. W. and Kalbfleisch, J. D. (1992), 'The effects of mixture distribution misspecification when fitting mixed-effects logistic models', *Biometrika* **79**(4), 755–762.

Neuhaus, J. M., McCulloch, C. E. and Boylan, R. (2011), 'A note on type ii error under random effects misspecification in generalized linear mixed models', *Biometrics* **67**(2), 654–656.

Nkiet, G. M. (2012), 'Direct variable selection for discrimination among several groups', *Journal of Multivariate Analysis* **105**(1), 151–163.

Pinheiro, J., Bates, D., DebRoy, S. and Sarkar, D. (2014), 'R core team (2014) nlme: linear and nonlinear mixed effects models. r package version 3.1-117', *Available at h ttp://CRAN. R-project. org/package= nlme* .

Pitman, E. (1939), 'Tests of hypotheses concerning location and scale parameters', *Biometrika* **31**(1/2), 200–215.

Plummer, M. (2008), 'Penalized loss functions for bayesian model comparison', *Biostatistics* **9**(3), 523–539.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006), 'Coda: convergence diagnosis and output analysis for mcmc', *R news* **6**(1), 7–11.

Puza, B. (2015), *Bayesian Methods for Statistical Analysis*, ANU Press.

R Core, T. (2017), 'R: A language and environment for statistical computing. vienna,

austria: R foundation for statistical computing; 2016'.

Reddy, T., Molenberghs, G., Njagi, E. N. and Aerts, M. (2016), 'A novel approach to estimation of the time to biomarker threshold: applications to hiv', *Pharmaceutical statistics* **15**(6), 541–549.

Reinsel, G. (1984), 'Estimation and prediction in a multivariate random effects generalized linear model', *Journal of the American Statistical Association* **79**(386), 406–414.

Rencher, A. C. (1998), *Multivariate statistical inference and applications*, Vol. 338, Wiley-Interscience.

Richardson, S. and Green, P. J. (1997), 'On bayesian analysis of mixtures with an unknown number of components (with discussion)', *Journal of the Royal Statistical Society: series B (statistical methodology)* **59**(4), 731–792.

Rietveld, M., van Der Valk, J., Bongers, I., Stroet, T., Slagboom, P. and Boomsma, D. (2000), 'Zygosity diagnosis in young twins by parental report', *Twin Research and Human Genetics* **3**(3), 134–141.

Rizopoulos, D. (2012), *Joint models for longitudinal and time-to-event data: With applications in R*, Chapman and Hall/CRC.

Rizopoulos, D., Hatfield, L. A., Carlin, B. P. and Takkenberg, J. J. (2014), 'Combining dynamic predictions from joint models for longitudinal and time-to-event data using bayesian model averaging', *Journal of the American Statistical Association* **109**(508), 1385–1397.

Roy, A. (2006), 'A new classification rule for incomplete doubly multivariate data using mixed effects model with performance comparisons on the imputed data', *Statistics in medicine* **25**(10), 1715–1728.

Roy, A. and Khattree, R. (2005), 'On discrimination and classification with multivariate repeated measures data', *Journal of Statistical Planning and Inference* **134**(2), 462–485.

Rubin, M. L., Chan, W., Yamal, J.-M. and Robertson, C. S. (2017), 'A joint logistic regression and covariate-adjusted continuous-time markov chain model', *Statistics in medicine* **36**(28), 4570–4582.

Shah, A., Laird, N. and Schoenfeld, D. (1997), 'A random-effects model for multi-

ple characteristics with possibly missing data', *Journal of the American Statistical Association* **92**(438), 775–779.

Tang, W. and Tu, X. (2012), *Modern clinical trial analysis*, Springer.

Tomasko, L., Helms, R. W. and Snapinn, S. M. (1999), 'A discriminant analysis extension to mixed models', *Statistics in medicine* **18**(10), 1249–1260.

Van Montfort, K., Oud, J. H. and Satorra, A. (2010), *Longitudinal research with latent variables*, Springer Science & Business Media.

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

Verbeke, G., Fieuws, S., Molenberghs, G. and Davidian, M. (2014), 'The analysis of multivariate longitudinal data: A review', *Statistical methods in medical research* **23**(1), 42–59.

Verbeke, G. and Lesaffre, E. (1996), 'A linear mixed-effects model with heterogeneity in the random-effects population', *Journal of the American Statistical Association* **91**(433), 217–221.

Verbeke, G. and Lesaffre, E. (1997), 'The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data', *Computational Statistics & Data Analysis* **23**(4), 541–556.

Wernecke, K.-D., Kalb, G., Schink, T. and Wegner, B. (2004), 'A mixed model approach to discriminant analysis with longitudinal data', *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **46**(2), 246–254.

Yerushalmy, J. (1947), 'Statistical problems in assessing methods of medical diagnosis, with special reference to x-ray techniques', *Public Health Reports (1896-1970)* pp. 1432–1449.

Zou, K. H. (2002), 'Receiver operating characteristic (roc) literature research', *On-line bibliography available from:¡ http://splweb. bwh. harvard. edu* **8000**.

Appendix A

Modified discriminant analysis step two

```
#~~~ second step ~~~#
# getProfiles to get patient visit time, age, contrast
# sensitivity, visual acuity and age for new patient id
#DNew.
IndProf1 <- getProfiles(t="visitmse", y=c("ageb","csse","vase","gender",
           "outcome"),id="pat", data=DNew)
   # w.prior is prior probability for each group.
w.prior<-c(0.5835,0.4165)

n1<-length(unique(DNew$pat)) ##number of unique patients
  P_hatGN <-rep(0,7)

for(i in 1:n1){   ##For each unique patient...
Pat<-IndProf1[[i]]##Select their individual profile
l<-dim(Pat)[1] ##Check how many observations it has

for(j in 1:l){##For each longitudinal observation...
#One column for each fixed effect parameter,
#one row for each observation of each marker
# the first 6 rows intercept and slopes for first marker CS
# the second 6 rows are intercepts and slopes for VA.
 X_new<-matrix(data=0,nrow=(j*2),ncol=6)
      X_new[1:j,1] <- rep(1,j) # intercept for CS
      X_new[1:j,2] <- rep(0,j)
      X_new[1:j,3] <- Pat[1:j,"ageb"]
      X_new[1:j,4] <- rep(0,j)
      X_new[1:j,5] <- Pat[1:j,"visitmse"]
      X_new[1:j,6] <- rep(0,j)
      X_new[(j+1):(j+j),1] <- rep(0,j)
      X_new[(j+1):(j+j),2] <- rep(1,j) # intercept for VA
```

```
      X_new[(j+1):(j+j),3] <- rep(0,j)
      X_new[(j+1):(j+j),4] <- Pat[1:j,"ageb"]
      X_new[(j+1):(j+j),5] <- rep(0,j)
      X_new[(j+1):(j+j),6] <- Pat[1:j,"visitmse"]
      Z_new<-matrix(data=0,nrow=(2*j),ncol=4)
#intercept and slope column for each
       marker (2x2),
#row for each observation time for each marker.
      Z_new[1:j,1] <- rep(1,j)
      Z_new[1:j,2] <- rep(0,j)
      Z_new[1:j,3] <- Pat[1:j,"visitmse"]
      Z_new[1:j,4] <- rep(0,j)
      Z_new[(j+1):(j+j),1] <- rep(0,j)
      Z_new[(j+1):(j+j),2] <- rep(1,j)
      Z_new[(j+1):(j+j),3] <- rep(0,j)
      Z_new[(j+1):(j+j),4] <- Pat[1:j,"visitmse"]
      # new observations for CS and VA.
      Y_new<-c(Pat$csse[1:j],Pat$vase[1:j])

      Iden = diag(j)
      #fixed effects parameters for Group 0
      B0 <- c(f0$coefficient$fixed)
      #fixed effects parameters for Group 1
      B1 <- c(f1$coefficient$fixed)
      #Muo is the mean of LMM model which is XB
      # covariate matrix for fixed effects multiple
      # vector of fixed effect
      Mu0<-X_new%*%B0
      Mu1<-X_new%*%B1
      # ZDZ0 is multiple covariate matrix for random effects
      # with covariance matrix for random effects
      ZDZ0 = Z_new%*%D0.mat%*%t(Z_new)
      ZDZ1 = Z_new%*%D1.mat%*%t(Z_new)
      # sigmaI0 is residual error matrix multiple identical matrix
      sigmaI0 = sigma0 %x% Iden
      sigmaI1 = sigma1 %x% Iden
      #Sigma0 is the variance of LMM model
      Sigma0 = ZDZ0 + sigmaI0
      Sigma1 = ZDZ1 + sigmaI1
      # SigmaPool is pool covariance matrix
      SigmaPool = (Sigma0*(n.D0*j-1)+Sigma1*(n.D1*j-1))/(n.D*j-2)
      # dMVN is used to get vector with evaluated values of the (log-)density
      #marg0Q to get evaluated values of the quadratic
      #(log-)density for Group 0
      marg0Q <- dMVN(Y_new,mean=Mu0,Sigma=Sigma0)
      marg1Q <- dMVN(Y_new,mean=Mu1,Sigma=Sigma1)
      #marg0L ito get evaluated values of the linear
      #(log-)density for Group 0
      marg0L <- dMVN(Y_new,mean=Mu0,Sigma=SigmaPool)
```

```
      marg1L <- dMVN(Y_new,mean=Mu1,Sigma=SigmaPool)

      p<-c(marg0Q,marg1Q,marg0L,marg1L)
      # P_hatGN is matrix include patient id, evaluated values of
      #quadratic density for Group 0 * prior probability for group 0,
      #evaluated values of quadratic density for Group 1
      #* prior probability for group 1, evaluated values of
      #linear density for Group 0 * prior probability for group 0,
     # evaluated values of linear density for Group 1
     # * prior probability for group 1, and patient's
     # Status

P_hatGN<-rbind(P_hatGN,c(floor(i),(Pat[j,1]),
(p[1:2]*w.prior)/sum(p[1:2]*w.prior),
                    (p[3:4]*w.prior)/
                    sum(p[3:4]*w.prior),Pat[j,6]))
  }}
```

Appendix B

Creating Simulated Datasets

Simulated dataset for Scenario 1, Chapter 4.

```
PBCSimulate <- function(s) {
    set.seed(s + 1606)
  ## Set up simulation parameters. Random Effects Mean
  # This vector means from Table 4.3.
  #For scenario 2, the first four means of the random effects is used.
  # This mean vector refers to Group 0
mu0 <- ParamEstPBC[c("b.Mean.1", "b.Mean.2", "b.Mean.3","b.Mean.4",
                     "b.Mean.5", "b.Mean.6", "b.Mean.7"), 1]
  # This mean vector refers to Group 1
mu1 <- ParamEstPBC[c("b.Mean.1", "b.Mean.2", "b.Mean.3","b.Mean.4",
                     "b.Mean.5", "b.Mean.6", "b.Mean.7"), 4]

  ## Random effects Covariance Matrix which from Table 4.3
  # The correlations of albumin and biliurbin markers that related to
  #scenario 2 are b.Corr.2.1, b.Corr.3.1,
  #b.Corr.4.1, b.Corr.3.2, b.Corr.4.2, and b.Corr.4.3.
D0Cov <- matrix(c(1, ParamEstPBC[c("b.Corr.2.1", "b.Corr.3.1",
                "b.Corr.4.1", "b.Corr.5.1", "b.Corr.6.1",
                "b.Corr.7.1"), 4], ParamEstPBC["b.Corr.2.1",4],
                1, ParamEstPBC[c("b.Corr.3.2", "b.Corr.4.2",
                "b.Corr.5.2", "b.Corr.6.2", "b.Corr.7.2"), 4],
                ParamEstPBC[c("b.Corr.3.1", "b.Corr.3.2"), 4],
                1, ParamEstPBC[c("b.Corr.4.3", "b.Corr.5.3",
                "b.Corr.6.3", "b.Corr.7.3"), 4], ParamEstPBC[c
                ("b.Corr.4.1", "b.Corr.4.2", "b.Corr.4.3"), 4], 1,
                 ParamEstPBC[c("b.Corr.5.4", "b.Corr.6.4",
                 "b.Corr.7.4"), 4], ParamEstPBC[c("b.Corr.5.1",
                 "b.Corr.5.2", "b.Corr.5.3", "b.Corr.5.4"), 4], 1,
                 ParamEstPBC[c("b.Corr.6.5", "b.Corr.7.5"), 4],
                 ParamEstPBC[c("b.Corr.6.1", "b.Corr.6.2",
                 "b.Corr.6.3", "b.Corr.6.4", "b.Corr.6.5"), 4], 1,
```

```
                ParamEstPBC[c("b.Corr.7.6"), 4], ParamEstPBC[c
                ("b.Corr.7.1", "b.Corr.7.2", "b.Corr.7.3",
                "b.Corr.7.4", "b.Corr.7.5", "b.Corr.7.6"), 4], 1),
                 nrow = 7, ncol = 7, byrow = TRUE)
# standard deviation for the random effects corresponds to Table 4.3.
# Again using the first four SD for Scenario 2.
SD <- ParamEstPBC[c("b.SD.1", "b.SD.2", "b.SD.3", "b.SD.4", "b.SD.5",
                    "b.SD.6", "b.SD.7"), 4]
# Here the correlation matrix is Converted to covariance matrices
D0 <- cor2cov(D0Cov, sd = SD)
   # alpha here relates to fixed effects parameters
   # scenario 2 is not include any fixed effects.
Alpha0 <- ParamEstPBC["alpha1", 1]
Alpha1 <- ParamEstPBC["alpha1", 4]

 ## Error variances.
    evar1 <- 0.159
    evar2 <- 0.169
    evar3 <- 0.157
    evar4 <- 0.198

  ## Simulate time points for 200 patients in Group 0
   # and 50 patients in Group 1, in total 250 patients.
    t0 <- rep(0, 250)
    t1 <- runif(250, 170, 200)
    t2 <- runif(250, 350, 390)
    t3 <- runif(250, 710, 770)
    t <- cbind(t0,
    t1, t2, t3)/30

    PatientData0 <- PatientData1 <- NULL
    # here start with Simulated dataset for Group 0.
for (i in 1:200) {
  ## Simulate random errors.
  Randerror <- c(rnorm(4, mean = 0, sd = evar1), rnorm(4, mean = 0,
                    sd = evar2), rep(0, 4), rep(0, 4))
  ## Simulate Patient random effect
  PatRan <- rMVN(1, mean = mu0, Sigma = D0)$x
  ## Simulate marker responses
  X <- rep(0, 16)
  X[13:16] <- t[i, ]
  Z <- matrix(data = 0, nrow = 16, ncol = 7)
  Z[1:4, 1] <- Z[5:8, 3] <- Z[9:12, 5] <- Z[13:16, 7] <- 1
  Z[1:4, 2] <- Z[5:8, 4] <- Z[9:12, 6] <- t[i, ]
  Yi <- X * Alpha0 + Z %*% PatRan + Randerror
  # Scenario 2 is not include Platelet and Spider.
  # that will be the same for group 1.
  Platelet <- rpois(4, exp(Yi[9:12]))
  Spiders <- rbinom(4, size = 1, exp(Yi[13:16])/(1 + exp(Yi[13:16])))
```

```
  #Here I combine the four markers for Scenario 1.
  tmp <- cbind(rep(i, 4), t[i, ], Yi[1:4], Yi[5:8], Platelet,
               Spiders, rep(0, 4))
  PatientData0 <- rbind(PatientData0, tmp)
    }
  ## Same for 50 patients in group 1.
for (i in 201:250) {
  ## Simulate random errors.
  Randerror <- c(rnorm(4, mean = 0, sd = evar3), rnorm(4, mean = 0,
                       sd = evar4), rep(0, 4), rep(0, 4))
  ## Simulate Patient random effect
  PatRan <- rMVN(1, mean = mu1, Sigma = D0)$x
  ## Simulate marker responses
  X <- rep(0, 16)
  X[13:16] <- t[i, ]
  Z <- matrix(data = 0, nrow = 16, ncol = 7)
  Z[1:4, 1] <- Z[5:8, 3] <- Z[9:12, 5] <- Z[13:16, 7] <- 1
  Z[1:4, 2] <- Z[5:8, 4] <- Z[9:12, 6] <- t[i, ]
  Yi <- X * Alpha1 + Z %*% PatRan + Randerror
  Platelet <- rpois(4, exp(Yi[9:12]))
  Spiders <- rbinom(4, size = 1, exp(Yi[13:16])/(1 + exp(Yi[13:16])))
  tmp <- cbind(rep(i, 4), t[i, ], Yi[1:4], Yi[5:8], Platelet,
               Spiders, rep(1, 4))
  PatientData1 <- rbind(PatientData1, tmp)
    }
 ## Combine the data groups together to create a simulated data set.
PatientData <- as.data.frame(rbind(PatientData0, PatientData1))
colnames(PatientData) <- c("id", "time", "Albumin", "bilirubin",
                           "Platelet", "Spiders", "Status")
    filename <- paste0("PBCsim", s, ".RData")
    save(PatientData, file = filename)
}
```

Appendix C

Performing Longitudinal Discriminant Analysis

```
library(mixAK)
# mod0 is the MGLMM fit to the patients in group 0.
>mod0 <- GLMM_MCMC(y = PBC_0Train[, c("Albumin", "bilirubin", "Platelet",
                  "Spiders")], dist = c("gaussian", "gaussian", "poisson
                  (log)","binomial(logit)"), id = PBC_0Train[, "id"], x =
                  list(Albumin = "empty", bilirubin = "empty", Platelet =
                  "empty", Spiders = PBC_0Train[, "time"]), z = list(
                  Albumin =PBC_0Train[,"time"],bilirubin = PBC_0Train[,
                  "time"], Platelet = PBC_0Train[, "time"], Spiders =
                  "empty"), random.intercept = c(Albumin = TRUE,bilirubin
                  = TRUE, Platelet = TRUE, Spiders = TRUE), prior.b = list
                  (Kmax = 1), nMCMC = c(burn = 5000, keep = 10000, thin =
                  10, info = 500), PED = FALSE)

# mod1 is the MGLMM fit to the patients in group 1.
>mod1 <- GLMM_MCMC(y = PBC_1Train[, c("Albumin", "bilirubin", "Platelet",
                  "Spiders")], dist = c("gaussian", "gaussian", "poisson
                  (log)", "binomial(logit)"), id = PBC_1Train[, "id"], x =
                  list(Albumin = "empty", bilirubin = "empty", Platelet =
                  "empty", Spiders = PBC_1Train[, "time"]), z = list(
                  Albumin = PBC_1Train[,  "time"], bilirubin = PBC_1Train[,
                  "time"], Platelet = PBC_1Train[, "time"], Spiders =
                  "empty"), random.intercept = c(Albumin = TRUE, bilirubin
                  = TRUE, Platelet = TRUE, Spiders = TRUE), prior.b = list
                  (Kmax = 1), nMCMC = c(burn = 5000, keep = 10000, thin =
                  10, info = 500), PED = FALSE)

# cluster is predict function to predict the status of
# new patients based on the patient data contained in
# the dataframe PBCNew.
>cluster <- GLMM_longitDA2(mod = list(mod0, mod1), w.prior = c(0.8, 0.2), y
```

```
                        = PBCNew[, c("Albumin", "bilirubin", "Platelet",
                        "Spiders")], id = PBCNew[, "id"], xz.common = TRUE,
                        x = list(Albumin = "empty", bilirubin = "empty",
                        Platelet = "empty", Spiders = PBCNew[, "time"]),
                        z = list(Albumin = PBCNew[, "time"], bilirubin =
                        PBCNew[, "time"], Platelet = PBCNew[, "time"],
                        Spiders = "empty"))

# Resul1 is matrix include patient id, marginal  posterior probability,
# conditional posterior probability, random effects posterior probability and
#finally the patient's status.
>Resul1 <- cbind(PBCNew$id[1], cluster$pi_marg, cluster$pi_cond,
                cluste$pi_reff, PBCNew$Status[1])
>colnames(Resul1) <- c("id", "Marg0", "Marg1", "Cond0", "Cond1", "RanEf0",
                      "RanEf1", "Status")
```

Appendix D

Calculating classification accuracy measurements

```
CVRes <- as.data.frame(CVRes)
Gp0 <- sum(CVRes$Status == 0)
Gp1 <- sum(CVRes$Status == 1)
 ## for each potential cutoff point
for (j in 1:l_p) {
# first five measurements are used to assess the
# accuracy of the marginal predictions.
 SensitivityM[NCV, j] <- sum(CVRes$Marg1 >= p[j] & CVRes$Status == 1)/Gp1
 SpecificityM[NCV, j] <- sum(CVRes$Marg1 < p[j] & CVRes$Status == 0)/Gp0
 PCCM[NCV, j] <- (sum(CVRes$Marg1 >= p[j] & CVRes$Status == 1) + sum(
                  CVRes$Marg1 < p[j] & CVRes$Status == 0))/(Gp0 + Gp1)
 PPVM[NCV, j] <- sum(CVRes$Marg1 >= p[j] & CVRes$Status == 1)/sum(
                     CVRes$Marg1 >= p[j])
 NPVM[NCV, j] <- sum(CVRes$Marg1 < p[j] & CVRes$Status == 0)/sum(
                     CVRes$Marg1 < p[j])
# The next five measurements are used to assess the
# accuracy of the conditional predictions.
 SensitivityC[NCV, j] <- sum(CVRes$Cond1 >= p[j] & CVRes$Status == 1)/Gp1
 SpecificityC[NCV, j] <- sum(CVRes$Cond1 < p[j] & CVRes$Status == 0)/Gp0
 PCCC[NCV, j] <- (sum(CVRes$Cond1 >= p[j] & CVRes$Status == 1) + sum(
                      CVRes$Cond1 < p[j] & CVRes$Status == 0))/(Gp0 + Gp1)
 PPVC[NCV, j] <- sum(CVRes$Cond1 >= p[j] & CVRes$Status == 1)/sum(
                     CVRes$Cond1 >= p[j])
 NPVC[NCV, j] <- sum(CVRes$Cond1 < p[j] & CVRes$Status == 0)/sum(
                     CVRes$Cond1 < p[j])
# Here the five measurements are used to assess the
# accuracy of the Random effects predictions.
 SensitivityR[NCV, j] <- sum(CVRes$RanEf1 >= p[j] & CVRes$Status == 1)/Gp1
 SpecificityR[NCV, j] <- sum(CVRes$RanEf1 < p[j] & CVRes$Status == 0)/Gp0
 PCCR[NCV, j] <- (sum(CVRes$RanEf1 >= p[j] & CVRes$Status == 1) + sum(
                      CVRes$RanEf1 < p[j] & CVRes$Status == 0))/(Gp0 + Gp1)
```

```
 PPVR[NCV, j] <- sum(CVRes$RanEf1 >= p[j] & CVRes$Status == 1)/sum(
                     CVRes$RanEf1 >= p[j])
 NPVR[NCV, j] <- sum(CVRes$RanEf1 < p[j] & CVRes$Status == 0)/sum(
                     CVRes$RanEf1 < p[j])
    }
  # For each prediction, AUC is calculated and the values
  # of the sensitivity etc at the optimal cutoffs on the ROC curves.
AUCM <- trapz(SpecificityM[NCV, ], SensitivityM[NCV, ])

# here calculate the optimal threshold which is
# balanced between sensitivity and specificity
D_M <- sqrt((1 - SensitivityM[NCV, ])^2 + (1 - SpecificityM[NCV, ])^2)
m_optM <- which.min(D_M)
Marginal[NCV, ] <- c(p[m_optM], SensitivityM[NCV, m_optM], SpecificityM
                    [NCV, m_optM], PCCM[NCV, m_optM], AUCM, PPVM[NCV,
                     m_optM], NPVM[NCV, m_optM])

AUCC <- trapz(SpecificityC[NCV, ], SensitivityC[NCV, ])
D_C <- sqrt((1 - SensitivityC[NCV, ])^2 + (1 - SpecificityC[NCV, ])^2)
m_optC <- which.min(D_C)
Conditional[NCV, ] <- c(p[m_optC], SensitivityC[NCV, m_optC], SpecificityC
                       [NCV, m_optC], PCCC[NCV, m_optC], AUCC, PPVC[NCV,
                        m_optC], NPVM[NCV, m_optC])

AUCR <- trapz(SpecificityR[NCV, ], SensitivityR[NCV, ])
D_R <- sqrt((1 - SensitivityR[NCV, ])^2 + (1 - SpecificityR[NCV, ])^2)
m_optR <- which.min(D_R)
Random[NCV, ] <- c(p[m_optR], SensitivityR[NCV, m_optR], SpecificityR
                  [NCV, m_optR], PCCR[NCV, m_optR], AUCR, PPVR[NCV,
                   m_optR], NPVR[NCV, m_optR])
```