2021-01-08

Classification Models for Multivariate Non-normal Repeated Measures Data

Brobbey, Anita

UNIVERSITY OF CALGARY

Classification Models for Multivariate Non-normal Repeated Measures Data

by

Anita Brobbey

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILMENT OF THE REQUIREMENT FOR THE

DEGREE OF DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN COMMUNITY HEALTH SCIENCES

CALGARY, ALBERTA

JANUARY, 2021

Abstract

Multivariate repeated measures data, in which multiple outcomes are repeatedly measured at two or more occasions, are commonly collected in several disciplines (e.g., medicine, ecology, environmental sciences), where investigators seek to discriminate between population groups or make predictions based on changes in multiple correlated outcomes over time. Repeated measures discriminant analysis have been developed and applied to address these research questions. These classification models, which have been mostly developed based on growth curve models, covariance pattern models, and mixed-effects models, are advantageous in that they can account for complex correlation structures in multivariate repeated measures data (e.g., within-outcome and between-outcome correlations) to improve their predictive accuracy. However, they largely rely on the assumption of multivariate normality, which is rarely satisfied in multivariate repeated measures data. To our knowledge, there has been limited investigation of the behavior of these existing models in multivariate non-normal repeated measures data.

The overarching goal of this research was to develop robust repeated measures discriminant analysis classifiers for multivariate non-normal repeated measures data. Specifically, we developed repeated measures discriminant analysis based on maximum trimmed likelihood estimators (MTLE) and generalized estimating equations (GEE) estimators and examine their accuracy in comparison to classifiers based on maximum likelihood estimation (MLE) using Monte Carlo methods. The simulation conditions examined, included population distribution, sample size, covariance structure (between-outcomes and within-outcome), covariance heterogeneity, repeated number of occasions, and number of outcome variables. The Monte Carlo study results indicated that the proposed methods increased overall mean classification accuracy

by 2% - 15% in multivariate non-normal repeated measures data compared to repeated measures discriminant analysis based on MLE under most scenarios. Data from two cohort studies were used to illustrate the implementation of the proposed repeated measures discriminant analysis methods.

The outcomes of this research includes novel multivariate classifiers for predicting group membership in multivariate normal and non-normal repeated measures data. This research contributes to the advancement of statistical science on methods for analyzing multivariate repeated measures data.

Acknowledgments

I thank the Almighty God for all his blessings throughout my doctoral education. I would like to express my profound gratitude to my supervisor Dr. Tolulope Sajobi, for his patience, invaluable assistance, and constructive criticisms throughout my study. This dissertation could not have been accomplished without his insights and knowledge into the statistical subjects. I have learnt a lot under his supervision, and he has prepared me well to become an independent researcher. Dr. Sajobi was not only my supervisor, he is a mentor, friend, and role model. I appreciate his continuous encouragement and guidance in the last four years. He was always available to help and advise me on research directions, struggles and personal issues. Moreover, I am very grateful to Dr. Sajobi for the funding he provided me toward my PhD and the teaching assistant opportunities he gave me throughout my doctoral education.

Special thanks to my co-supervisor, Dr. Samuel Wiebe, for his constructive suggestions, support and more importantly, it was a real honour and privilege to gain from his experience and his exceptional clinical knowledge. His critical questions guided me to think more about the clinical implications of my research questions, methodologies and findings in solving real-life problems. I deeply appreciate the continuous support, guidance and contributions of my supervisory committee members, Drs. Tyler Williamson and Alberto Nettel-Aguirre. Their valuable comments, insightful suggestions and intellectual support have brought this study to a successful completion. Thank you, Alberto for introducing me to Dr. Susan Samuel. I am grateful to Dr. Samuel and her team at the Alberta Children's Hospital for all the support and for giving me a chance to work as a statistical analyst in an interdisciplinary environment. Working with them filled me with a lot of joy and inspired me to work hard to complete this journey.

My warmest and deepest gratitude to my husband and my dear parents for their love, constant support, and encouragement throughout this journey. Finally, my sincere thanks go to all graduate students in Sajobi Research Methods & Data Analytics (sRMDA) Lab, friends and loved ones who contributed in various ways to make my Ph.D dream a reality. Thank you all!

Dedication

To my husband, Kwaku Kyeremeh and my dad, Kofi Brobbey

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AIC	Akaike information criterion
AR-1	First-order autoregressive
CPM	Covariance pattern model
CS	Compound symmetry
FSA	Feasible Solution Algorithm
GEE	Generalized estimating equations
GLM	Generalized linear model
LDA	Linear discriminant analysis
QDA	Quadratic discriminant analysis
QIC	Quasi-likelihood under the independence model criterion
MAR	Missing at random
MCA	Mean classification accuracy
MCAR	Missing completely at random
MCD	Minimum covariance determinant
MTLE	Maximum trimmed likelihood estimators
MVE	Minimum volume ellipsoid
UNAR	Unstructured between-outcomes and first-order autoregressive within-outcome Kronecker correlation structures
UN	Unstructured correlation structure
UNCS	Unstructured between-outcomes and compound symmetry within-outcome Kronecker correlation structures

Epigraph

The best view comes after the hardest climb

Anonymous

1. Introduction

1.1 Background

Multivariate repeated measures data, where measurements are collected at two or more occasions for multiple variables[1-16], have been used for discriminating between population groups. For example, Fieuws et al. used longitudinally collected biochemical and physiological markers to develop a linear discriminant analysis (LDA) rule to predict 10-year success of graft in patients who received a kidney transplant [1]. Using repeated measures of prostate-specific antigen (PSA), free testosterone index (FTI), and body mass index (BMI), discriminant analysis classifier has been developed to estimate probabilities of prostate cancer onset in a population of cancer patients[3, 9]. Marshall and Baron[6] also developed a classification for determining pregnant women at risk of birth complications using longitudinally collected information on two biomarkers in a sample of pregnant women.

The majority of these classifiers for multivariate repeated measures data have been developed based on mixed-effects regression[1, 2, 4, 7-9, 12-15] and covariance pattern models[10, 11] which allow for the use of parsimonious means and covariance structures. These models are developed based on the assumption of multivariate normality, which is often not tenable in repeated measures studies such as studies of health-related quality of life where outcomes are typically skewed or heavy-tailed[17, 18]. For example, Fieuws et al. commented that a trivariate linear mixed model that was used to analyze longitudinal information on systolic blood pressure, body mass index, and blood triglycerides for hypertension prediction, was not appropriate to describe all longitudinal profiles in their renal graft failure data study because some of their data were non-normally distributed, and used a generalized linear mixed model [1, 19].

Given the demand for multivariate repeated measures classification models[1, 2, 4, 7-9, 11-14, 16] and the increasing collection of multivariate repeated measures health data, which are mostly characterized by non-normal distributions[1, 2, 14, 16], there is the need for accurate repeated measures classifiers that overcome the limitations of multivariate normality assumptions and account for the complex correlation structures in data especially in small-sampled data. Developing such classification models will be useful for classifying individuals into one of two (or more) populations using multivariate repeated measures designs involving continuous, discrete and mixed type outcomes, which are routinely collected in several disciplines.

1.2 Research Questions & Objectives

The overarching aim of this research is to develop accurate classification models for discriminating between population groups in multivariate repeated measures data when the assumption of multivariate normality of the outcome variables is not tenable in small-sampled data. The study's research questions are as follows

1. How accurate are existing repeated measures discriminant analysis classifiers when applied to discriminate between study samples of multivariate non-normal repeated measures data?

2. How accurate are repeated measures discriminant analysis based on trimmed estimators and multivariate Generalized estimating equations (GEE) in comparisons to the conventional discriminant analysis models based on MLE for classification in multivariate non-normal repeated measures?

3. What is the impact of mis-specification of correlation structures on the accuracy of repeated measures discriminant analysis based on GEE when used for classification in multivariate repeated measures data? How is the impact of mis-specification influenced by outcome variable distribution?

The study objectives are to:

1. ***develop*** repeated measures classification models based on robust estimation methods for discriminating between population groups in multivariate repeated measures data characterized by non-normal distributions

2. ***develop*** discriminant analysis estimation procedure suitable for modelling multivariate discrete, continuous, and mixed type multivariate repeated measures data together with covariates, and

3. ***investigate*** the impact of correlation structure misspecification on classification accuracy in GEEs discriminant analysis under a variety of simulation generation conditions using Monte Carlo methods.

In addition, the implementation of these developed methods in the study will be demonstrated using example datasets from population-based cohorts of patients with new onset epilepsy and inflammatory bowel diseases.

1.3 Organization of Thesis

This thesis is structured as a manuscript-based dissertation that consists of three manuscripts, which are, as at the time of completing the dissertation, under review in peer-reviewed journals. Chapter 2 provides a review of the literature of the relevant methodologies used in this dissertation. This includes the literature on repeated measures discriminant analysis, covariance structure mis-specification in repeated measures data analysis, and robustness of statistical estimators. Chapter 3 describes the results of an investigation of the accuracy of repeated measures discriminant analysis based on maximum trimmed likelihood estimation and its classification performance in comparison to classifiers based on maximum likelihood estimation (MLE). Data from the Manitoba Inflammatory Bowel Disease (IBD) cohort study was used to demonstrate the implementation of the procedures. In Chapter 4, a new class of repeated measures discriminant analysis classifiers based on multivariate generalized estimation equations was developed for classification in multivariate repeated measures data characterized by different types of outcomes (e.g., binary, count, ordinal). The accuracy of this class of classifiers in comparison to the conventional repeated measures discriminant analysis based on MLE was examined using Monte Carlo methods. Data from a prospective longitudinal cohort of children with new-onset epilepsy were used to illustrate the implementation of these methods. This manuscript is currently under review in *Statistical Methods in Medical Research* at the time of submission of this dissertation. Chapter 5 details the results of a study that examined the impact of mis-specification of GEE working correlation structure on the classification accuracy of repeated measures discriminant analysis based on multivariate GEE under a variety of different distributions and number of outcomes. This manuscript is currently under review in *Communications in Statistics.*

The dissertation closes with a discussion about the implications of these findings and suggestions for future research in Chapter 6.

2. Literature Review

In this chapter, we outline the general framework (formulation) of the discrimination problem, present the main approaches of classical discriminant analysis, and literature on discriminant analysis procedures for multivariate repeated measures data, which includes procedures based on covariance pattern, mixed-effects, and GEE models. This chapter introduces the concept of robustness of statistical estimators, existing robust procedures for multivariate non-normal data used in this research. These include trimmed estimators (Minimum Covariance Determinant and Minimum Volume Ellipsoid).

2.1 Multivariate Repeated Measures Discriminant Analysis

Let $\mathbf{y}_{ij}$ be the $pq \times 1$ vector of observed measurements corresponding to p repeated measurements on each of the q outcome variables for the ith study participant in the jth population $(i = 1, \dots\dots, n_j; j = 1,2; N = n_1 + n_2)$. The vectors are structured such that the repeated measurements are nested within each variable. While this manuscript focuses on the analysis of two-population designs, the procedures have been generalized to multi-population problems [20-22]. Assume that $\mathbf{y}_{ij} \sim N_{pq}(\boldsymbol{\mu}_j, \boldsymbol{\Omega}_j)$, where $\boldsymbol{\mu}_j$ and $\boldsymbol{\Omega}_j$ are the population mean and covariance for the jth group and are estimated by $\widehat{\boldsymbol{\mu}}_j$ and, $\widehat{\boldsymbol{\Omega}}_j$ respectively. Denote π_j as the prior probability, or the membership probability of population j, that is the probability for an observation to come from population j. The optimal classifier (i.e, the Bayes rule) is based on conditional probability, which by the Bayes theorem takes the following form

$$P(\mathbf{y} \in \mathrm{j} \mid \boldsymbol{Y} = \mathbf{y}) = \frac{P(\boldsymbol{Y} = \mathbf{y} \mid \mathbf{y} \in \mathrm{j}) P(\mathbf{y} \in \mathrm{j})}{P(\boldsymbol{Y} = \mathbf{y})} \tag{2.1}$$

$$= \frac{\pi_j f_j(\mathbf{y})}{\sum_{j=1}^{2} \pi_j f_j(\mathbf{y})}$$

where $j = 1,2$ (number of populations), $f_j(\mathbf{y})$ is the likelihood (conditional density function), and π_j is the prior probabilities for population j. The classification decision function can be written as

$$\underset{j}{\operatorname{argmax}}\, P(\mathbf{y} \in \mathrm{j} \mid \boldsymbol{Y} = \mathbf{y}) = \underset{j}{\operatorname{argmax}}\, \pi_j f_j(\mathbf{y}) \tag{2.2}$$

Suppose $\boldsymbol{Y} \mid \mathbf{y} \in \mathrm{j} \sim N_{pq}\left(\boldsymbol{\mu}_j, \boldsymbol{\Omega}_j\right)$, where $\boldsymbol{\mu}_j$ and $\boldsymbol{\Omega}_j$ are the mean vector and covariance matrix respectively for population j. Taking the logarithm of the classification function with simple calculations reveal that,

$$d_j(\mathbf{y}) = \left[\mathbf{y} - \boldsymbol{\mu}_j\right]^T \left(\boldsymbol{\Omega}_j\right)^{-1} \left[\mathbf{y} - \boldsymbol{\mu}_j\right] + \log\left|\boldsymbol{\Omega}_j\right| - 2\log\left(\pi_j\right) + \text{constant} \tag{2.3}$$

where the first term is the so-called Mahalanobis distance between $\mathbf{y}$ and $\boldsymbol{\mu}_j$. The quadratic discriminant analysis (QDA) classification, when the population covariances are not equal (i.e., $\boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_2$) assign $\mathbf{y}$ to population j if $d_j(\mathbf{y})$ achieves the maximum among $[d_1(\mathbf{y}), d_2(\mathbf{y})]$. If further assume a common covariance matrix of the two populations under the assumption of homoscedasticity ($\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2 = \boldsymbol{\Omega}$), we can simply QDA as linear discriminant analysis (LDA)

$$d_j(\mathbf{y}) = \left[\mathbf{y} - \boldsymbol{\mu}_j\right]^T (\boldsymbol{\Omega})^{-1} \left[\mathbf{y} - \boldsymbol{\mu}_j\right] + \log|\boldsymbol{\Omega}| - 2\log\left(\pi_j\right) + \text{constant} \tag{2.4}$$

$$= -\mathbf{y}^T \boldsymbol{\Omega}^{-1} \boldsymbol{\mu}_j + \left(\frac{1}{2} \boldsymbol{\mu}_j^T \boldsymbol{\Omega}^{-1} \boldsymbol{\mu}_j - \log\left(\pi_j\right)\right) + \text{constant}$$

It is also well known that with a common covariance structure among populations, if both populations have equal membership probabilities, this rule (2.4) coincides with Fisher's linear

discriminant rule[22]. As $\boldsymbol{\mu}_j$, $\boldsymbol{\Omega}_j$ and π_j are in practice unknown, they have to be estimated from the sampled data. To estimate $\boldsymbol{\mu}_j$ and $\boldsymbol{\Omega}_j$, one usually uses the population mean $\bar{\mathbf{y}}_{ij}$ and the population empirical covariance matrix $\mathbf{S}_j$ (MLEs), yielding the classical discriminant rule. Two choices are popular for the estimates of the membership probabilities π_j. Either the π_j are considered to be constant over all populations, yielding $\hat{\pi}_j$= ½ for each poulation or estimated as the relative frequencies of the observations in each population, thus $\hat{\pi}_j = \frac{n_j}{N}$. The accuracy of the classification rules are commonly evaluated by the overall mean classification accuracy, i.e. the proportion of correctly specified individuals. The overall classification accuracy is estimated by

$$\text{Overall classification accuracy} = \frac{\text{correct classifications}}{\text{Total sample size}} = \frac{n_{11} + n_{22}}{N} \qquad (2.5)$$

where n_{11} and n_{22} are the number of study participants correctly assigned to populations 1 and 2, respectively. The classical linear and quadratic discriminant analyses are popular discriminant analysis models because of their simplicity and flexible assumptions. However, these models are inherently less accurate when there is contamination due to outliers or non-normal distributions such as those obtained in skewed or heavy-tailed distributions[23-26]. Therefore, it is important to consider robust alternatives to these estimators and in the last two decades several affine equivariant estimators possessing a high breakdown point have been proposed[27-30]. Moreover, classical discriminant analysis procedures require balanced data, do not include covariate information, and cannot be applied to high dimensional data in which sample size is less than the product of the number of repeated measurements and the number of outcome variables[31, 32].

Early research about multivariate repeated measures data focused on procedures based on growth curve models, covariance pattern models, and mixed-effects models. Discriminant analysis

procedure based on the growth curve model[33] extended the concept of discriminant analysis to multivariate outcome curves observed over a specified time interval. The multivariate repeated measures growth curve model involves fitting outcome curves by linear interpolation between successive observations and classifying an individual outcome curve to the group curve it resembles most. It requires that the outcome curves in the training sample are fully observed over the considered time interval; however, does not require equally spaced observations and makes no assumptions about the nature of the variables. In addition to growth curve models, there have been a number of developments in discriminant analysis procedures based on covariance pattern and mixed-effects models for multivariate repeated measures data.

2.1.1 Covariance Pattern Model

Repeated measures discriminant analysis based on the covariance pattern model have also been described for multivariate repeated measures data[4, 10, 34]. Covariance pattern model (CPM) assumes all outcomes follow the multivariate normal distribution, and multivariate linear regression model can be used but assumes the variance-covariance matrix to be of a certain form[35]. The regression model for CPM for each population membership in matrix form can be written as

$$\mathbf{y}_{ij} = \mathbf{X}_i\boldsymbol{\beta}_j + \boldsymbol{e}_{ij} \qquad (2.6)$$

where $\mathbf{y}_{ij}$ is a pq x 1 vector of q correlated outcomes that are each repeatedly measured at p occasions in the jth population, and $\mathbf{X}_i$ is the corresponding pq x Kq block diagonal covariate (design) matrix, and $\boldsymbol{e}_{ij} \sim \mathrm{N}_{pq}(\mathbf{0}, \boldsymbol{\Omega}_j)$ is a vector containing the error components. Population means are computed from estimates of the parameters, that is, $\widehat{\boldsymbol{\mu}}_j = \mathrm{E}(\mathbf{y}_{ij}) = \mathbf{X}_i\widehat{\boldsymbol{\beta}}_j$, and the $pq \times pq$

variance-covariance matrix ($\boldsymbol{\Omega}_j$) has a *functional form*. No random effects are included in the regression model, so these models do not distinguish the variance term in within-subjects and between-subjects variance. The covariance matrix $\boldsymbol{\Omega}_j$ of the model or residuals provides information about both the association within each outcome variable over time and also the association between the outcome variables. For fully balanced data set, completely unspecified covariance matrix $\boldsymbol{\Omega}_j$ leads to a total of $pq\ (pq + 1)/2$ unknown parameters to be estimated for any statistical inference, where p is the number of time points at which measurements have been taken [36, 37]. Estimation of all parameters will require a very large sample, which may not always be feasible ($n_j < pq$) or may be cost prohibitive.

To reduce the dimension of the unknown parameters of the covariance matrix, a more parsimonious structure is sometimes used, such as a Kronecker product (or separable matrix) of the covariance matrix $\boldsymbol{\Omega}_j = \boldsymbol{V}_j \otimes \boldsymbol{\Sigma}_j$, where $\boldsymbol{V}_j$ and $\boldsymbol{\Sigma}_j$ are $p \times p$ and $q \times q$ positive definite matrices for the p time points and q outcomes, respectively, and $\otimes$ is the Kronecker product sign [36-40]. Even for unstructured $p \times p$ and $q \times q$ covariance matrices, a reduction in the number of covariance parameters is obtained[36-43]. The number of unknown parameters to be estimated is only $p(p+1)/2 + q(q+1)/2 - 1$ under a Kronecker product structure which is much less than $pq\ (pq + 1)/2$ in an unstructured variance–covariance matrix[38]. Further assumed structure for the repeated measurements of each outcome, such as a first-order autoregressive (AR-1) or compound symmetry (CS) covariance, can lead to an even more parsimonious covariance model; so that the number of unknown parameters to be estimated further reduces to $1+ q(q+1)/2$ [36, 38, 40]. The structuring of covariance (CS and AR-1) provides flexible model for covariance, resulting in improvement in precision, particularly in the analysis based on small sample[36, 38, 40]. However, the Kronecker product model implies that the correlation matrix $\boldsymbol{V}$ of the repeated measures on a given

outcome variable is assumed to be the same for all outcome variables and the variance-covariance matrix $\boldsymbol{\Sigma}$ between the measurements on all outcome variables at a given time point is assumed constant for all time points, which may not be realistic in many applications [36, 37]. Thus, the choice of appropriate covariance structure is crucial, and it is vital to test the appropriateness of the covariance structure on the data before any statistical analysis[11]. Inferences of interest are easily influenced by the correlation structure's assumptions, however unstructured correlation structure might cause convergence problems as the number of parameters to be estimated grows rapidly [44].

The main advantage of using a structured variance-covariance matrix over an unstructured one is that the number of unknown parameters decreases substantially, assisting analysis of small sample data[10]. The covariance procedures can result in efficient classification rules in high-dimensional data but can result in decreased classification accuracy when the mean and/or covariance structure is/are misspecified[10]. The inclusion of covariates, specification of mean and parsimonious covariance structure in these models further improve classification accuracy compared to the classical discriminant analysis based on unstructured means and population covariances. However, none of the classifiers developed based on covariance pattern models allow for incomplete longitudinal data or non-normal distributions.

2.1.2 Mixed Effects Models

Many researchers have proposed discriminant analysis models based on the use of mixed-effects models for multivariate repeated measures data[1, 2, 4, 7-9, 12, 13]. Covariance pattern models do not address the incompleteness of the data or the issue of missing values. Discarding data with missing components can result in appreciable information loss[45]. To jointly model multivariate repeated measures data using mixed-effects models, a mixed effect model is defined for each outcome variable and then the outcome variables are related through the random-effects[35] for each population.

$$\mathbf{y}_{ij} = \mathbf{X}_i\boldsymbol{\beta}_j + \mathbf{Z}_i\mathbf{d}_{ij} + \boldsymbol{e}_{ij}$$

where $\mathbf{X}_i$ is the pq x Kq block diagonal covariate matrix for fixed effects, $\mathbf{Z}_i$ is the corresponding pq x Lq covariate matrix for random effects, $\boldsymbol{\beta}_j \sim (\boldsymbol{\beta}'_{1j}, \boldsymbol{\beta}'_{2j}, \ldots, \boldsymbol{\beta}'_{qj})'$ the vectors of fixed effects, $\boldsymbol{e}_{ij} \sim \mathrm{N}_{pq}(\mathbf{0}, \boldsymbol{\Omega})$ is a vector containing the error components, and $\boldsymbol{d}_{ij} \sim \mathrm{N}_{pq}(\mathbf{0}, \mathbf{D})$, the joint distribution of *i*th subject-specific random effects pertaining to all outcomes. The covariance matrix $\mathrm{var}(\boldsymbol{d}_{ij}) = \mathbf{D}$ is assumed to be a generally unstructured positive definitive matrix. The relationship between the outcome variables can be induced through the random-effects in one of two ways. The first approach is to include random-effects that are common between the outcomes. These models are referred to as shared random effects models. The shared random -effects models reflect the belief that common set of underlying characteristics of the individual governs the outcome processes[35]. This assumption can reduce the computational problem substantially, since the number of random effects in the model does not increase with the number of outcome variables. However, one drawback of the shared random effects approach is that it implies that there is a perfect positive correlation between the shared random effects which is typically unrealistic[35]. The second approach is to include unique random effects for each outcome variable, and then model

the relationship between the outcome variables through the covariances of the random-effects[1, 4, 35]. These are referred to as correlated random-effects models. In contrast to the shared random effects approach, this model allows the correlation between random effects to be positive or negative. By examining these correlations, one can get an idea of the relationship between the trajectories of the outcomes. It is also possible to calculate a measure of the how correlation changes over time ("evolution of association")[1].

Various discriminant analysis procedures based on mixed-effects models have been made in recent years for multivariate repeated measures data. Multivariate linear and non-linear mixed-effects models that assume unstructured[1, 4] and Kronecker product structure[7-9, 31, 46] for the variance-covariance matrix have been introduced. For example, Marshall et al.[8] investigated discriminant analysis models based on bivariate nonlinear mixed-effects models with Kronecker product covariance for classification in incomplete multivariate longitudinal data. They showed that incorporating the Kronecker product covariance structure for multiple outcomes resulted in more accurate prediction model estimates compared to unstructured covariance structure. These multivariate mixed-effects models that assume Kronecker variance-covariance matrix address the issue of small sample size. Generalized linear mixed-effects models have been extended in multivariate repeated measures studies for different outcomes (continuous, counts and binary)[1, 2, 14, 16].

Each of the mixed-effects models references mentioned above except Komarek et al.[4] and Hughes et al.[2, 16] assumed that the random effects follow a multivariate normal distribution. However, under misspecification of the random effects distribution, estimates of the model parameters may become seriously biased [47]. To make the model more robust against misspecification of the random-effects distribution, a normal mixture is assumed for the random

effects to obtain a robust model that is further used in the LDA procedure[2, 4, 16]. Few studies have assumed multivariate *t*-distribution and multinomial distribution for the random effects to identify different classes of subjects in multivariate repeated measures data[48-52]. Many of these mixed-effects models can be fitted using software packages such as the SAS procedure MIXED for linear models, GLIMMIX for generalized linear models and NLMIXED for nonlinear models.

There are several advantages in using multivariate mixed-effect models to analyze multivariate repeated measures data. Multivariate mixed-effect models are efficient for dealing with incomplete data. Because an individual's contribution to the likelihood function is calculated one subject at a time, it is possible to work with the available data for that subject, ignoring the missing data. Valid inferences can be obtained even with incomplete information, under missing at random (MAR) assumption. Mixed-effects models for multivariate repeated measures data have been extended to include variables of different types. This is because the relationship between the variables is handled through the random effects rather than the residuals. However, when the number of parameters increases with number of repeated measures and a collection of outcomes, the random-effects approaches are more likely to be computationally intensive and unstable[1, 4, 35]. In addition, it is difficult to evaluate the marginal likelihood of jointly generalized linear mixed-effects models when the outcome is non-normal.

Despite the variety of classification models developed for classification in multivariate repeated measures data, existing classifiers mostly relied on the assumption of multivariate normality. So far, there have been limited investigations of multivariate repeated measures classification models that are robust to model mis-specification and/or non-normal data distributions[44].

2.1.3 Generalized Estimating Equations

Generalized estimating equations (GEEs) are well-known marginal models used to analyze longitudinal discrete and continuous outcomes in clinical trials and biomedical studies[53-57]. GEE is used to characterize the marginal expectation of outcomes as a function of explanatory variables. While the mixed-effect model is an individual-level approach that adopts random effects to capture the correlation between the observations of the same subject[58], GEE is a population-level approach based on a quasi-likelihood function and provides the population estimates of the parameters[53, 59]. The parameter estimates of GEE are consistent and asymptotically normally distributed even when the "working" correlation structure of outcomes is mis-specified under mild regularity conditions.

Traditional GEEs have been extended to modeling of multiple outcomes[60]. In multivariate repeated measures binary data, Shelton et al. built a multivariate GEE approach SAS macro, and their method has been implemented in R package as well[61, 62]. Rochon[63] also applied the multivariate GEE model to simultaneously analyze a mixture of binary and continuous types of repeated measures. Using Kronecker product to decompose the working correlation matrix that captures between- and within-outcome relationships with a smaller number of correlation parameters, multivariate GEE models for multivariate repeated measures continuous data have been generalized to a class of multivariate GEE models for multivariate repeated measures data with same-type or mixed outcomes[64,65]. The multivariate GEE models has been implemented in an R package JGEE[65].

Multivariate GEE is an attractive approach because it relaxes the distribution assumption and only requires the correct specification of marginal means and variances as well as the link function which connects the covariates of interest and marginal means. Multivariate GEE can be implemented with several commonly available statistical software (e.g., SAS, Stata, S-Plus and

R)[66]. GEE methodology leads to valid inferences and claims its inferential optimality, when the missingness mechanism is missing completely at random (MCAR) (i.e. the complete cases can be viewed as a random sample from the underlying population). However, GEE can be inefficient due to ignored data especially in multivariate problems with highly arbitrary patterns of missing data[67, 68]. Weighted GEE and imputation approaches have proposed when the underlying data is missing at random (MAR). Despite the attractiveness of multivariate GEE, which is focused on its ability to analyze both discrete and continuous outcome variables, and accommodate different types of covariates, researchers have not explored repeated measures discriminant analysis based on multivariate GEE.

2.2 Covariance or Correlation structure Mis-specification

Multivariate repeated measures data comprise of two sources of variations (outcomes and occasions), and these variations must be taken into account while analyzing these kinds of data. A classical multivariate approach to the modeling and analysis of these data would be assuming unstructured covariance structure, yielding maximum-likelihood estimates provided the sample size is large. In this case, a large sample size is required for estimation of unstructured covariance structure since the number of unknown parameters to be estimated increases very rapidly with the increase in dimension of either the number of outcomes *q*, or the number of repeated occasions *p*[36-43]. "While it is robust not to assume knowledge of the covariance structure, this can result in rather weak inference in the sense that too many degrees of freedom are used up in estimating the covariance parameters, leaving too few for the parameters of interest"[69], Crowder and Hand (4,p.60). Therefore, for reasons of parsimony or more efficient mean estimation, or because the sample size may be insufficiently large for the covariance structure to be positive definite almost

surely, it may be desirable or necessary to assume that the covariance structure has a more structured form[70].

A common and natural structure to consider is the Kronecker product structure, i.e., $\boldsymbol{V} \otimes \boldsymbol{\Sigma}$, for two positive definite matrices $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$. Several authors have used this Kronecker product structure variance-covariance matrix in their analyses of multivariate repeated measures data[71-74] and in classification problems[10, 11, 75]. This Kronecker product structure separates the covariance structure of all the variables into covariance structures attributable to each of the two factors (outcomes and occasions), hence this structure is commonly said to be *separable*. Separability imposes a number of constraints on the variances and correlations among the observed variables[39]. Because separability imposes quite severe constraints on the covariance matrix, it may not hold for some datasets and it is important to test for this assumption[39, 76]. In addition, the choice of an appropriate covariance structure is crucial for multivariate repeated measures data in the context of classification since it increases the misclassification error rate. Thus, it is vital to test the appropriate covariance structure on the data before any statistical analysis[36].

In addition, hypotheses testing problems for multivariate repeated measures data using Kronecker product structure with both unstructured components have been widely studied by many authors[37, 39, 41-43], and Kronecker product structure with a CS or AR-1 correlation structures on the first component have also been widely studied to avoid identifiability problem[36, 40]. Likewise, Kronecker covariance structure assuming both components as structured CS or AR-1 which is useful for spatio-temporal repeated measurements have been studied[77]. For example, for modeling the covariance of multivariate environmental monitoring data obtained repeatedly over time and space, or for modeling covariance structure of glucose measurement at 15 different regions (p=15) in both hemispheres(q=2) of the brain[78].

All these authors used likelihood ratio test (LRT) statistic for testing separability of a covariance structure for multivariate repeated measures data. There are three types of LRT: (1) biased tests based on an asymptotic chi-square null distribution. The MIXED procedure of SAS software can be used to test the hypotheses for separable covariance structure with the first component as CS or AR-1 correlation or unstructured covariance structures using biased LRT[79]; (2) unbiased/modified LRT statistic in which the test statistic is modified in order to match the theoretical chi-square distribution to test the separability of variance–covariance structure; and (3) unbiased/unmodified tests based on an empirical null distribution (END). Hypotheses tests for separable structures are well developed area, and biased and unbiased/unmodified LRTs are available. However, the LRT statistic is reliable with very large samples, which may be limited in the real-life applications because we have only finite samples. Exploiting the ENDs of the LRT statistic overcome the problem of the accuracy of the asymptotic approximation under the null distribution of the unmodified LRT statistic for testing separable covariance structure for small or moderate sample sizes[39, 77]. However, the ENDs of the LRT statistics are quite different from their limiting chi-square distributions for small sample size. Therefore, the LRT fails in practical use because its distribution is very different from its limiting chi-square distribution for small samples. In addition, the LRT cannot be used for $n_j < pq$ for the unstructured variance-covariance matrix as alternative hypothesis. Nonetheless, researchers still use the theoretical chi-square distribution even for small samples as exact tests are not available in such cases.

In multivariate repeated measures data applications, one can fit linear models for a classification problem with separable covariance structure when $n_j < pq$ using MIXED procedure of SAS[11, 80]. However, before applying MIXED procedure of SAS for classification rules, one must test whether the data have separable covariance structure[80]. Unfortunately, all the above-mentioned

available LRT tests need the assumption $n_j > pq$, which is often not possible in applied setting given the limitations on data collection.

Rao's score test (RST) has been proposed as an alternative to LRT approach which avoids this limitation[38, 81]. The unmodified RST procedure test a separable covariance structure with the first component as a CS correlation matrix, which essentially means that all measurements for any characteristic within the same subject are equi-correlated. An advantage of RST is that it only exploits the null hypothesis, and thus does not need the assumption $n_j > pq$ as LRT does. When both components of the separable covariance structure are unstructured, the RST requires a sample size $n_j > \max(p, q)$, which can be large for many repeated measures (p)[38]. However, when the first separable component is the CS correlation structure, RST only requires a sample size $n_j > q$, which is independent of the number of repeated measures. Given the increasing collection of multivariate repeated measures data on which separability could be assessed, testing separability of a covariance structure using RST when $n_j > q$ is a substantial improvement over the LRT.

In quasi-likelihood framework, quasi-likelihood under the independent model information criterion (QIC)[82] was proposed as a modification of Akaike information criterion (AIC)[83] to select an appropriate working correlation structure among several candidates in GEE models. Similarly, to the AIC, the QIC is a trade-off between a good fit to the model (as measured by the quasi-likelihood), and a penalty for complexity measured by trace. However, QIC tends to favor the independence working structure, because the quasi-likelihood is formed under the working independence structure and hence utilizes little information about the correlation for GEE. As a remedy, Hin and Wang[84] suggested Correlation Information Criterion (CIC), that uses the penalty term in the QIC as a criterion for the selection of working correlation structure.

2.3 Statistical Robustness

Many assumptions commonly made in classical statistics such as normality, independence, and linearity are often not fulfilled in practice[85, 86]. Statistical procedures (in particular, those optimized for an underlying normal distribution) are excessively sensitive to seemingly minor deviations from the assumptions, and alternative "robust" procedures have been proposed [27, 86]. Statistical robustness signifies insensitivity to small deviations from idealized assumptions [85, 86]. In particular, distributional robustness of statistical estimators means insensitivity to small deviations from the shape of the true underlying distribution (usually normal) and tolerance to outliers [86]. The goals of robust estimators are: to describe the structure best fitting the bulk of the data; to identify and mitigate outliers and leverage points [85].

As mentioned by Peter Huber [27, 86]; robust, distribution-free, and non-parametric seem to be closely related properties but actually are not. Robust statistical estimators should not be confused with nonparametric estimators, although a few nonparametric procedures happen to be very robust [85, 87]. Robust statistics work in a "neighborhood" of parametric models[85]. Robust estimators consider that parametric models are only approximations to reality and are not only valid under strict parametric models but also in a neighborhood of such parametric models[85]. Therefore, robust estimators allow approximate fulfillment of strict assumptions, while nonparametric estimators make weak but strict assumptions (like symmetry and absolute continuity) [85, 87]. For example, the sample mean and the sample median are nonparametric estimates of the mean and the median, but the mean is not robust to outliers.

In estimation, optimality under the ideal model is commonly measured by the efficiency of the estimator, while near-optimality under contamination is displayed by measures of its resistance, or robustness[27, 86]. Robustness provides methods that trade-off some efficiency at the

ideal model to gain resistance against the effects of deviations [27, 86]. Robust estimators have optimal or nearly optimal efficiency at the assumed model, are resistant to small deviations from the model assumptions, and does not suffer a breakdown in case of large deviations [88].

2.3.1 Measures of Robustness

There are several measures of robustness of statistical estimators attempting to quantify the change, including the *influence function* and *breakdown point*. The influence function (IF) measures the effect of *an infinitesimally small fraction* of contamination on the estimator, hence *a local robustness measure*[85, 89]. A desirable property of the IF is boundedness. Boundedness ensures that a small fraction of contamination or outliers can have only a limited effect on the estimate or describes a function which does not go to infinity as the number of outliers become arbitrarily large[85, 88, 89]. The breakdown point of an estimator is defined as the proportion of outliers or contamination that an estimator can handle before becoming arbitrarily large (or breaks down) [90]. The higher the breakdown point of an estimator, the more robust the estimator. The breakdown point takes values from 0% to 50% [88]. The breakdown point cannot exceed 50% because if more than half of the observations are contaminated, then it is not possible to distinguish between the underlying distribution and the contaminating distribution [88, 91.] The breakdown point of the sample mean is zero, which means that a single outlier may throw the estimator completely off, while for the sample median it is 50%. Therefore, the median is a robust measure of central tendency while the mean is not [90, 92].

Estimation procedures that are robust to departures from multivariate normal distributions have been adopted for repeated measures prediction models. These include (a) transformation of data, and (b) robust estimators of parameters. Transformation is seen as an easy-to-implement

remedy to address the assumption of normality; however, applying a non-linear (eg: logarithmic, inverse) transformation to the outcome not only normalizes the residuals, but also distorts the scale of the transformed variable as well as alter the fundamental relationships among variables [93, 94]. Hence interpretation of the covariate effect on the transformed outcome can be complicated[93].

Robust estimators such as M-estimators [27, 28], S-estimators[29], minimum covariance determinant (MCD) estimators and minimum volume ellipsoid (MVE)[30] have been adopted to develop robust discriminant analysis for predicting population memberships, but for multivariate data collected at a single point (cross-sectional data) [95, 96] and univariate repeated measures data[97]. These robust estimators are especially useful in the multivariate normal model for the robust estimation of mean vectors $\boldsymbol{\mu}_j$ and covariance matrices $\boldsymbol{\Omega}_j$, even linear mixed models can be formalized as a multivariate normal model[88, 98]. The dispersion function for the multivariate normal model can be the determinant of the covariance matrix $\det(\boldsymbol{\Omega}_j) = |\boldsymbol{\Omega}_j|$ and Mahalanobis distances can be defined as

$$d_i = \sqrt{(\mathbf{y}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Omega}_j^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_j)} \tag{2.7}$$

The Mahalanobis distance is a natural measure of 'outlyingness' of an observation[91]. Consequently, and provided that the parameters $\boldsymbol{\mu}_j$ and $\boldsymbol{\Omega}_j$ in d_i are estimated robustly (example, via MCD and MVE), Mahalanobis distances can be used to detect multivariate outliers in that outliers correspond to large Mahalanobis distances[29, 88]. Therefore, high breakdown estimates such as S-estimators, MCD and MVE have been recommended for non-normal high dimensional data[29,30, 91, 99].

2.3.2 MCD and MVE Estimators

The minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) estimators introduced by Rousseeuw have received a considerable attention by scientific community and widely used in practice[30]. These are both affine equivariant estimators with bounded influence function properties and high breakdown points of ([n_j/2]-m + l)/ n_j approaching ½ for large n_j[91] and m-dimensional data. The MCD mean and covariance estimates are obtained from a subset of the data covering h ($h < n_j$) of the original data with the minimum covariance matrix determinant among all possible subsets of size h. A recommendable and common choice for h that yields maximum breakdown is: h= *[(n_j+m+1)/2]*, which asymptotically reaches half of the data, but any integer h within the interval [(n_j + m + 1)/2, n_j] can be chosen[91]. If $h = n_j$ then the MCD estimates reduce to the sample mean and covariance matrix of the full dataset. The MCD approach is similar to the MVE and has the same objective function. The only difference is in the constraint used where MCD minimizes the determinant of the covariance matrix based on the h data, while MVE minimizes the volume of the ellipsoid on h data. Thus, the MVE mean and covariance estimates are the centre and scatter of the ellipsoid with minimum volume covering at least h points of the data respectively. The MCD estimator is more attractive than MVE because it has a better convergence rate of $n_j^{-1/2}$compared to $n_j^{-1/3}$ of MVE[100, 101] and MCD gives the exact solution[102, 103].

Several different algorithms have been proposed in attempt to increase the computational efficiency of MVE and MCD because to obtain approximate values of these estimators is not only expensive but could be impossible for large sample sizes with large number of outcomes and

repeated occasions. These algorithms include: the Feasible Solution Algorithm (FSA) which is computationally heavy and relatively slow[104, 105] and the most commonly used algorithm, FAST-MCD[106] which obviate the need to examine all possible subsets of the data and the FAST algorithm has been modified for MVE[107]. The Fast algorithms are available in many computer packages such as MATLAB, R, SAS, and S-Plus.

2.4 Summary of Review

Marginal models, and mixed-effects models have been used to capture the dependencies in multivariate repeated measures data analyses in in the context of classification problems. The term marginal models include among others, covariance pattern models(CPM) and generalized estimating equations (GEE)[108]. The correlation structures induced by CPM and GEE are similar, but CPM apply only to continuous normally distributed outcome while GEE can be applied to a broad range of outcome variables often encountered in empirical applications (e.g., continuous, ordinal, polychotomous, dichotomous). Despite this attractive feature of GEE, researchers have not explored the application of GEE approach in discriminant analysis. Also, the selection of the correlation structure for the repeated measurements is not as critical for GEE as for mixed-effects models since the parameter estimates are consistent and asymptotically normally distributed even under mis-specified "working" correlation structure of outcomes in GEE approach. However, in the presence of missing data, GEE is only valid under the strong assumption of missing completely at random (MCAR)[109] but not missing at random assumption (MAR). However, mixed-effects models are efficient for dealing with incomplete data. Therefore, valid inferences can be obtained in mixed-effects models even with incomplete information under MAR. Multiple imputation GEEs and other approaches have been proposed as elegant ways to ensure validity of the inference under

MAR[109, 110]. A disadvantage of mixed-effects model is that the dimension of random-effects quickly increases as more outcomes and random-effects are added to the model, increasing the computational burden.

While it is crucial to model the dependencies in multivariate repeated measures data, estimation of all parameters will require a very large sample which may not always be feasible. To reduce the dimension of the unknown parameters of the covariance matrix, a more parsimonious structure such as a Kronecker product of the covariance matrix is sometimes used. To test the appropriateness of an assumed covariance structure on the data before any statistical analysis, many authors have widely studied hypotheses testing problems for multivariate repeated measures data using Kronecker product structure via likelihood tests [37, 39, 41-43] and Rao's score test [38, 81].

Finally, robust estimators have also been adopted to robustify marginal and mixed-effects models in the analysis of non-normal repeated measures data[96, 98, 111-116]. High breakdown robust estimates such as S-estimators, MVE and MCD have been recommended for non-normal high dimensional data instead of M-estimators. However, most of these robust estimators have been adopted to develop robust discriminant analysis for predicting group memberships in univariate repeated measures data[112-116] and have not been extended to multivariate repeated measures discriminant analysis.

References

1. Fieuws S, Verbeke G, Maes B, et al. Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics* 2007; 9: 419-431.
2. Hughes DM, Komárek A, Czanner G, et al. Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types. *Statistical methods in medical research* 2018; 27: 2060-2080.
3. Inoue LYT, Etzioni R, Morrell C, et al. Modeling disease progression with longitudinal markers. *Journal of the American Statistical Association* 2008; 103: 259-270.
4. Komárek A, Hansen BE, Kuiper EM, et al. Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statistics in medicine* 2010; 29: 3267-3283.
5. Li Y, Wang Y, Wu G, et al. Discriminant analysis of longitudinal cortical thickness changes in alzheimer's disease using dynamic and network features. *Neurobiology of aging* 2012; 33: 427. e415-427. e430.
6. Marshall G and Barón AE. Linear discriminant models for unbalanced longitudinal data. *Statistics in medicine* 2000; 19: 1969-1981.
7. Marshall G, De la Cruz-Mesía R, Barón AE, et al. Non-linear random effects model for multivariate responses with missing data. *Statistics in Medicine* 2006; 25: 2817-2830.
8. Marshall G, De la Cruz-Mesía R, Quintana FA, et al. Discriminant analysis for longitudinal data with multiple continuous responses and possibly missing data. *Biometrics* 2009; 65: 69-80.
9. Morrell CH, Brant LJ, Sheng S, et al. Screening for prostate cancer using multivariate mixed-effects models. *Journal of applied statistics* 2012; 39: 1151-1175.
10. Roy A and Khattree R. On discrimination and classification with multivariate repeated measures data. *Journal of Statistical Planning and Inference* 2005; 134: 462-485.
11. Roy A and Khattree R. Classification of multivariate repeated measures data with temporal autocorrelation. *J Appl Stat Sci* 2007; 15: 283-294.
12. Brant LJ, Sheng SL, Morrell CH, et al. Screening for prostate cancer by using random-effects models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2003; 166: 51-62.
13. De La Cruz-Mesia R and Quintana FA. A model-based approach to bayesian classification with applications to predicting pregnancy outcomes from longitudinal β-hcg profiles. *Biostatistics* 2006; 8: 228-238.
14. Fieuws S, Verbeke G and Molenberghs G. Random-effects models for multivariate repeated measures. *Statistical methods in medical research* 2007; 16: 387-397.
15. Hughes DM, Komárek A, Czanner G, et al. Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types. *Statistical methods in medical research* 2016: 0962280216674496.
16. Hughes DM, Komárek A, Bonnett LJ, et al. Dynamic classification using credible intervals in longitudinal discriminant analysis. *Statistics in medicine* 2017; 36: 3858-3874.
17. Ferro MA, Camfield CS, Levin SD, et al. Trajectories of health-related quality of life in children with epilepsy: A cohort study. *Epilepsia* 2013; 54: 1889-1897.
18. Speechley KN, Ferro MA, Camfield CS, et al. Quality of life in children with new-onset epilepsy: A 2-year prospective cohort study. *Neurology* 2012; 79: 1548-1555.

19. Morrell C, Brant L, Sheng S, et al. Using multivariate mixed-effects models to predict hypertension. In: *Proc Joint Stat Meeting Biometrics Sec* 2003, pp.2916-2921.
20. Filzmoser P, Joossens K and Croux C. Multiple group linear discriminant analysis: Robustness and error rate. *Compstat 2006-proceedings in computational statistics*. Springer, 2006, pp.521-532.
21. Croux C, Filzmoser P and Joossens K. Classification efficiencies for robust linear discriminant analysis. *Statistica Sinica* 2008: 581-599.
22. Fisher RA. The use of multiple measurements in taxonomic problems. *Annals of eugenics* 1936; 7: 179-188.
23. Lachenbruch PA and Goldstein M. Discriminant analysis. *Biometrics* 1979: 69-85.
24. Marks S and Dunn OJ. Discriminant functions when covariance matrices are unequal. *Journal of the American Statistical Association* 1974; 69: 555-559.
25. Flury BW and Schmid MJ. Quadratic discriminant functions with constraints on the covariance matrices: Some asymptotic results. *Journal of multivariate analysis* 1992; 40: 244-261.
26. Wahl PW and Kronmal RA. Discriminant functions when covariances are unequal and sample sizes are moderate. *Biometrics* 1977: 479-484.
27. Huber PJ. Robust estimation of a location parameter. *The annals of mathematical statistics* 1964; 35: 73-101.
28. Maronna RA. Robust m-estimators of multivariate location and scatter. *The annals of statistics* 1976: 51-67.
29. Rousseeuw P and Yohai V. Robust regression by means of s-estimators. *Robust and nonlinear time series analysis*. Springer, 1984, pp.256-272.
30. Rousseeuw PJ. Multivariate estimation with high breakdown point. *Mathematical statistics and applications* 1985; 8: 37.
31. Roy A. A new classification rule for incomplete doubly multivariate data using mixed effects model with performance comparisons on the imputed data. *Statistics in medicine* 2006; 25: 1715-1728.
32. Tomasko L, Helms RW and Snapinn SM. A discriminant analysis extension to mixed models. *Statistics in medicine* 1999; 18: 1249-1260.
33. Albert A. Discriminant analysis based on multivariate response curves: A descriptive approach to dynamic allocation. *Statistics in medicine* 1983; 2: 95-106.
34. Roy A and Khattree R. Classification of multivariate repeated measures data with temporal autocorrelation. *Journal of Applied Statistical Science* 2007; 15: 283-294.
35. Verbeke G, Fieuws S, Molenberghs G, et al. The analysis of multivariate longitudinal data: A review. *Statistical methods in medical research* 2014; 23: 42-59.
36. Roy A and Khattree R. On implementation of a test for kronecker product covariance structure for multivariate repeated measures data. *Statistical Methodology* 2005; 2: 297-306.
37. Srivastava MS, von Rosen T and Von Rosen D. Models with a kronecker product covariance structure: Estimation and testing. *Mathematical Methods of Statistics* 2008; 17: 357-370.
38. Filipiak K, Klein D and Roy A. Score test for a separable covariance structure with the first component as compound symmetric correlation matrix. *Journal of Multivariate Analysis* 2016; 150: 105-124.
39. Lu N and Zimmerman DL. The likelihood ratio test for a separable covariance matrix. *Statistics & probability letters* 2005; 73: 449-457.

40. Roy A and Khattree R. Testing the hypothesis of a kronecker product covariance matrix in multivariate repeated measures data. *SAS Users Group International, Proceedings of the Statistics and Data Analysis Section* 2005: 199-130.
41. Roy A. A note on testing of kronecker product covariance structures for doubly multivariate data. In: *Proceedings of the American Statistical Association, Statistical Computing Section* 2007, pp.2157-2162.
42. Roy A and Khattree R. Tests for mean and covariance structures relevant in repeated measures based discriminant analysis. *Journal of Applied Statistical Science* 2003; 12: 91-104.
43. Werner K, Jansson M and Stoica P. On estimation of covariance matrices with kronecker product structure. *IEEE Transactions on Signal Processing* 2008; 56: 478-491.
44. Cho H. The analysis of multivariate longitudinal data using multivariate marginal models. *Journal of Multivariate Analysis* 2016; 143: 481-491.
45. Fitzmaurice GM, Laird NM and Ware JH. *Applied longitudinal analysis*. John Wiley & Sons, 2012.
46. Marshall G, la Cruz-Mesía D, Quintana FA, et al. Discriminant analysis for longitudinal data with multiple continuous responses and possibly missing data. *Biometrics* 2009; 65: 69-80.
47. Litière S, Alonso A and Molenberghs G. 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* 2008; 27: 3125-3144.
48. Muthén BO. Beyond sem: General latent variable modeling. *Behaviormetrika* 2002; 29: 81-117.
49. Nagin DS. Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological methods* 1999; 4: 139.
50. Nagin DS and Land KC. Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed poisson model. *Criminology* 1993; 31: 327-362.
51. Nagin DS and Tremblay RE. Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological methods* 2001; 6: 18.
52. Thum YM. Hierarchical linear models for multivariate outcomes. *Journal of Educational and Behavioral Statistics* 1997; 22: 77-108.
53. Crowder M. On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* 1995; 82: 407-410.
54. Friedman LM, Furberg CD, DeMets DL, et al. *Fundamentals of clinical trials*. Springer, 2015.
55. Liang K and Zeger S. A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrika* 1986; 73: 13-22.
56. Liang K-Y and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73: 13-22.
57. Zeger SL and Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986: 121-130.
58. Lu B, Preisser JS, Qaqish BF, et al. A comparison of two bias-corrected covariance estimators for generalized estimating equations. *Biometrics* 2007; 63: 935-941.
59. Wedderburn RW. Quasi-likelihood functions, generalized linear models, and the gauss—newton method. *Biometrika* 1974; 61: 439-447.
60. O'Brien LM and Fitzmaurice GM. Analysis of longitudinal multiple-source binary data using generalized estimating equations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2004; 53: 177-193.

61. Shelton BJ, Gilbert GH, Liu B, et al. A sas macro for the analysis of multivariate longitudinal binary outcomes. *Computer Methods and Programs in Biomedicine* 2004; 76: 163-175.
62. Asar Ö and İlk Ö. Mmm: An r package for analyzing multivariate longitudinal data with multivariate marginal models. *Computer Methods and Programs in Biomedicine* 2013; 112: 649-654.
63. Rochon J. Analyzing bivariate repeated measures for discrete and continuous outcome variables. *Biometrics* 1996: 740-750.
64. Gao F, Thompson P, Xiong C, et al. Analyzing multivariate longitudinal data using sas. In: *Proceedings of the Thirty-first Annual SAS Users Group International Conference* 2006, pp.187-131. Citeseer.
65. Inan G. Jgee: Joint generalized estimating equation solver. R package version, 2015.
66. Horton NJ and Lipsitz SR. Review of software to fit generalized estimating equation regression models. *The American Statistician* 1999; 53: 160-169.
67. Kang J and Yang Y. Joint modeling of mixed count and continuous longitudinal data. *Analysis of mixed data*. Chapman and Hall/CRC, 2013, pp.91-108.
68. Lipsitz SR, Fitzmaurice GM, Ibrahim JG, et al. Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: An application to acquired immune deficiency syndrome data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009; 172: 3-20.
69. Crowder MJ and Hand DJ. *Analysis of repeated measures*. CRC Press, 1990.
70. Wolfinger RD. Heterogeneous variance: Covariance structures for repeated measures. *Journal of agricultural, biological, and environmental statistics* 1996: 205-230.
71. Chaganty NR and Naik DN. Analysis of multivariate longitudinal data using quasi-least squares. *Journal of Statistical Planning and Inference* 2002; 103: 421-436.
72. Naik DN and Rao SS. Analysis of multivariate repeated measures data with a kronecker product structured covariance matrix. *Journal of Applied Statistics* 2001; 28: 91-105.
73. Boik RJ. Scheffés mixed model for multivariate repeated measures: A relative efficiency evaluation. *Communications in Statistics-Theory and Methods* 1991; 20: 1233-1255.
74. Galecki AT. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics-Theory and Methods* 1994; 23: 3105-3119.
75. Roy A and Khattree R. Discrimination and classification with repeated measures data under different covariance structures. *Communications in Statistics—Simulation and Computation®* 2005; 34: 167-178.
76. Lu N and Zimmerman D. On likelihood-based inference for a separable covariance matrix. *Statistics and Actuarial Science Dept, Univ of Iowa, Iowa City, IA, Tech Rep* 2004; 337.
77. Roy A and Leiva R. Likelihood ratio tests for triply multivariate data with structured correlation on spatial repeated measurements. *Statistics & Probability Letters* 2008; 78: 1971-1980.
78. Worsley K, Evans A, Strother S, et al. A linear spatial correlation model, with applications to positron emission tomography. *Journal of the American statistical association* 1991; 86: 55-67.
79. Khattree R and Naik DN. *Applied multivariate statistics with sas software*. SAS Institute Inc., 2018.
80. Roy A and Khattree R. Classification rules for repeated measures data from biomedical research. *Computational Methods in Biomedical Research* 2007: 323-370.

81. Filipiak K, Klein D and Roy A. A comparison of likelihood ratio tests and rao's score test for three separable covariance matrix structures. *Biometrical Journal* 2017; 59: 192-215.
82. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics* 2001; 57: 120-125.
83. Akaike H. Information theory and an extension of the maximum likelihood principle. *Selected papers of hirotugu akaike*. Springer, 1998, pp.199-213.
84. Hin LY and Wang YG. Working-correlation-structure identification in generalized estimating equations. *Statistics in medicine* 2009; 28: 642-658.
85. Hampel F, Ronchetti E, Rousseeuw P, et al. Robust statistics, j. *Wiley& Sons, New York* 1986.
86. Huber PJ. Robust statistics. 1981. Wiley, New York.
87. Hampel FR, Ronchetti EM, Rousseeuw PJ, et al. *Robust statistics: The approach based on influence functions*. John Wiley & Sons, 2011.
88. Heritier S, Cantoni E, Copt S, et al. *Robust methods in biostatistics*. John Wiley & Sons, 2009.
89. Hampel FR. The influence curve and its role in robust estimation. *Journal of the american statistical association* 1974; 69: 383-393.
90. Hampel FR. A general qualitative definition of robustness. *The Annals of Mathematical Statistics* 1971: 1887-1896.
91. Rousseuw PJ and Leroy AM. Robust regression and outlier detection. Wiley, New York, 1987.
92. Donoho DL and Huber PJ. The notion of breakdown point. *A festschrift for Erich L Lehmann* 1983; 157184.
93. Lipsitz SR, Ibrahim J and Molenberghs G. Using a box–cox transformation in the analysis of longitudinal data with incomplete responses. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2000; 49: 287-296.
94. Stevens SS. On the theory of scales of measurement. 1946.
95. Croux C and Dehon C. Robust linear discriminant analysis using s-estimators. *Canadian Journal of Statistics* 2001; 29: 473-493.
96. He X and Fung WK. High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis* 2000; 72: 151-162.
97. Hubert M and Van Driessen K. Fast and robust discriminant analysis. *Computational Statistics & Data Analysis* 2004; 45: 301-320.
98. Copt S and Victoria-Feser M-P. High-breakdown inference for mixed linear models. *Journal of the American Statistical Association* 2006; 101: 292-300.
99. Hawkins DM and McLachlan GJ. High-breakdown linear discriminant analysis. *Journal of the American Statistical Association* 1997; 92: 136-143.
100. Butler R, Davies P and Jhun M. Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics* 1993: 1385-1400.
101. Croux C and Haesbroeck G. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis* 1999; 71: 161-190.
102. Hadi AS. Identifying multiple outliers in multivariate data. *Journal of the Royal Statistical Society: Series B (Methodological)* 1992; 54: 761-771.
103. Hubert M, Rousseeuw PJ and Van Aelst S. Multivariate outlier detection and robustness. *Handbook of Statistics* 2005; 24: 263-302.

104. Hawkins DM. The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data. *Computational Statistics & Data Analysis* 1994; 17: 197-210.
105. Hawkins DM and Olive DJ. Improved feasible solution algorithms for high breakdown estimation. *Computational statistics & data analysis* 1999; 30: 1-11.
106. Rousseeuw PJ and Driessen KV. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 1999; 41: 212-223.
107. Maronna RA, Martin RD and Yohai VJ. *Robust statistics: Theory and methods (with r)*. John Wiley & Sons, 2019.
108. Fitzmaurice G, Davidian M, Verbeke G, et al. *Longitudinal data analysis*. CRC press, 2008.
109. Robins JM, Rotnitzky A and Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association* 1995; 90: 106-121.
110. Little RJ and Rubin DB. *Statistical analysis with missing data*. John Wiley & Sons, 2019.
111. Cantoni E and Ronchetti E. Robust inference for generalized linear models. *Journal of the American Statistical Association* 2001; 96: 1022-1030.
112. Wang YG, Lin X and Zhu M. Robust estimating functions and bias correction for longitudinal data analysis. *Biometrics* 2005; 61: 684-691.
113. Preisser JS and Qaqish BF. Robust regression for clustered data with application to binary responses. *Biometrics* 1999; 55: 574-579.
114. He X, Fung WK and Zhu Z. Robust estimation in generalized partial linear models for clustered data. *Journal of the American Statistical Association* 2005; 100: 1176-1184.
115. Pinheiro JC, Liu C and Wu YN. Efficient algorithms for robust estimation in linear mixed-effects models using the multivariate t distribution. *Journal of Computational and Graphical Statistics* 2001; 10: 249-276.
116. Sinha SK. Robust analysis of generalized linear mixed models. *Journal of the American Statistical Association* 2004; 99: 451-460.

Chapter 3

Robust Trimmed Likelihood Discriminant Analysis for Multivariate Repeated Data

Anita Brobbey A., Lix LM., Nettel-Aguirre A., Williamson T., Wiebe S., Sajobi T., Robust Trimmed Likelihood Discriminant Analysis for Multivariate Repeated Data. *Journal of Modern Applied Statistical Methods* (Minor revision requested) (under review).

AB's contribution to this manuscript includes development of new robust discriminant analysis procedures, design and implementation of simulation study, manuscript preparation, and analysis of example dataset. LML, ANA, TW and TS provided statistical and methodological expertise support. SW revised the manuscript for important clinical content and insights. All co-authors reviewed results and revised the manuscript for important intellectual content. AB assumes responsibility for the integrity of the manuscript. This manuscript in its entirety is included in Chapter 3

Abstract

Repeated measures discriminant analyses have been developed for distinguishing between two or more independent populations in multivariate repeated measures designs, in which multiple outcomes are repeatedly measured at two or more measurement occasions. However, these models, which are based on structured covariances, rely on the assumption of multivariate normality. This study developed repeated measures linear discriminant analysis (RMLDA) and repeated measures quadratic discriminant analysis (RMQDA) based on maximum trimmed likelihood estimators (MTLE) for classifying repeatedly measured observations characterized by multivariate non-normal distributions. Monte Carlo methods were used to compare the accuracy of repeated measures discriminant analysis procedures based on maximum likelihood estimators (MLE) and MTLE under a variety of simulation generation conditions, including population distribution, covariance structure, covariance heterogeneity, between-variable and within-variable correlations and number of outcome variables. There were negligible differences in the mean accuracy of repeated measures discriminant analysis based on MLE and MTLE when the data were sampled from multivariate normal distribution. But, the MTLE procedures had between 2% and 13% higher overall mean classification accuracy than MLE procedures for multivariate heavy-tailed distributions. Repeated measures discriminant analysis based on robust estimators are recommended for discriminating between population in multivariate repeated measures designs characterized by non-normal distributions.

Keywords*:* Repeated measures, longitudinal data, robust methods, covariance structure, trimmed estimators, non-normality, Outliers

3.1 Introduction

Multivariate repeated measures data, in which multiple outcomes are repeatedly measured at two or more occasions, are commonly collected in several disciplines including medicine, ecology, and environmental sciences, where investigators seek to understand changes in multiple correlated outcomes over time or different occasions[1-6]. Multivariate repeated measures data are particularly useful for studying evolutions in subjects' outcomes over time on multiple characteristics[7]. For example, Fieuws and Verbeke [1] reported data on a cohort of patients who having undergone kidney transplant, were longitudinally monitored at irregularly spaced intervals over a 10 year period. The repeated collection of multiple biochemical and physiological markers, which constitute multivariate repeated measures data, were used to predict 10-year success of graft. Multivariate repeated measures data are inherently challenging to analyze because they are typically characterized by non-normal distributions, and high-dimensional data. Moreover, such data have complex correlation structures[8, 9]. Classical classification and prediction models developed for data collected in a cross-sectional study are not appropriate to address the complexities observed in multivariate repeated measures data [8, 9] .

Repeated measures discriminant analysis, which assume parsimonious mean and covariance structures, have been proposed for discriminating between population groups in multivariate repeated measures data. These procedures have been primarily developed based on mixed-effects regression models, covariance pattern models, and growth curve models[10-14] . For example, Roy and Khattree developed repeated measures discriminant analysis procedures based on structured means and Kronecker product variance-covariance matrix of unstructured between-outcome correlation matrix and compound symmetric (CS) or first-order autoregressive (AR-1) within-outcome correlation for predicting population membership in multivariate repeated

measures data[12, 13]. One approach that has been widely used in applied behavioral research is growth curve modeling analysis. Discriminant analysis have been extended to the study of multivariate outcome curves that can be used to classify a given patient's response curve (example: linear and quadratic shape) to the prognostic group it resembles most[15]. However, misspecification of the functional form of the growth curve can potentially lead to biased parameter estimates, misleading conclusions and lower accuracy[16, 17]. Similarly, repeated measures discriminant analysis have been developed based on multivariate linear and non-linear mixed-effects models that assume no structure [1, 18] and a Kronecker product structure [6, 19-21] for the within-outcome and between-outcome covariance matrices. Repeated measures discriminant analysis based on mixed-effects models are known to be advantageous in that they can accommodate time-varying and time-invariant covariates in addition to the longitudinally measured outcomes to improve classification accuracy. Generalized linear mixed models have been extended in multivariate repeated measures data studies for different type outcomes (continuous, counts and binary) [1, 2, 22]. However, when the number of parameters increases with number of outcomes and repeated occasions, the random-effect approaches are more likely to be computationally intensive and unstable. In addition, under misspecification of the common assumption of multivariate normal distribution for random effect parameters, estimates of the mixed-effects model parameters may become seriously biased[25] and consequently, the performance of the discriminant procedure may also be affected[2].

Most repeated measures discriminant analysis procedures assume the data are sampled from a multivariate normal distribution, which may not be tenable. Multivariate repeated measures data are frequently characterized by multivariate skewed or heavy-tailed distributions[23]. So far, there has been limited investigations of repeated measures discriminant analysis procedures that are robust (i.e., insensitive) to departures from the assumptions of multivariate normality for

discriminating between population groups in multivariate non-normal repeated measures data. The lack of these repeated measures discriminant analysis procedures that are robust to violation of the distributional assumptions has limited their application to several applied research settings where multivariate repeated measures data are routinely collected (e.g., cancer screening).

This study develops robust discriminant analysis models for multivariate non-normal repeated measures data. Specifically, we examined the accuracy of repeated measures discriminant analysis based on maximum trimmed likelihood estimation (MTLE) methods[24, 25] under a variety of simulation generation conditions. The manuscript is organized as follows. In section 3.2, we describe the mathematical framework for repeated measures discriminant analysis procedures and their robust extensions. The results of a Monte Carlo simulation study, which is designed to assess the performance of repeated measures discriminant analysis based on MLE and MTLE are discussed in section 3.3. Data from the Manitoba Inflammatory Bowel Disease (IBD) cohort study were used to demonstrate the application of these repeated measures discriminant analysis procedures in section 3.4, while a discussion of the key findings and their implications are described in section 3.5.

3.2 Repeated Measures Data Analysis

Let $\mathbf{y}_{ij} = (\mathbf{y}_{ij1}, \mathbf{y}_{ij2}, \ldots, \mathbf{y}_{ijq})$ be a pq x 1 vector of q outcomes, each repeatedly measured at p occasions for the ith individual in the jth population, sampled from a multivariate normal distribution such that $\mathbf{y}_{ij} \sim N_{pq}(\boldsymbol{\mu}_j, \boldsymbol{\Omega}_j)$, where $\boldsymbol{\mu}_j$ and $\boldsymbol{\Omega}_j$ is assumed to be pq x 1 mean vector and pq x pq positive definite covariance matrix respectively. When $\boldsymbol{\mu}_j$ and $\boldsymbol{\Omega}_j$ are unknown and completely unspecified, a total of $pq + pq(pq+1)/2$ unknown parameters must be estimated. This number increases very rapidly as p and q increases. Estimation of so many parameters will require a very large sample, which may not always be feasible. A parsimonious approach to parameter

estimation is to assume that $\mathbf{\Omega}_j$ has a Kronecker product structure: $\mathbf{\Omega}_j = \mathbf{V}_j \otimes \mathbf{\Sigma}_j$, where $\mathbf{V}_j$ and $\mathbf{\Sigma}_j$ are $p \times p$ and $q \times q$ positive definite matrices respectively, and $\otimes$ denotes the Kronecker product function[12, 13]. The matrix $\mathbf{V}_j$ is the correlation matrix of the repeated measures on a given outcome variable and it is assumed to be the same for all outcome variables. The matrix $\mathbf{\Sigma}_j$ represents the variance-covariance matrix between the measurements on all outcome variables at a given time point and this is assumed constant for all time points. Suppose that no structures whatsoever are assumed on $\mathbf{V}_j$ and $\mathbf{\Sigma}_j$ except that they are positive definite matrices; then, the classifier has the form

$$\lambda(\mathbf{y}_i) = \arg\max_j \ln\left(\pi_j f_j(\mathbf{y}_i)\right) \tag{3.1}$$

where

$$f_j(\mathbf{y}_i) = (2\pi)^{-\frac{pq}{2}} |\mathbf{V}_j|^{-\frac{p}{2}} |\mathbf{\Sigma}_j|^{-\frac{q}{2}} \exp\left[-\frac{1}{2} D_j(\mathbf{y}_i)\right] \tag{3.2}$$

π_j is the prior probability that an observation $\boldsymbol{y}_i$ is from class j, and $D_j(\mathbf{y}_i) = (\mathbf{y}_i - \boldsymbol{\mu}_j)' (\boldsymbol{V}_j^{-1} \otimes \mathbf{\Sigma}_j^{-1})(\mathbf{y}_i - \boldsymbol{\mu}_j)$ is the squared Mahalanobis distance between the multiple outcome vector $\boldsymbol{y}_i$ and the population mean, $\boldsymbol{\mu}_j$. The parameters $\boldsymbol{\mu}_j$, $\mathbf{V}_j$ and $\mathbf{\Sigma}_j$ are unknown and should be estimated, relying on training samples from the different populations. Specifically, discriminant analysis rule for two populations allocate $\boldsymbol{y}_i$ to population 1 if $\hat{\lambda}_{12}(\boldsymbol{y}_i) \leq 0$ where

$$\hat{\lambda}_{12}(\mathbf{y}_i) = D_1^*(\mathbf{y}_i) - D_2^*(\mathbf{y}_i) + 2 \log \frac{\hat{\pi}_2}{\hat{\pi}_1} \tag{3.3}$$

with $D_j^*(\mathbf{y}_i) = (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_j)' \, \hat{\mathbf{\Omega}}_j^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}}_j) + \log |\, \hat{\mathbf{\Omega}}_j \,|$ and $\hat{\pi}_1$ and $\hat{\pi}_2$ are the *a priori* probabilities that observations belong to populations 1 and 2, respectively. For the conventional discriminant analysis, the parameters $\boldsymbol{\mu}_j$, $\mathbf{V}_j$ and $\mathbf{\Sigma}_j$ are estimated using MLE. Based on the choice of covariance structures, estimates of the Mahalanobis distance and classification rule can be derived[11, 12]. The

homoscedastic model is obtained when the covariance components are homogeneous, that is, $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2 = \boldsymbol{\Omega}$, the pooled covariance matrix for j = 1,2. The above classifier implies classification of a subject with multiple outcome vector $\mathbf{y}_i$ in the first population, if and only if

$$\left(\mathbf{y}_i - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2}\right)' (\mathbf{V}^{-1} \otimes \boldsymbol{\Sigma}^{-1})(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) > \log \frac{\pi_2}{\pi_1} \tag{3.4}$$

which is the linear discriminant analysis (LDA) function, and quadratic discriminant analysis (QDA) function when $\mathbf{V}_1 \neq \mathbf{V}_2$ as

$$(\mathbf{y}_i - \boldsymbol{\mu}_2)'(\mathbf{V}_2^{-1} \otimes \boldsymbol{\Sigma}_2^{-1})(\mathbf{y}_i - \boldsymbol{\mu}_2) - (\mathbf{y}_i - \boldsymbol{\mu}_1)'(\mathbf{V}_1^{-1} \otimes \boldsymbol{\Sigma}_1^{-1})(\mathbf{y}_i - \boldsymbol{\mu}_1) \tag{3.5}$$

$$> \log \left|\frac{\boldsymbol{\Omega}_1}{\boldsymbol{\Omega}_2}\right| + 2 \log \frac{\pi_2}{\pi_1}$$

However, conventional repeated measures discriminant analysis procedures rely on the assumption of multivariate normal distribution, which may not be tenable in multivariate repeated measures data, which are usually characterized by non-normal distributions[2, 26].

3.2.1 Robust Repeated Measures Discriminant Analysis

An alternate approach to overcome these limitations in conventional repeated measures discriminant analysis procedures involves the development of robust repeated measures discriminant analysis procedures based on maximum trimmed likelihood estimators (MTLE) of mean, $\boldsymbol{\mu}_j$ and covariance components, $\mathbf{V}_j$ and $\boldsymbol{\Sigma}_j$. In MTLE, the contribution of each $\boldsymbol{y}_i$ to the (log)likelihood function scores ($\ell(\boldsymbol{\theta}; \boldsymbol{y}_i)$ are ranked from the smallest to the highest based estimated parameters and the loglikelihood function scores at the extreme tails are assigned smaller or no weights. Depending on the weights assigned to observations at the tails, different robust estimators could be derived. Specifically, in this study we developed robust repeated measures discriminant analysis based on minimum covariance determinant (MCD) and minimum volume

ellipsoid (MVE) estimators, which are special cases of MTLE[24, 25]. For any given value of $\boldsymbol{\theta}$, there exists an ordering of $\boldsymbol{y}$ of individuals such that,

$$\ell(\boldsymbol{\theta}; \boldsymbol{y}_1) \geq \ell(\boldsymbol{\theta}; \boldsymbol{y}_2) \geq \cdots \geq \ell(\boldsymbol{\theta}; \boldsymbol{y}_n) \tag{3.6}$$

where $\ell(\boldsymbol{\theta}; \boldsymbol{y}_i) = \ln f(\boldsymbol{y}_i; \boldsymbol{\theta})$ is the contribution of the *i*th observation to the log-likelihood function. Note, the original indices of the observations may not satisfy the likelihood ordering in (3.6) for all values of $\boldsymbol{\theta}$. If, for a given value of $\boldsymbol{\theta}$, the above ordering is not satisfied, the indices of the observations can be changed so that (3.6) is satisfied[24, 25]. The ordering of the observations may be different for different values of $\boldsymbol{\theta}$. The trimmed log likelihood function is given as

$$\sum_{i=1}^{h} \ell(\boldsymbol{\theta}; \boldsymbol{y}_i) \tag{3.7}$$

where h is the trimming parameter. The MTLE $\widehat{\boldsymbol{\theta}}(h)$ is obtained by maximizing the trimmed log likelihood function. The key idea is to trim the $n - h$ points that are the most unlikely from the estimation of the likelihood function. Special cases of MTLE includes MLE, MCD, and MVE. When $h = n$, $\widehat{\boldsymbol{\theta}}(n)$, we obtain the MLE of $\boldsymbol{\theta}$ and for $h < n$, the MCD and MVE estimators are MTLEs of $\boldsymbol{\theta} = (\widehat{\boldsymbol{\mu}}_h, \widehat{\boldsymbol{\Omega}}_h)$ for h observations yielding the desired robust estimates [24, 27]. The parameter h has to be set manually and h can be taken as low as (n /2) + 1. The farther h is from n, the more robust but the less efficient are the estimators. The MCD approach is similar to the MVE in that it searches for a portion of the data that minimizes the impact of outlying observations on estimation of means and covariance parameters. However, MVE seeks to minimize the volume of an ellipsoid created by the retained data, whiles MCD minimizes the determinant of the variance-covariance matrix. For example; the location estimate of MVE is the center of the minimum volume ellipsoid covering (at least) h of the data whiles for MCD, the location estimate is the mean of h of the data for which the determinant of the covariance matrix is minimal.

3.3 Simulation Study

A Monte Carlo simulation study was conducted to examine the accuracy of robust repeated measures discriminant analysis procedures in comparison to the conventional repeated measures discriminant analysis based on MLE estimators. Specifically we investigated the following procedures: (a) Repeated measures discriminant analysis that assumes structured means and Kronecker correlation matrix of unstructured between-outcomes and within-outcome AR-1 correlation matrices (st-UNAR), (b) Repeated measures discriminant analysis that assumes unstructured means and Kronecker correlation matrix of unstructured between-outcomes and within-outcome AR-1 correlation matrices (un-UNAR), (c) Repeated measures discriminant analysis that assumes structured means and Kronecker correlation matrix of unstructured between-outcomes and within-outcome CS correlation matrices (st-UNCS) and (d) Repeated measures discriminant analysis that assumes unstructured means and Kronecker correlation matrix of between-outcomes and within-outcome CS correlation matrices (un-UNCS). The parameters of each repeated measures discriminant analysis procedure were estimated using MLE, MVE and MCD estimators. Moreover, repeated measures LDA was used for classification when population covariances were assumed homogeneous, while repeated measures QDA was adopted when population covariances were assumed heterogeneous.

The following simulation generation conditions were investigated: (a) number of different outcomes (q), (b) number of repeated occasions (p), (c) total sample size (n), (d) population sizes (n_1, n_2), € Covariance pattern and magnitude of correlation among the repeated measurements (ρ), (f) mean configuration, (g) Covariance heterogeneity, and (h) population distribution. All procedures were investigated for two independent groups. The number of repeated measurements was set at $p = 3$, and 5 whilst the number of different outcomes was set at $q = 3$, and 7. Previous

studies about repeated measures discriminant analysis procedures have considered p ranging from 3 to 10, an increase in classification accuracy was quite significant when p increases from three to five[11, 12]. Total sample sizes of $n = 100$, 140 and 200 were investigated. This is consistent with previous simulation studies that examined the accuracy of repeated measures discriminant analysis based on parsimonious covariance structures with n ranging between 60 and 200. Moreover, consistent with previous studies, we examined the impact of equal and unequal group sizes [11, 12,28, 29, 30 31]. For $n = 100$, we set $(n_1, n_2) = (50, 50)$, and $(40, 60)$. Similar equal (1:1) and unequal (2:3) group size ratios were investigated when $n = 140$ and 200. A variety of mean configurations of different forms have been previously investigated in the development of repeated measures discriminant analysis procedures [11, 12]. In this study, four configurations for $\boldsymbol{\mu}_1$ were selected for each pair of p and q. (see Table 3.1) and $\boldsymbol{\mu}_2$ was the null vector for all conditions[11, 12, 32]. The descriptions of the four configurations for $\boldsymbol{\mu}_1$ in Table 3.1 are as follows; configuration I-III had no change in mean pattern over time for constant, monotonic increasing and quadratic mean patterns among repeated outcomes respectively, and configuration IV was assumed to have monotonic increasing mean pattern over time for non-constant means among the repeated measurements.

Furthermore, the accuracy of repeated measures discriminant analysis procedures is known to be influenced by the magnitude and pattern of within- and between-outcome correlations[33]. Therefore, we investigated the following components of the assumed Kronecker variance-covariance matrices: $\boldsymbol{\Omega}_j = \mathbf{V}_j \otimes \boldsymbol{\Sigma}_j$, $j = 1, 2$ where $\boldsymbol{\Sigma}_j$ was assumed to be a $q \times q$ unstructured variance-covariance matrix with a common variance of 60 among outcome variables and the $p \times p$ correlation matrix $\mathbf{V}_j = \mathbf{V}_j(\rho_j)$ was assumed to follow a AR-1 or CS structure with ρ_j chosen as

0.3 and 0.7, representing moderate to strong autocorrelation in the data[11, 12] (See Table 3.2 for more details).

In order to assess the performance of the discriminant function, random samples were generated from both multivariate normal and multivariate non-normal distributions. With specified mean, $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Omega}_j$, *pq*-variate normal distribution populations were generated using the mvrnorm() function from the **MASS** R package[34], multivariate lognormal distribution were generated using the rlnorm() function from the **compositions** R package [35], multivariate t distribution were generated using the rmvt() function from the **mvnfast** R package[36], and multivariate Cauchy distribution were generated using the rmsc() function from the **sn** R package [36]. For robust estimators, the proportion of trimmed data was fixed at 10% symmetric trimming. Some researchers have investigated 5%-25% trimming[37-39]. Even though one study argues 20% trimming[38], another recommend no more than 10% trimming to achieve optimal results[39]. The FASTMCD and FASTMVE algorithms[40-42] were used to define a subsample of observations for the trimmed means and covariances. More specifically, robust estimates of the LDA and QDA procedures that assumed unstructured or structured means and structured covariances were derived by maximizing the likelihood of the 90% best subsample of original observations using the fast algorithms. These means and covariance have high robustness properties[40, 43].

Fixed-effects analysis of variance (ANOVA) model was used to assess the relative importance of different simulation factors on the variations in the average classification accuracy for each procedure[44, 45]. The percentage of explained variance attributable to each main effect and interactions were evaluated using η^2, an R^2 equivalent in regression analysis[46]. The classification was performed on the generated samples from each of the two populations. Previous classification research have employed the classification accuracy or the error rate (1-accuracy) metric to

discriminate between two or more groups[11, 12, 28]. Thus, the overall classification accuracy (correctly classified / total sample) was used as performance metric in this study and the standard errors were also calculated. For each procedure and each method of estimation, a total of 1440 combination of simulation factors was investigated. There were 1000 replications for each combination. The Monte Carlo study was conducted using R version 3.6.3.

3.3.1 Simulation Study Results

Simulation factors such as population distribution, and mean configurations accounted for most of the variation in the classification accuracies (Table 3.3), with proportion of explained variation ranging 40.8 - 56% and 9.2 - 11% respectively. In addition, the estimation methods (MLE and MVE) accounted for some variation in the classification accuracies for all procedures ranging from 0.6% to 1.7%

Tables 3.4 and 3.5 describe the mean classification accuracies and standard errors of repeated measures LDA and repeated measures QDA based on MLE and MVE, respectively, by population distribution and number of outcomes. The mean classification accuracy of the repeated measures discriminant analysis procedures were highest when the data were sampled from a multivariate normal distribution and lowest when the data were sampled from extremely heavy-tailed distribution, regardless of the type of estimation method adopted, number of outcome variables, or mean configuration. In particular, there were negligible differences in the mean classification accuracy of repeated measures discriminant analysis procedures based on MLE and those based on robust estimators when the data were sampled from a multivariate normal distribution, or a moderately multivariate heavy-tailed distribution. However, the robust estimators' procedures were more accurate than MLE when the data were sample from a

multivariate heavy-tailed distribution. For example, the mean classification accuracy of st-UNCS based on MLE and MVE were 0.86 and 0.85 when $q = 7$ and data were sampled from a multivariate normal distribution with outcome variables. Whereas, the mean accuracy of the former and latter procedures were 0.54 and 0.70 when $q = 7$ and the data were sampled from a multivariate Cauchy distribution, respectively (Table 3.4). Similar patterns were observed in repeated measures quadratic discriminant analysis procedures (Table 3.5).

Furthermore, when the data were sampled from a multivariate normal distribution, the mean accuracy of each repeated measures LDA procedures increased as the number of outcomes increased, regardless of the method or estimation. However, the increase in classification accuracy as q increased was smaller when the data were sampled from a multivariate non-normal distribution. For example, when the data were sampled from a multivariate normal distribution, the increase in classification accuracy of the un-UNAR procedure based on MLE and MVE was about 0.07 as q increased from 3 to 7. But when data were sampled from a multivariate lognormal distribution, there were negligible differences in the classification accuracies for these procedures as q increased. In contrast, the mean classification accuracy of the repeated measures QDA procedures decreased as q increased for almost all the investigated population data distributions, except for the multivariate lognormal and Cauchy distributions (Table 3.5).

However, for multivariate lognormal and Cauchy distributions, smaller to no change in mean classification accuracy was observed for all procedures for MLE. For the un-UNAR procedure when data were sampled from multivariate lognormal, 0.54 mean classification accuracy was observed when number of outcomes were both three and seven, whereas for data sampled from multivariate Cauchy distribution mean classification accuracy were 0.54 and 0.56, respectively for MLE (Table 3.4). In contrast, the decreased in mean classification accuracy for

non-normal distributions were much lower for repeated measures discriminant analysis based on MLE compared to the robust methods. Moreover, higher mean classification accuracies were observed in these non-normal distributions under robust methods. For the un-UNAR procedure under MVE, when data were sampled from multivariate lognormal, 0.56 mean accuracy was observed when number of outcomes, $q = 3$ compared to 0.57 accuracy when $q = 7$, whereas for data sampled from multivariate Cauchy distribution mean accuracies were 0.70 and 0.73, respectively (Table 3.4). Again, similar observations were seen for all procedures in robust methods. For both MLE and robust methods mean accuracies, smaller to no change was observed for structured and unstructured mean for all procedures, irrespective of population distributions and number of outcomes.

For Table 3.5, when QDA classifier was adopted for unequal group covariance, opposite effect of the number of outcomes seen in Table 3 was observed, that is $q = 3$ had a higher mean accuracy rate than $q = 7$ in normal distribution and t-distribution for un-UNAR, un-UNCS and st-UNCS except st-UNAR for both MLE and robust methods. For example: for the un-UNAR procedure under MVE, the mean accuracy for $q = 3$ was 0.82, whereas for $q = 7$, it was 79, whiles for the st-UNAR procedure, the mean accuracy rate for $q = 3$ was 0.84, whereas for $q = 7$, it was 0.86. Also, higher mean accuracies were observed for the st-UNAR procedure compared to the other procedures. In addition, we observed higher increase in classification accuracy for increase number of outcomes in Table 3.5 for multivariate lognormal distribution under MLE compared to Table 3. For example, for the un-UNAR procedure under MLE, when data were sampled from multivariate lognormal, 0.65 accuracy was observed when number of outcomes, $q = 3$ compared to 0.70 accuracy rate when $q = 7$. With regards to MLE and robust methods, higher mean accuracies were observed for Cauchy distribution based on robust methods compared

to MLE, but smaller to no increase mean accuracies were observed in other population distributions (Table 3.4).

Tables 3.6 and 3.7 describe the overall mean classification accuracy of the repeated measures LDA and QDA procedures by population distribution and mean configuration, respectively. There were negligible differences in the accuracy of the repeated measures discriminant analysis based on MLE and robust estimators when the data were sampled from multivariate normal or multivariate *t* distribution. However, the robust procedures were significantly more accurate than MLE when the data were sampled from multivariate Cauchy distribution. For example, the mean classification accuracy of st-UNAR based on MLE and robust estimators were 0.72 and 0.71, when the data were sample from a multivariate t-distributions with mean configuration I, respectively. Whereas the mean accuracy of the former and latter procedures were 0.53 and 0.64 when the data were sampled from a multivariate Cauchy distribution with the same mean configuration I.

On the other hand, the impact of choice of mean configuration on the accuracy of the repeated measures LDA models was confounded by the population distribution. Specifically, when the data were sampled from a multivariate normal distribution, the mean classification accuracy of the repeated measures LDA procedures were lowest under mean configuration I, which assumed no change in mean pattern over time for constant mean among repeated outcomes, but highest under mean configuration IV, which assumed unstructured means among the repeated measurements regardless of the estimation methods. However, there were negligible differences in the classification accuracy of the procedures based on MLE estimators across all the mean configurations when the data were sampled from a multivariate log-normal or Cauchy distribution. In contrast, accuracy of the procedures based on robust estimators varied across mean

configurations when the data were sampled from a multivariate lognormal or multivariate Cauchy distribution. For example, the mean accuracy of the st-UNAR procedure based on MLE increased by 0.16 (were 0.72 and 0.88) across the mean configurations when the data were sampled from multivariate normal distribution, whereas there was negligible change in mean accuracy of this procedure across the mean configurations when the data were generated from a multivariate log-normal distribution. In contrast, the change in mean classification accuracy for st-UNAR procedure based on robust estimators across the mean configurations were 0.16 and 0.13 when the data were sampled from multivariate normal and multivariate Cauchy distributions, respectively (Table 3.6). Thus, the mean accuracy of the procedures based on robust estimators increased across the mean configurations when the data were sampled from multivariate Cauchy distribution compared to procedures based on MLE (Table 3.6). Similar results were obtained for repeated measures quadratic discriminant analysis (Table 3.7). Results for repeated measures discriminant analysis based MCD and MVE were similar, hence we reported results for MVE estimation to avoid repetition.

3.4 Application: Manitoba Inflammatory Bowel Disease Study

Multivariate repeated measures data from the Manitoba Inflammatory Bowel Disease (IBD) Cohort Study, a prospective longitudinal cohort study to investigate the determinants of disease outcomes in community dwelling individuals living with Crohn's disease or ulcerative colitis, were used to demonstrate the application of these methods. Data were collected at six-month intervals, after baseline, using self-report instruments. Study participants were rated as having active ($n_1 = 214$) or inactive ($n_2 = 127$) disease based on self-reported IBD symptoms at study entry. Details about the Manitoba IBD Cohort Study have been previously published

elsewhere[47, 37]. Differences between active and inactive disease groups on a disease-specific measure of quality of life, the IBD questionnaire (IBDQ)[28], were investigated in the first year of the study (i.e., three measurement occasions, at baseline (0 months), 6 months, and 12 months, p=3). The primary research question is to be able to identify active and inactive disease groups at one year of diagnosis using their longitudinal profiles of quality of life. Multivariate repeated measures data collected on the four IBDQ domains (q=4) namely emotional health (IBDQ-eh), systematic symptoms (IBDQ-ss), social function (IBDQ-sf) and bowel symptoms (IBDQ-bws) over the one-year period were used to discriminate between both groups of participants.

Of the 389 participants who provided data at baseline (month 0), 213 had complete IBDQ domains at the end of the first year. Among the 213 participants, 133 were participants with active IBD. Table 3.8 and Figure 3.1 describe the differences on each domain for active and inactive participants in the Manitoba IBD Cohort Study. Participants with inactive disease had higher quality of life scores on all four domains than participants with active disease (Figure 3.1). The group means and descriptive measures of multivariate skewness and kurtosis for the IBDQ data are reported in Table 3.7. The expected Mardia's multivariate skewness is 0 and kurtosis is 24 for a multivariate normal distribution of 4 variables[48]. P-value smaller than 0.05 indicated significant skewness or kurtosis. At least one of these tests was significant, thus the underlying joint population was non-normal. Overall, the multivariate skewness and kurtosis suggested a moderate departure from the assumption of a normal distribution in the active disease group when compared with the inactive group (Table 3.8). A non-constant trend was observed in the group means for both the active and inactive disease group (Table 3.8) and moderate difference was observed in group covariances. Hence, we used RMQDA assuming Kronecker product covariance. The advantage of imposing the Kronecker product structure on the data is that it reduces the number of

parameters to estimate, which results in greater precision of the estimates since *n/pq* is small for both active (~11) and inactive (~7) disease groups.

In estimating parameters for the proposed robust RMQDA method (MVE), the symmetric trimming parameter was chosen to be 10%, and compared to RMQDA based on MLE. Results of these approaches were reported in Table 3.9. Overall, we observed a 1% to 3% increase in all robust methods compared to MLE with 10% trimming. As observed from the simulation, these robust procedures may not always be more efficient than repeated measures discriminant analysis based on MLE for moderate departures from a multivariate normal distribution.

In addition, we investigated the influence of class imbalance on the accuracy of these proposed models, for which additional simulation condition results are provided in the Appendix. Tables 3.10 and 3.11 contain class-specific accuracies of repeated measures LDA procedures based on MLE and robust MVE by number of outcomes and sample sizes for normal and Cauchy distributions respectively. Conclusions and observations from the additional simulation results remained the same as the initial simulations. Thus, class imbalance did not influence the proposed repeated measures models.

3.5 Discussion

This study investigated repeated measures discriminant analysis procedures that assume structured and unstructured means with Kronecker covariances based on maximum trimmed estimators for discriminating between population groups. As expected, the classification accuracies of the repeated measures discriminant analysis procedures were highest in multivariate normal distributions but lowest when the data were sampled from a multivariate Cauchy distribution. Repeated measures discriminant analysis procedures based on MTLE were more

accurate than the conventional repeated measures discriminant analysis based on MLEs when data were sampled from multivariate lognormal and Cauchy distributions[24, 32]. However, there were negligible differences in the mean classification accuracies of the MTLE and MLE procedures under multivariate normal and moderately heavy-tailed distributions. Furthermore, our results also showed that the impact of population group mean separation i.e., distance and data dimensions on the classification accuracy of the conventional repeated measures discriminant analysis procedures could be masked by the departure from the assumption of multivariate normality. In contrast, the impact of both group means separation and data dimension on the accuracy of the repeated measures discriminant analysis procedures based on MTLE was not confounded by departure from the assumption of multivariate normality. A common criticism of trimmed estimators is that they are less powerful in small-sampled studies under multivariate normal distributions[49]. However, our simulation study showed negligible differences in the accuracy of repeated measures discriminant analysis procedures based on MTLE and those based on MLE in small-sampled conditions[32].

Of note is the finding that, we observed similar classification accuracy of the investigated repeated measures discriminant analysis procedures based on parsimonious means/or covariance matrices, regardless of the method of estimation. This can be attributed to the fact that all these procedures were investigated in scenarios where the underlying means and covariance structures were correctly specified. It is most likely that the predictive performances of these procedures might vary especially in multivariate repeated measures data in which population means and covariances are unstructured where the assumption of parsimony (i.e. Kronecker product assumption for group means and/or covariance) are violated. While previous research studies have suggested that repeated measures discriminant analysis procedures often result in decreased

classification accuracy when the means and covariances are misspecified[32, 50], there is limited investigation of the robustness of the repeated measures discriminant analysis based on MTLE to model mis-specification. Future research investigations will examine the robustness of repeated measures discriminant analysis procedures based on MTLE to misspecification of group means and covariance structure.

This study has some limitations. Our simulation only investigated the classification performance of the investigated models in multivariate normal and multivariate heavy-tailed distributions but not in multivariate skewed distribution. Previous investigations have shown that trimmed estimators are particularly more efficient in data with moderate to significant heavy-tailed distributions[32]. Second, the assumption of complete multivariate repeated measures data in which there is no missing data on all outcomes and at all measurement occasions might not be realistic in multivariate repeated measures data often encountered in applied research. In clinical settings, missing data often occur in multivariate repeated measures studies because patients miss some of their regular appointments or because some variables may not be measured at particular visits. Repeated measures discriminant analysis based on mixed-effects models have been proposed for incomplete multivariate repeated measures data but the misspecification of the common assumption of the random effects parameter as multivariate normal distribution may seriously affect the accuracy of discriminant analysis classification rules. [25] Pattern mixture and selection models have been proposed to adjust for potential bias in models when it cannot be assumed that the mechanism of missingness is ignorable[51 52, 53]. Further research will investigate the development of repeated measures discriminant analysis procedures based on these models with imputations and further developments in which mixed-effects models can be extended to these robust trimmed methods for classification.

In addition, this study relied on the assumption of Kronecker product structure covariance to capture the relationship among multivariate repeated measures. Various researches have used Kronecker product covariance structures to address sample size and computational issues in multivariate repeated measures[12, 13, 54-56]. While Kronecker structures provide a parsimonious model approach to parameter estimation, the accuracy of the resulting repeated measures discriminant analysis procedures may be reduced when the means and/or covariance structure of the data is misspecified[57]. It is important that the choice of these repeated measures discriminant analysis procedures be guided first by a preliminary examination of the appropriate means and/or covariance structure in the multivariate repeated measures data[21, 58]. For example, several procedures have been developed for testing hypotheses Kronecker product covariance structures in multivariate repeated measures for such purposes[13, 56, 59,60].

In summary, this study proposes a new class of repeated measures discriminant analysis procedures based on MTLE, which overcomes the inherently restrictive distributional assumption of multivariate normality when discriminating between populations groups in multivariate repeated measures data characterized by multivariate non-normal distributions. These procedures are useful for developing classification models for both short-term and long-term outcomes in complex data.

References

1. Fieuws S, Verbeke G, Maes B, et al. Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics* 2007; 9: 419-431.
2. Hughes DM, Komárek A, Czanner G, et al. Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types. *Statistical methods in medical research* 2018; 27: 2060-2080.
3. Inoue LYT, Etzioni R, Morrell C, et al. Modeling disease progression with longitudinal markers. *Journal of the American Statistical Association* 2008; 103: 259-270.
4. Li Y, Wang Y, Wu G, et al. Discriminant analysis of longitudinal cortical thickness changes in alzheimer's disease using dynamic and network features. *Neurobiology of aging* 2012; 33: 427. e415-427. e430.
5. Marshall G and Barón AE. Linear discriminant models for unbalanced longitudinal data. *Statistics in medicine* 2000; 19: 1969-1981.
6. Morrell CH, Brant LJ, Sheng S, et al. Screening for prostate cancer using multivariate mixed-effects models. *Journal of applied statistics* 2012; 39: 1151-1175.
7. Diggle P. *Analysis of longitudinal data*. Oxford University Press, 2002.
8. Verbeke G, Fieuws S, Molenberghs G, et al. The analysis of multivariate longitudinal data: A review. *Statistical methods in medical research* 2014; 23: 42-59.
9. Galecki AT. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics-Theory and Methods* 1994; 23: 3105-3119.
10. Lix L and Sajobi T. Discriminant analysis for repeated measures data: A review. *Front Psychol* 2010; 1: 146.
11. Roy A and Khattree R. Discrimination and classification with repeated measures data under different covariance structures. *Communications in Statistics—Simulation and Computation®* 2005; 34: 167-178.
12. Roy A and Khattree R. On discrimination and classification with multivariate repeated measures data. *Journal of Statistical Planning and Inference* 2005; 134: 462-485.
13. Roy A and Khattree R. On implementation of a test for kronecker product covariance structure for multivariate repeated measures data. *Statistical Methodology* 2005; 2: 297-306.
14. Tomasko L, Helms RW and Snapinn SM. A discriminant analysis extension to mixed models. *Statistics in medicine* 1999; 18: 1249-1260.
15. Albert A. Discriminant analysis based on multivariate response curves: A descriptive approach to dynamic allocation. *Statistics in medicine* 1983; 2: 95-106.
16. Usami S and Murayama K. Time-specific errors in growth curve modeling: Type-1 error inflation and a possible solution with mixed-effects models. *Multivariate behavioral research* 2018; 53: 876-897.
17. Wu W, West SG and Taylor AB. Evaluating model fit for growth curve models: Integration of fit indices from sem and mlm frameworks. *Psychological methods* 2009; 14: 183.
18. Komárek A, Hansen BE, Kuiper EM, et al. Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statistics in Medicine* 2010; 29: 3267-3283.
19. Marshall G, De la Cruz-Mesía R, Barón AE, et al. Non-linear random effects model for multivariate responses with missing data. *Statistics in Medicine* 2006; 25: 2817-2830.

20. Marshall G, De la Cruz-Mesía R, Quintana FA, et al. Discriminant analysis for longitudinal data with multiple continuous responses and possibly missing data. *Biometrics* 2009; 65: 69-80.
21. Roy A. A new classification rule for incomplete doubly multivariate data using mixed effects model with performance comparisons on the imputed data. *Statistics in medicine* 2006; 25: 1715-1728.
22. Fieuws S, Verbeke G and Molenberghs G. Random-effects models for multivariate repeated measures. *Statistical methods in medical research* 2007; 16: 387-397.
23. Hughes DM, Komárek A, Czanner G, et al. Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types. *Statistical methods in medical research* 2018; 27: 2060-2080.
24. Hadi AS and Luceño A. Maximum trimmed likelihood estimators: A unified approach, examples, and algorithms. *Computational Statistics & Data Analysis* 1997; 25: 251-272.
25. Maronna RA. Robust m-estimators of multivariate location and scatter. *The Annals of Statistics* 1976; 4: 51-67.
26. Yan F, Lin X and Huang X. Dynamic prediction of disease progression for leukemia patients by functional principal component analysis of longitudinal expression levels of an oncogene. *The Annals of Applied Statistics* 2017; 11: 1649-1670.
27. Cheng T-C and Biswas A. Maximum trimmed likelihood estimator for multivariate mixed continuous and categorical data. *Computational Statistics & Data Analysis* 2008; 52: 2042-2065.
28. Barön AE. Misclassification among methods used for multiple group discrimination-the effects of distributional properties. *Statistics in medicine* 1991; 10: 757-766.
29. He X and Fung WK. High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis* 2000; 72: 151-162.
30. Williams BK and Titus K. Assessment of sampling stability in ecological applications of discriminant analysis. *Ecology* 1988; 69: 1275-1285.
31. Fitzmaurice G, Laird N and Ware J. Modelling the mean: Parametric curves. *Applied Longitudinal Analysis Hoboken, New Jersey, USA: John Wiley & Sons Inc* 2004: 141-147.
32. Sajobi TT, Lix LM, Dansu BM, et al. Robust descriptive discriminant analysis for repeated measures data. *Computational Statistics & Data Analysis* 2012; 56: 2782-2794.
33. Thomas DR and Zumbo BD. Using a measure of variable importance to investigate the standardization of discriminant coefficients. *Journal of Educational and Behavioral Statistics* 1996; 21: 110-130.
34. Ripley B. Ebooks corporation. Stochastic simulation. Wiley Online Library, 1987.
35. Aitchison J. The statistical analysis of compositional analysis. Chapman & Hall, London, 1986.
36. Azzalini A. *The skew-normal and related families*. Cambridge University Press, 2013.
37. Stigler SM. Do robust estimators work with real data? *The Annals of Statistics* 1977: 1055-1098.
38. Wilcox RR. *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. Springer, 2010.
39. Ramsey PH and Ramsey PP. Optimal trimming and outlier elimination. *Journal of Modern Applied Statistical Methods* 2007; 6: 2.
40. Rousseeuw PJ and Driessen KV. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 1999; 41: 212-223.
41. Maronna RA, Martin RD and Yohai VJ. *Robust statistics: Theory and methods (with r)*. John Wiley & Sons, 2019.

42. Todorov V and Filzmoser P. An object-oriented framework for robust multivariate analysis. 2009.
43. Rousseeuw PJ and Leroy AM. *Robust regression and outlier detection*. John wiley & sons, 2005.
44. Liu Y and Zumbo BD. The impact of outliers on cronbach's coefficient alpha estimate of reliability: Visual analogue scales. *Educational and Psychological Measurement* 2007; 67: 620-634.
45. Liu Y and Zumbo BD. Impact of outliers arising from unintended and unknowingly included subpopulations on the decisions about the number of factors in exploratory factor analysis. *Educational and psychological measurement* 2012; 72: 388-414.
46. Cohen J. Statistical power analysis for the behavioral sciences, 2nd edn. Á/l. Erbaum Press, Hillsdale, NJ, USA, 1988.
47. Bernstein CN, Rawsthorne P, Cheang M, et al. A population-based case control study of potential risk factors for ibd. *American Journal of Gastroenterology* 2006; 101: 993-1002.
48. Mardia KV. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 1970; 57: 519-530.
49. Srivastava DK and Mudholkar GS. Trimmed t2: A robust analog of hotelling's t2. *Journal of Statistical Planning and Inference* 2001; 97: 343-358.
50. Sajobi TT, Lix LM, Li L, et al. Discriminant analysis for repeated measures data: Effects of mean and covariance misspecification on bias and error in discriminant function coefficients. *Journal of Modern Applied Statistical Methods* 2011; 10: 15.
51. Hedeker D and Gibbons RD. Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological methods* 1997; 2: 64.
52. Little RJ. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* 1993; 88: 125-134.
53. Little RJ. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the american statistical association* 1995; 90: 1112-1121.
54. Naik DN and Rao SS. Analysis of multivariate repeated measures data with a kronecker product structured covariance matrix. *Journal of Applied Statistics* 2001; 28: 91-105.
55. Krzyśko M and Skorzybut M. Discriminant analysis of multivariate repeated measures data with a kronecker product structured covariance matrices. *Statistical papers* 2009; 50: 817-835.
56. Lu N and Zimmerman DL. The likelihood ratio test for a separable covariance matrix. *Statistics & probability letters* 2005; 73: 449-457.
57. Zhou J and Qu A. Informative estimation and selection of correlation structure for longitudinal data. *Journal of the American Statistical Association* 2012; 107: 701-710.
58. Roy A. A note on testing of kronecker product covariance structures for doubly multivariate data. In: *Proceedings of the American Statistical Association, statistical computing section* 2007, pp.2157-2162.
59. Srivastava MS, von Rosen T and Von Rosen D. Models with a kronecker product covariance structure: Estimation and testing. *Mathematical Methods of Statistics* 2008; 17: 357-370.
60. Filipiak K, Klein D and Roy A. A comparison of likelihood ratio tests and rao's score test for three separable covariance matrix structures. *Biometrical Journal* 2017; 59: 192-215.

Table 3.1: Four mean configuration structures assumed for population 1 ($\boldsymbol{\mu}_1$) in the Monte Carlo Study

Configuration	*p*	*q*=3	*q*=7
I		$\mathbf{1}_p \otimes$ (30,30,30)	$\mathbf{1}_p \otimes$ (30,30,30,30,30,30,30)
II	3 or 5	$\mathbf{1}_p \otimes$ (27,29,31)	$\mathbf{1}_p \otimes$ (30,31,32,33,34,35,36)
III		$\mathbf{1}_p \otimes$ (30,25,30)	$\mathbf{1}_p \otimes$ (30,27,24,21,24,27,30)
IV	3	(1,1.1,1.2) $\otimes$ (30,25,30)	(1,1.1,1.2)$\otimes$ (30,27,24,21,24,27,30)
	5	(1,1.1,1.2,1.3,1.4)$\otimes$ (30,25,30)	(1,1.1,1.2,1.3,1.4)$\otimes$ (30,27,24,21,24,27,30)

Note: For population 2, $\boldsymbol{\mu}_2 = \mathbf{1}_\mathrm{p} \otimes 25\mathbf{1}_\mathrm{q}$ for all conditions; *q*=number of outcome variables; *p*=number of repeated occasions

Table 3.2: Configuration of unstructured between-outcomes covariance matrix $\boldsymbol{\Sigma}_1$ given within-outcome correlation coefficient (ρ) for the Monte Carlo Study

Within-outcome correlation coefficient (ρ)	0.3	0.7
q=3	$\boldsymbol{\Sigma}_1 = 60\begin{pmatrix} 1 & 0.15 & 0.3 \\ 0.15 & 1 & 0.45 \\ 0.3 & 0.45 & 1 \end{pmatrix}$	$\boldsymbol{\Sigma}_1 = 60\begin{pmatrix} 1 & 0.65 & 0.66 \\ 0.65 & 1 & 0.7 \\ 0.66 & 0.7 & 1 \end{pmatrix}$
q=7	$\boldsymbol{\Sigma}_1 = 60\begin{pmatrix} 1 & 0.35 & 0.2 & 0.2 & 0.3 & 0.25 & 0.3 \\ 0.35 & 1 & 0.3 & 0.3 & 0.37 & 0.32 & 0.34 \\ 0.2 & 0.3 & 1 & 0.31 & 0.35 & 0.32 & 0.34 \\ 0.2 & 0.3 & 0.31 & 1 & 0.3 & 0.25 & 0.33 \\ 0.3 & 0.37 & 0.35 & 0.3 & 1 & 0.28 & 0.34 \\ 0.25 & 0.32 & 0.32 & 0.25 & 0.28 & 1 & 0.35 \\ 0.3 & 0.34 & 0.34 & 0.33 & 0.34 & 0.35 & 1 \end{pmatrix}$	$\boldsymbol{\Sigma}_1 = 60\begin{pmatrix} 1 & 0.65 & 0.66 & 0.7 & 0.72 & 0.65 & 0.75 \\ 0.65 & 1 & 0.7 & 0.7 & 0.67 & 0.72 & 0.74 \\ 0.66 & 0.7 & 1 & 0.71 & 0.75 & 0.72 & 0.74 \\ 0.7 & 0.7 & 0.71 & 1 & 0.7 & 0.65 & 0.73 \\ 0.72 & 0.67 & 0.75 & 0.7 & 1 & 0.68 & 0.74 \\ 0.65 & 0.72 & 0.63 & 0.65 & 0.68 & 1 & 0.75 \\ 0.75 & 0.74 & 0.74 & 0.73 & 0.74 & 0.75 & 1 \end{pmatrix}$

Note: $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ or $\boldsymbol{\Sigma}_1 = 3\boldsymbol{\Sigma}_2$

Table 3.3: Estimated percentage of variation explained (95% C.I) from Analysis of Variance

Simulation Condition	un-UNAR	un-UNCS	st-UNAR	st-UNCS
p	*	0.3	*	0.2
q	0.6	0.6	1.1	0.6
Covariance structure	*	*	*	*
ρ	2.0	1.2	1.2	1.4
n	*	*	*	*
Mean Configuration	10.5	11.0	9.2	10.5
Population Distribution	51.4	49.8	56.0	50.2
Covariance ratio (QDA vs LDA)	0.9	*	3.3	1.1
Estimation (MLE vs MVE)	0.6	1.7	1.5	1.5
Population Distribution x Covariance ratio	5.1	4.9	3.9	5.0
Population Distribution x Mean Configuration	5.0	5.2	4.1	5.0

Note: *= Estimated percentage of variation explained close to zero; C.I = confidence interval; p= number of repeated occasions; p = number of responses; ρ= coefficient of correlation; n =sample size; un-UNAR = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; un-UNCS = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; st-UNAR = structured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; st-UNCS = structured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; MLE=Maximum likelihood estimator; MVE= minimum volume ellipsoid; LDA= Linear Discriminant Analysis; QDA= Quadratic Discriminant Analysis

Table 3.4: Overall Mean Accuracy of Repeated Measures LDA procedures based on MLE and Robust Estimator, Estimator, MVE (standard error) by population distribution, Number of Outcomes for equal group covariance

Distribution	q	MLE				MVE			
		un-UNAR	un-UNCS	st-UNAR	st-UNCS	un-UNAR	un-UNCS	st-UNAR	st-UNCS
Normal	3	0.77(0.03)	0.77(0.03)	0.77(0.03)	0.78(0.03)	0.77(0.03)	0.77(0.03)	0.76(0.03)	0.78(0.03)
	7	0.84(0.03)	0.85(0.03)	0.84(0.03)	0.86(0.03)	0.84(0.03)	0.85(0.03)	0.84(0.03)	0.85(0.03)
T	3	0.77(0.03)	0.77(0.03)	0.77(0.03)	0.78(0.03)	0.77(0.03)	0.77(0.03)	0.76(0.03)	0.77(0.03)
	7	0.84(0.03)	0.85(0.03)	0.84(0.03)	0.86(0.03)	0.84(0.03)	0.85(0.03)	0.84(0.03)	0.85(0.03)
Lognormal	3	0.54(0.03)	0.53(0.03)	0.54(0.03)	0.53(0.03)	0.56(0.04)	0.56(0.04)	0.56(0.04)	0.56(0.04)
	7	0.54(0.03)	0.54(0.03)	0.54(0.03)	0.54(0.03)	0.57(0.04)	0.57(0.04)	0.57(0.04)	0.57(0.04)
Cauchy	3	0.54(0.05)	0.54(0.05)	0.54(0.05)	0.54(0.05)	0.70(0.04)	0.70(0.04)	0.70(0.04)	0.70(0.04)
	7	0.56(0.06)	0.56(0.06)	0.56(0.06)	0.56(0.07)	0.73(0.05)	0.73(0.05)	0.73(0.05)	0.73(0.05)

Note: un-UNAR = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; un-UNCS = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; st-UNAR = structured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; st-UNCS = structured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; MLE=Maximum likelihood estimator; MVE= minimum volume ellipsoid; LDA= Linear Discriminant Analysis

Table 3.5: Overall Mean Accuracy of Repeated Measures QDA procedures based on MLE and Robust Estimator, MVE (standard error) by population distribution, Number of outcomes for unequal group covariance

Distribution	*q*	MLE				MVE			
		un-UNAR	un-UNCS	st-UNAR	st-UNCS	un-UNAR	un-UNCS	st-UNAR	st-UNCS
Normal	3	0.83(0.04)	0.83(0.03)	0.85(0.03)	0.83(0.03)	0.82(0.04)	0.83(0.03)	0.84(0.03)	0.83(0.03)
	7	0.80(0.04)	0.81(0.03)	0.86(0.03)	0.81(0.03)	0.79(0.04)	0.81(0.03)	0.86(0.03)	0.80(0.03)
T	3	0.83(0.04)	0.83(0.03)	0.85(0.03)	0.83(0.03)	0.82(0.04)	0.83(0.03)	0.84(0.03)	0.83(0.03)
	7	0.80(0.04)	0.81(0.03)	0.86(0.03)	0.81(0.03)	0.79(0.04)	0.81(0.03)	0.86(0.03)	0.80(0.03)
Lognormal	3	0.65(0.11)	0.69(0.04)	0.68(0.05)	0.69(0.04)	0.68(0.07)	0.68(0.04)	0.67(0.05)	0.68(0.04)
	7	0.70(0.04)	0.70(0.04)	0.70(0.04)	0.70(0.04)	0.69(0.04)	0.69(0.04)	0.69(0.04)	0.69(0.04)
Cauchy	3	0.53(0.08)	0.54(0.05)	0.54(0.05)	0.54(0.05)	0.66(0.05)	0.66(0.05)	0.66(0.05)	0.66(0.05)
	7	0.55(0.06)	0.56(0.07)	0.56(0.06)	0.56(0.07)	0.67(0.06)	0.67(0.06)	0.67(0.06)	0.67(0.06)

Note: un-UNAR = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; un-UNCS = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; st-UNAR = structured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; st-UNCS = structured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; MLE=Maximum likelihood estimator; MVE= minimum volume ellipsoid; QDA= Quadratic Discriminant Analysis

Table 3.6: Overall Mean Accuracy of Repeated Measures LDA procedures based on MLE and Robust Estimator, MVE (standard error) by population distribution, mean configuration for equal group covariance

Distribution	Mean Configuration	MLE				MVE			
		un-UNAR	un-UNCS	st-UNAR	st-UNCS	un-UNAR	un-UNCS	st-UNAR	st-UNCS
Normal	I	0.72(0.03)	0.72(0.04)	0.72(0.03)	0.73(0.04)	0.71(0.04)	0.72(0.04)	0.71(0.03)	0.72(0.04)
	II	0.76(0.03)	0.78(0.03)	0.76(0.03)	0.79(0.03)	0.76(0.03)	0.78(0.03)	0.76(0.03)	0.78(0.03)
	III	0.79(0.03)	0.77(0.03)	0.79(0.03)	0.79(0.03)	0.79(0.03)	0.79(0.03)	0.79(0.03)	0.79(0.03)
	IV	0.88(0.02)	0.89(0.02)	0.88(0.02)	0.89(0.02)	0.88(0.03)	0.89(0.03)	0.87(0.03)	0.89(0.03)
T	I	0.72(0.03)	0.72(0.04)	0.72(0.03)	0.73(0.04)	0.71(0.04)	0.72(0.04)	0.71(0.03)	0.72(0.04)
	II	0.76(0.03)	0.78(0.03)	0.76(0.03)	0.79(0.03)	0.76(0.03)	0.79(0.03)	0.76(0.03)	0.79(0.03)
	III	0.79(0.03)	0.77(0.03)	0.79(0.03)	0.79(0.03)	0.79(0.03)	0.79(0.03)	0.79(0.03)	0.79(0.03)
	IV	0.88(0.02)	0.89(0.02)	0.88(0.03)	0.89(0.02)	0.88(0.03)	0.89(0.03)	0.87(0.03)	0.89(0.03)
Lognormal	I	0.54(0.03)	0.53(0.03)	0.54(0.03)	0.54(0.03)	0.56(0.04)	0.56(0.04)	0.56(0.04)	0.56(0.04)
	II	0.54(0.03)	0.54(0.03)	0.54(0.03)	0.54(0.03)	0.56(0.04)	0.56(0.04)	0.56(0.04)	0.56(0.04)
	III	0.54(0.03)	0.54(0.03)	0.54(0.03)	0.54(0.03)	0.56(0.04)	0.56(0.04)	0.56(0.04)	0.56(0.04)
	IV	0.54(0.04)	0.54(0.03)	0.54(0.04)	0.54(0.03)	0.57(0.04)	0.57(0.04)	0.57(0.04)	0.57(0.04)
Cauchy	I	0.53(0.04)	0.53(0.04)	0.53(0.04)	0.53(0.04)	0.64(0.04)	0.64(0.04)	0.64(0.04)	0.64(0.05)
	II	0.54(0.05)	0.54(0.05)	0.54(0.05)	0.54(0.05)	0.69(0.05)	0.69(0.05)	0.68(0.05)	0.69(0.05)
	III	0.54(0.05)	0.54(0.05)	0.54(0.05)	0.54(0.05)	0.69(0.05)	0.69(0.05)	0.69(0.05)	0.69(0.05)
	IV	0.57(0.04)	0.57(0.03)	0.57(0.04)	0.57(0.03)	0.78(0.04)	0.78(0.04)	0.77(0.04)	0.78(0.04)

Note: un-UNAR = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; un-UNCS = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; st-UNAR = structured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; st-UNCS = structured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; MLE=Maximum likelihood estimator; MVE= minimum volume ellipsoid; MVE= minimum volume ellipsoid

Table 3.7: Overall Mean Accuracy of Repeated Measures QDA procedures based on MLE and Robust Estimator, MVE (standard error) by population distribution, mean configuration for unequal group covariance

Distribution	Mean Configuration	MLE				MVE			
		un-UNAR	un-UNCS	st-UNAR	st-UNCS	un-UNAR	un-UNCS	st-UNAR	st-UNCS
Normal	I	0.69(0.03)	0.70(0.02)	0.76(0.03)	0.70(0.02)	0.69(0.03)	0.69(0.03)	0.75(0.03)	0.69(0.03)
	II	0.78(0.04)	0.77(0.04)	0.82(0.03)	0.77(0.04)	0.77(0.04)	0.76(0.04)	0.81(0.03)	0.77(0.04)
	III	0.81(0.04)	0.80(0.04)	0.81(0.03)	0.80(0.04)	0.80(0.04)	0.80(0.04)	0.81(0.03)	0.80(0.04)
	IV	0.90(0.04)	0.92(0.03)	0.94(0.02)	0.92(0.03)	0.90(0.05)	0.92(0.03)	0.94(0.02)	0.92(0.03)
T	I	0.69(0.03)	0.70(0.02)	0.76(0.03)	0.70(0.03)	0.69(0.03)	0.69(0.03)	0.75(0.03)	0.69(0.03)
	II	0.78(0.04)	0.77(0.04)	0.82(0.03)	0.77(0.04)	0.77(0.04)	0.76(0.04)	0.81(0.04)	0.77(0.04)
	III	0.81(0.04)	0.80(0.04)	0.81(0.03)	0.80(0.04)	0.80(0.04)	0.80(0.04)	0.81(0.04)	0.80(0.04)
	IV	0.90(0.04)	0.92(0.03)	0.94(0.02)	0.92(0.03)	0.90(0.05)	0.92(0.03)	0.94(0.02)	0.92(0.03)
Lognormal	I	0.68(0.08)	0.69(0.04)	0.69(0.04)	0.69(0.04)	0.68(0.05)	0.68(0.04)	0.68(0.04)	0.68(0.04)
	II	0.68(0.07)	0.70(0.04)	0.69(0.04)	0.70(0.04)	0.69(0.05)	0.68(0.04)	0.68(0.04)	0.68(0.04)
	III	0.67(0.07)	0.68(0.04)	0.68(0.04)	0.68(0.04)	0.67(0.05)	0.67(0.04)	0.67(0.04)	0.67(0.04)
	IV	0.68(0.08)	0.70(0.04)	0.70(0.05)	0.70(0.04)	0.69(0.05)	0.69(0.04)	0.68(0.04)	0.69(0.04)
Cauchy	I	0.53(0.06)	0.54(0.05)	0.54(0.05)	0.54(0.05)	0.62(0.05)	0.62(0.05)	0.63(0.05)	0.62(0.05)
	II	0.54(0.07)	0.54(0.05)	0.55(0.05)	0.55(0.05)	0.64(0.05)	0.64(0.05)	0.65(0.05)	0.64(0.05)
	III	0.54(0.07)	0.54(0.05)	0.54(0.05)	0.54(0.05)	0.65(0.05)	0.65(0.05)	0.64(0.05)	0.65(0.05)
	IV	0.55(0.08)	0.56(0.04)	0.56(0.05)	0.56(0.03)	0.70(0.05)	0.71(0.04)	0.71(0.04)	0.71(0.04)

Note: un-UNAR = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; un-UNCS = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; st-UNAR = structured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; st-UNCS = structured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; MLE=Maximum likelihood estimator; MVE= minimum volume ellipsoid; QDA= Quadratic Discriminant Analysis

Table 3.8: Descriptive Statistics of IBDQ Domains in active and inactive IBD participants in Manitoba IBD Cohort Study

Measurement Occasion	IBDQ Domains	Active n=133			Inactive n=80		
		Mean (SD)	*Skewness (p-value)*	*Kurtosis (p-value)*	*Mean (SD)*	*Skewness (p-value)*	*Kurtosis (p-value)*
0-month	Emotional health	5.03(1.12)			5.46(1.15)		
	Systemic symptoms	4.78(1.30)	2.35* (<0.001)	24.9* (<0.001)	4.94(1.23)	6.07* (<0.001)	29.6* (<0.001)
	Social function	5.76(1.35)			6.30(0.87)		
	Bowel symptoms	5.03(1.11)			5.69(0.88)		
6-month	Emotional health	5.12(1.03)			5.71(0.89)		
	Systemic symptoms	4.34(1.21)	2.72* (<0.001)	25.45 (0.09)	5.16(1.15)	14.01* (<0.001)	39.74* (<0.001)
	Social function	5.92(1.19)			6.55(0.76)		
	Bowel symptoms	5.09(1.04)			5.88(0.90)		
12-month	Emotional health	5.11(1.11)			5.79(0.91)		
	Systemic symptoms	4.32(1.28)	4.59* (<0.001)	31.93* (<0.001)	5.22(1.14)	16.28* (<0.001)	38.33* (<0.001)
	Social function	5.81(1.30)			6.58(1.00)		
	Bowel symptoms	5.08(1.11)			5.96(0.90)		

**Note: ***p-value < 0.05, the joint distribution of the variables has significant skewness or kurtosis

Table 3.9: Overall Classification Accuracy of Conventional and Robust QDA procedures for IBD data

	un-UNAR	un-UNCS	st-UNAR	st-UNCS
MLE	0.50	0.60	0.64	0.51
MVE (10%)	0.53	0.63	0.65	0.52

Note: un-UNAR = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; un-UNCS = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; st-UNAR = structured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; st-UNCS = structured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; MLE=Maximum likelihood estimator; MVE= minimum volume ellipsoid

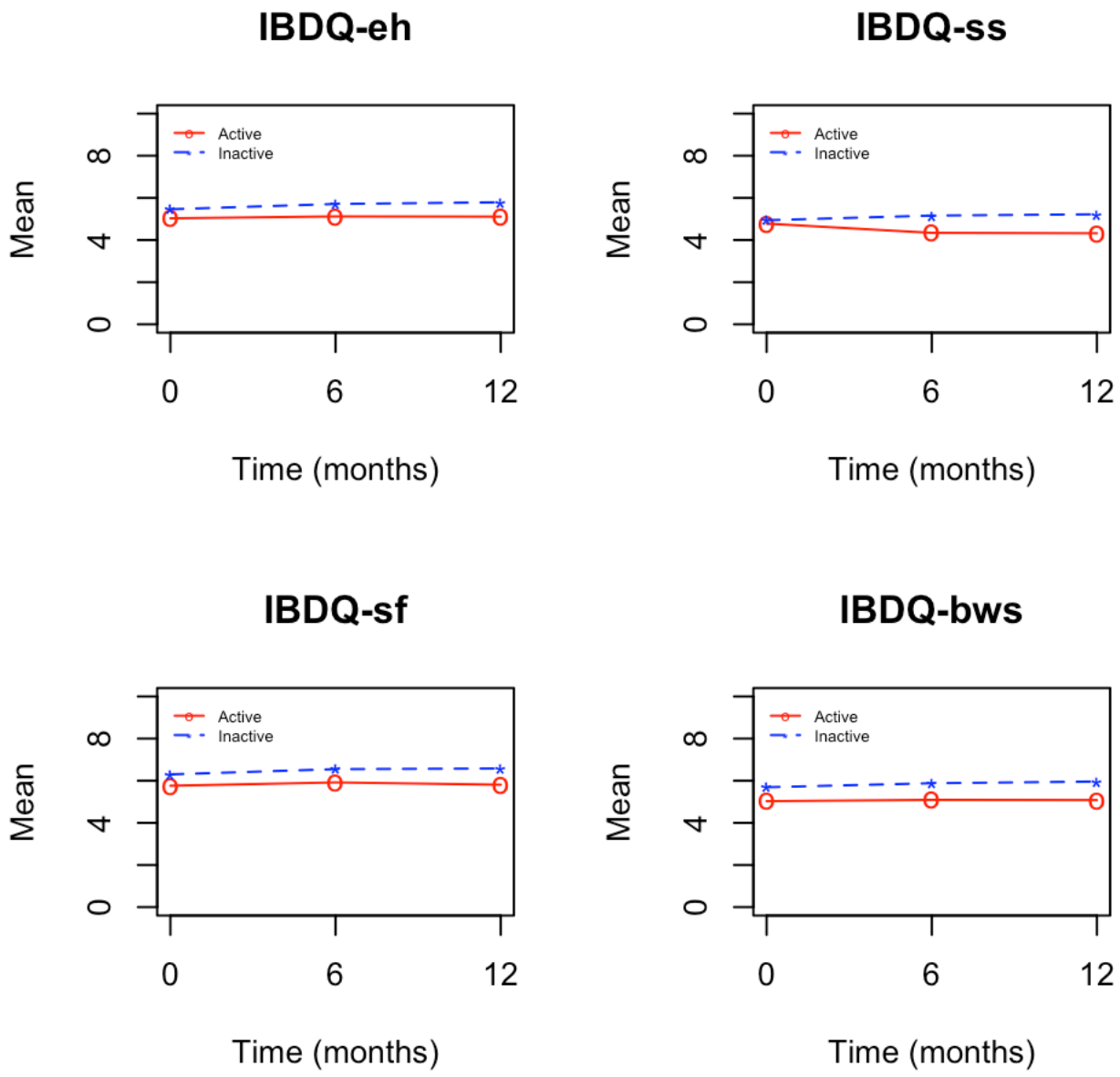

Figure 3.1: Observed mean longitudinal profiles of an indicator of whether a participant had active (Red) or inactive (Blue) IBD in each of the four IBDQ domains: emotional health (IBDQ-eh), systematic symptoms (IBDQ-ss), social function (IBDQ-sf) and bowel symptoms (IBDQ-bws)

Appendix

Table 3. 10: Class-Specific Accuracies of Repeated Measures LDA procedures based on MLE and Robust Estimator (MVE) for Normal distribution by Number of Outcomes and Sample Sizes

Methods			un-UNAR		un-UNCS		st-UNAR		st-UNCS	
		Sample Size (n_1, n_2)	Pop1	Pop2	Pop1	Pop2	Pop1	Pop2	Pop1	Pop2
Number of outcomes(*q=3)*										
MLE	Equal	50,50	0.70	0.85	0.70	0.85	0.69	0.84	0.71	0.85
		70,70	0.71	0.83	0.71	0.83	0.71	0.83	0.72	0.84
		100,100	0.72	0.82	0.72	0.82	0.71	0.82	0.72	0.83
	Unequal	40,60	0.70	0.85	0.70	0.85	0.70	0.85	0.71	0.85
		56,84	0.71	0.84	0.71	0.84	0.70	0.83	0.71	0.84
		80,120	0.71	0.82	0.72	0.82	0.71	0.82	0.72	0.83
MVE	Equal	50,50	0.69	0.85	0.70	0.85	0.69	0.84	0.70	0.85
		70,70	0.70	0.83	0.71	0.84	0.70	0.83	0.71	0.84
		100,100	0.71	0.82	0.72	0.83	0.71	0.82	0.72	0.83
	Unequal	40,60	0.69	0.85	0.69	0.85	0.69	0.84	0.70	0.85
		56,84	0.70	0.83	0.71	0.84	0.70	0.83	0.71	0.84
		80,120	0.71	0.82	0.72	0.83	0.70	0.81	0.72	0.83
Number of outcomes(*q=7)*										
MLE	Equal	50,50	0.78	0.91	0.79	0.91	0.78	0.91	0.79	0.92
		70,70	0.79	0.90	0.79	0.90	0.79	0.89	0.80	0.91
		100,100	0.80	0.88	0.80	0.89	0.80	0.88	0.81	0.90
	Unequal	40,60	0.78	0.91	0.79	0.91	0.78	0.91	0.80	0.92
		56,84	0.79	0.89	0.79	0.90	0.79	0.89	0.80	0.91
		80,120	0.79	0.88	0.80	0.89	0.79	0.88	0.81	0.90
MVE	Equal	50,50	0.78	0.91	0.79	0.92	0.78	0.90	0.79	0.92
		70,70	0.78	0.89	0.80	0.91	0.78	0.89	0.80	0.90
		100,100	0.79	0.88	0.80	0.89	0.79	0.88	0.80	0.90
	Unequal	40,60	0.78	0.91	0.79	0.92	0.78	0.90	0.79	0.92
		56,84	0.78	0.89	0.79	0.91	0.78	0.89	0.79	0.90
		80,120	0.79	0.88	0.80	0.90	0.79	0.88	0.80	0.90

Note: un-UNAR = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; un-UNCS = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; st-UNAR = structured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; st-UNCS = structured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; MLE=Maximum likelihood estimator; MVE= minimum volume ellipsoid; LDA= Linear Discriminant Analysis;Pop1=Population 1; Pop2=Population 2

Table 3.11: Class-Specific Accuracies of Repeated Measures LDA procedures based on MLE and Robust Estimator (MVE) for Cauchy distribution by Number of Outcomes and Sample Sizes

Methods			un-UNAR		un-UNCS		st-UNAR		st-UNCS	
		Sample Size (n_1, n_2)	Pop1	Pop2	Pop1	Pop2	Pop1	Pop2	Pop1	Pop2
Number of outcomes(*q=3)*										
MLE	Equal	50,50	0.47	0.71	0.47	0.71	0.47	0.71	0.47	0.70
		70,70	0.48	0.69	0.48	0.68	0.48	0.69	0.48	0.68
		100,100	0.47	0.66	0.47	0.66	0.47	0.66	0.47	0.66
	Unequal	40,60	0.47	0.69	0.47	0.69	0.47	0.69	0.47	0.69
		56,84	0.47	0.68	0.47	0.66	0.47	0.67	0.47	0.66
		80,120	0.47	0.66	0.46	0.65	0.47	0.66	0.46	0.65
MVE	Equal	50,50	0.59	0.80	0.59	0.80	0.60	0.80	0.59	0.80
		70,70	0.62	0.79	0.62	0.79	0.61	0.78	0.62	0.79
		100,100	0.63	0.77	0.63	0.77	0.63	0.77	0.63	0.78
	Unequal	40,60	0.58	0.79	0.58	0.79	0.58	0.79	0.58	0.79
		56,84	0.61	0.78	0.61	0.79	0.61	0.78	0.61	0.79
		80,120	0.63	0.77	0.63	0.77	0.63	0.77	0.63	0.77
Number of outcomes(*q=7)*										
MLE	Equal	50,50	0.49	0.74	0.48	0.74	0.49	0.74	0.48	0.74
		70,70	0.48	0.72	0.48	0.72	0.48	0.72	0.48	0.73
		100,100	0.48	0.71	0.48	0.71	0.48	0.71	0.48	0.72
	Unequal	40,60	0.48	0.72	0.48	0.72	0.48	0.73	0.48	0.72
		56,84	0.48	0.72	0.48	0.73	0.48	0.72	0.48	0.73
		80,120	0.48	0.70	0.48	0.70	0.48	0.70	0.48	0.71
MVE	Equal	50,50	0.58	0.82	0.58	0.83	0.58	0.82	0.58	0.83
		70,70	0.64	0.82	0.65	0.82	0.64	0.82	0.65	0.82
		100,100	0.67	0.81	0.68	0.82	0.67	0.81	0.68	0.82
	Unequal	40,60	0.54	0.81	0.54	0.82	0.54	0.81	0.54	0.82
		56,84	0.61	0.81	0.62	0.82	0.61	0.81	0.62	0.82
		80,120	0.67	0.81	0.68	0.81	0.67	0.80	0.68	0.81

Note: un-UNAR = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; un-UNCS = unstructured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; st-UNAR = structured means-Kronecker product of unstructured between-outcomes and within-outcome AR-1 correlation matrices; st-UNCS = structured means-Kronecker product of unstructured between-outcomes and within-outcome CS correlation matrices; MLE=Maximum likelihood estimator; MVE= minimum volume ellipsoid; LDA= Linear Discriminant Analysis; Pop1=Population 1; Pop2=Population 2

Chapter 4

Repeated Measures Discriminant Analysis using Multivariate Generalized Estimation Equations

Brobbey A., Wiebe S., Nettel-Aguirre A., Josephson CB., Williamson T., Lix LM., Sajobi T. Repeated measures discriminant analysis using multivariate generalized estimation equations. *Statistical Methods in Medical Research* (Under review).

AB's contribution to this manuscript includes development of new discriminant analysis procedures, design and implementation of simulation study, manuscript preparation, and analysis of example dataset. ANA, TW, LML and TS provided statistical and methodological expertise support. SW and CBJ revised the manuscript for important clinical content and insights. All co-authors provided supervisory support, reviewed the results and critically revised the manuscript. AB assumes responsibility for the integrity of the manuscript. This manuscript in its entirety is included in Chapter 4.

.

Abstract

In bio-medical sciences, studies are often designed to investigate changes in multivariate repeated measures, where one or more outcomes are measured repeatedly over time in the participating subjects. Many statistical procedures have been proposed for the analysis of multivariate repeated measures data and their extension to discriminant analysis. However, most of these procedures rely on the assumptions of multivariate normality and/or correct specification of the correlation and/or mean structures which may not be tenable in multivariate repeated measures designs which are characterized by binary, ordinal, or mixed types of outcome distributions. This study investigates the accuracy of repeated measures discriminant analysis based on the multivariate generalized estimating equation (GEE) framework for classification in multivariate repeated measures designs with the same or different types of outcomes repeatedly measured over time. Monte Carlo methods were used to compare the classification accuracy of repeated measures discriminant analysis procedures based on multivariate GEE, and repeated measures discriminant analysis based on maximum likelihood estimators (MLE) under diverse simulation generation conditions, which included number of repeated measure occasions, number of outcomes, sample size, correlation structures, and type of outcome distribution. Repeated measures discriminant analysis based on multivariate GEE exhibited higher mean classification accuracy than repeated measures discriminant analysis based on MLE especially in multivariate non-normal distributions. Three repeatedly measured outcomes namely severity of epilepsy, current number of anti-epileptic drugs (AEDs), and parent-reported quality of life in children with epilepsy were classified into remission and refractory groups within two years.

Keywords*:* discriminant analysis, multivariate repeated measures data, generalized estimating equation, multivariate non-normal distribution, classification

4.1 Introduction

In more recent years, relevant work has been done in capturing the longitudinal nature of clinical data and using it for classification via discriminant analysis. These research studies include discriminant analysis extensions to repeated measures data with multiple outcomes [1-8]. When multiple outcomes need to be analyzed, a joint model is required, which extends beyond the correlation between repeated measurements of one outcome. Rather, the model should also allow for a correlation structure between the different outcomes. Utilizing the correlation structure across outcomes with a multivariate model, could increase the classification accuracy[9]. Classical discriminant analysis does not model the correlation structure and thus the information regarding the possible structure in the correlation for repeated measurements taken on the same individual and between outcomes is lost[10-13]. Moreover, classical discriminant analysis is based on multivariate normality assumption to guarantee an optimal solution. Equal correlation structures is assumed in the population groups[10] for linear discriminant analysis (LDA) whilst quadratic discriminant analysis (QDA) allows for unequal covariance structures between the population groups[11-13].

Most LDA methodologies in multivariate repeated measures data are based on mixed effects model. Multivariate linear and non-linear mixed-effects models that assumes unstructured[1, 14] and parsimonious structure[6, 9, 15, 16] for the variance-covariance matrix have been introduced. For instance, several continuous markers and a multivariate linear mixed model was used to evaluate a prognosis of primary biliary cirrhosis patients[14] and non-linear mixed-effects model to distinguish between women with and without pregnancy abnormalities[15]. Similarly, three continuous markers were used to classify patients suffering from prostate cancer[6]. Generalized linear mixed models have been extended in multivariate repeated measures studies for different

type of outcomes (continuous, counts and binary)[1, 2, 17]. Most mixed-effects model LDA assume that the random effects follow a multivariate normal distribution. Moreover, the dimension of random-effects quickly increases as more outcomes and more measurements occasions are added to the model, increasing the computational burden and instability[1, 7, 14]. In addition, it is difficult to evaluate the marginal likelihood of jointly generalized linear mixed models when the outcome is non-normal.

Contrary to mixed effects models approaches, some researchers have utilized generalized estimating equations (GEE) based on multiple marginal models of multiple outcomes. To avoid the specification of the full likelihood function especially for discrete data, multivariate GEE[18] is a suitable approach for parameter estimation for repeated measures data without full specification of the likelihood. GEEs offer a computationally non-intensive parameter estimation algorithm and the resulting parameter estimates have population-averaged interpretation. A joint modeling of multiple outcome variables is based on straightforward extension of univariate GEEs with correlation structure across outcomes which provides separate set of regression parameters for each outcome variable[19, 20]. Specifically, GEEs directly specify a marginal mean model for each outcome and induce the correlation between measurements of outcomes through a working correlation matrix. GEEs are less sensitive to covariance misspecification compared to mixed effects models [18, 21].

This study examines the accuracy of discriminant analysis based on multivariate GEE framework for classification in multivariate repeated measures designs with same/different types of outcomes. The manuscript is organized as follows. In sections 2, we describe the GEEs framework for multivariate repeated measures data. The proposed approach, the extension of the multivariate GEE framework to discriminant analysis, is presented in Section 3. In section 4, we

summarize the results of a Monte Carlo simulation study to assess the validity of the proposed GEE repeated measures discriminant analysis approach under diverse simulation scenarios. Data from a multivariate longitudinal study of children with epilepsy were used to demonstrate the application of these procedures in section 5. Finally, a discussion of the key findings from the study and its implications are described in section 6.

4.2 Generalized Estimating Equations for Multivariate Repeated Measures Data

Suppose we have a random sample of *n* individuals. For each individual $i = 1, \ldots, n$, let $\mathbf{y}_i = (\mathbf{y}'_{i1}, \mathbf{y}'_{i2}, \ldots, \mathbf{y}'_{iq})'$ be a pq x 1 vector of q correlated outcomes that are each repeatedly measured at *p* occasions, and $\mathbf{X}_i = \mathbf{X}_{i*} \otimes \mathbf{I}_q$ is a corresponding pq x Kq block diagonal covariate matrix, where $\mathbf{X}_{i*} = (\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots \mathbf{X}_{ik}, \ldots, \mathbf{X}_{iK})$ is a p x K matrix of covariates and $\mathbf{X}_{ik} = (\mathbf{X}_{i1k}, \ldots\ldots, \mathbf{X}_{ipk})$, $\mathbf{I}_q$ is an q x q identity matrix, and $\otimes$ is the Kronecker product sign. For the analysis of multivariate correlated data, the marginal mean vector $\boldsymbol{\mu}_i = (\boldsymbol{\mu}'_{i1}, \boldsymbol{\mu}'_{i1}, \ldots, \ldots, \boldsymbol{\mu}'_{iq})'$ is associated with *K* covariates through a generalized linear model (GLM) as follows:

$$\boldsymbol{\mu}_i = \mathbf{f}_l(\mathbf{X}_i\boldsymbol{\beta}), \tag{4.1}$$

where, $\mathbf{f}_l(\cdot)$, $l = 1,2, \ldots, q$ is the inverse outcome-specific link function, $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \ldots, \ldots, \boldsymbol{\beta}'_q)'$, where $\boldsymbol{\beta}_q = (\boldsymbol{\beta}_{q1}, \boldsymbol{\beta}_{q2}, \ldots \boldsymbol{\beta}_{qk}, \ldots, \boldsymbol{\beta}_{qK})'$ is the $pq \times 1$ dimensional vector of the qth outcome regression coefficients with population-averaged interpretations. The $pq \times pq$ marginal covariance matrix is:

$$\boldsymbol{\Omega}_i = \phi\boldsymbol{\Sigma}_i, \tag{4.2}$$

where ϕ is a scale parameter that can be known or estimated and $\boldsymbol{\Sigma}_i$ is an $pq \times pq$ working covariance matrix, which results in a a total of $pq(pq + 1)/2$ unknown parameters to be estimated for any statistical inference [22, 23] which may not always be feasible (pq is close to n).

To reduce the dimension of the unknown parameters of the correlation matrix, a parsimonious structure is sometimes used, such as a Kronecker product correlation matrix such that

$$\boldsymbol{\Sigma}_i = \mathbf{A}_i^{1/2}(\mathbf{R}_q(\boldsymbol{\alpha}) \otimes \mathbf{R}_p(\boldsymbol{\rho}))\mathbf{A}_i^{1/2} \tag{4.3}$$

where $\mathbf{A_i}$ is an $pq \times pq$ block diagonal matrix, which contains the marginal variance of outcomes on the main diagonals, $\mathbf{R}_q(\alpha)$ is a $q \times q$ working correlation matrix of the outcomes with the parameter vector α, and $\mathbf{R}_p(\rho)$ is a $p \times p$ the working correlation matrix for a given outcome at different time points with the parameter ρ. This structure reduces the number of correlation parameters to estimate[23-27]. Consequently, $\mathbf{R}_q(\boldsymbol{\alpha})$ and $\mathbf{R}_p(\boldsymbol{\rho})$ denote between-outcomes correlation matrix and within-outcome correlation matrix respectively. Further assuming a structured working correlation, such as exchangeable (EX), first-order autoregressive (AR-1), or unstructured (UN), for $\mathbf{R}_q(\boldsymbol{\alpha})$ and exchangeable (EX) or unstructured (UN) structures for $\mathbf{R}_p(\boldsymbol{\rho})$ can lead to an even more parsimonious model [22, 28, 29]. The parsimonious structure provides flexible model for the complex correlation, particularly when sample size is small[22, 28, 29]. Inferences of interest are easily influenced by the correlation structure's assumptions and unstructured correlation structure might cause convergence problems as the number of parameters to be estimated grows rapidly [30]. In the quasi-likelihood framework with repeated measures outcomes, the regression coefficients $\boldsymbol{\beta}$ can be estimated by solving the generalized estimating equations (GEEs)

$$\mathrm{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{D}_i' \, \boldsymbol{\Omega}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \tag{4.4}$$

where $\mathbf{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$ is the block diagonal matrix of derivatives of the mean with respect to the regression parameters, $\boldsymbol{\mu}_i$ is the marginal mean vector, and $\boldsymbol{\Omega}_i$ is the working covariance matrix. Specifically, $\mathrm{U}(\boldsymbol{\beta}) = \mathbf{0}$ are solved with a Fisher-Scoring algorithm such that

$$\widehat{\boldsymbol{\beta}} = \widetilde{\boldsymbol{\beta}} + \left(\sum_{i=1}^{n} \widetilde{\mathbf{D}}_i' \widetilde{\boldsymbol{\Omega}}_i^{-1} \widetilde{\mathbf{D}}_i\right)^{-1} \left(\sum_{i=1}^{n} \widetilde{\mathbf{D}}_i' \widetilde{\boldsymbol{\Omega}}_i^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)\right) \tag{4.5}$$

Under mild regularity conditions, the parameter estimates are consistent and asymptotically normally distributed even when the "working" correlation structure of outcomes is mis-specified, and the variance-covariance matrix can be estimated using a robust "sandwich" variance estimator[31]. The asymptotic covariance matrix of the non-vanishing (non-zero) component of $\widehat{\boldsymbol{\beta}}$ via the sandwich estimator formula is [31, 32]:

$$\widehat{\mathrm{cov}}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{n} \widehat{\boldsymbol{D}}_i' \widehat{\boldsymbol{\Omega}}_i^{-1} \widehat{\boldsymbol{D}}_i\right)^{-1} \widehat{\boldsymbol{M}}_* \left(\sum_{i=1}^{n} \widehat{\boldsymbol{D}}_i' \widehat{\boldsymbol{\Omega}}_i^{-1} \widehat{\boldsymbol{D}}_i\right)^{-1}, \tag{4.6}$$

with

$$\widehat{\boldsymbol{M}}_* = \sum_{i=1}^{n} \widehat{\boldsymbol{D}}_i' \widehat{\boldsymbol{\Omega}}_i^{-1} \widehat{\mathrm{cov}}(\mathbf{y}_i) \widehat{\boldsymbol{\Omega}}_i^{-1} \widehat{\boldsymbol{D}}_i \tag{4.7}$$

and $\widehat{\mathrm{cov}}(\mathbf{y}_i) = (\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_i)'$ is an estimator of the true variance-covariance matrix of $\mathbf{y}_i$[18, 31]. Note that if $\boldsymbol{\Omega}_i$ is correctly specified, $\boldsymbol{\Omega}_i = \mathrm{cov}(\mathbf{y}_i)$[33, 34]. Moreover, GEE requires the correct specification of marginal mean and variance as well as the link function, which connects the covariates of interest and the marginal means.

4.3 GEE Extension to Multivariate Repeated Measures Discriminant Analysis

Following the GEE notation, we assume that the *i*th individual in the *j*th population (j =1,2) with multivariate repeated outcomes $\mathbf{y}_{ij}$, has a marginal mean $\boldsymbol{\mu}_j$, and variance

covariance matrix $\boldsymbol{\Omega}_j$ assumed to be *pq* x *pq* positive definite. Analogously, with estimations of $\widehat{\boldsymbol{\mu}}_j = \boldsymbol{f}_q(\boldsymbol{X}_i\widehat{\boldsymbol{\beta}}_j)$ and the variance covariance matrix $\widehat{\boldsymbol{\Omega}}_j$ from the GEE model in population *j* using a pre-defined structure, the homoscedastic model is obtained when the variance components are homogeneous, that is, $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2 = \boldsymbol{\Omega}$, the pooled covariance matrix. Based on LDA, a randomly selected *i*th individual with multiple outcome vector $\mathbf{y}_i$ is classified in the first population, if

$$\left(\mathbf{y}_i - \frac{\widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_2}{2}\right)' \widehat{\boldsymbol{\Omega}}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2) > \log \frac{\hat{\pi}_2}{\hat{\pi}_1} \tag{4.8}$$

where $\widehat{\boldsymbol{\mu}}_j$ and $\widehat{\boldsymbol{\Omega}}_j^{-1}$are the GEE estimates from (1) and (2), $\hat{\pi}_1$ and $\hat{\pi}_2$ are the *a priori* probabilities that observations belong to populations 1 and 2. Otherwise, is the individual is classified into the second population. For QDA (i.e., $\boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_2$), the *i*th subject with multiple outcome vector $\mathbf{y}_i$ is classified in the first population, if

$$(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_2)'\widehat{\boldsymbol{\Omega}}_2^{-1}(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_2) - (\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_1)'\widehat{\boldsymbol{\Omega}}_1^{-1}(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_1) > \log \left|\frac{\widehat{\boldsymbol{\Omega}}_1}{\widehat{\boldsymbol{\Omega}}_2}\right| + 2\log \frac{\hat{\pi}_2}{\hat{\pi}_1} \tag{4.9}$$

otherwise, it is classified into the second population.

4.4 Simulation Study

A Monte Carlo simulation study was conducted to examine the accuracy of linear and quadratic GEE discriminant analysis procedures that assume Kronecker product structured covariances compared to conventional discriminant analysis based on MLE for multivariate repeated measures data. The following conditions were investigated: (a) number of repeated measurements (p), (b) total sample size (N), (c) group sizes (n_1, n_2), (d) pattern and magnitude of correlation among the repeated measurements (ρ), (e) mean configuration, (f) covariance heterogeneity, and (g) population distribution. All procedures were investigated for two

independent groups. The number of repeated occasions/ time points was set at $p = 3$ and 5, and number of outcomes was set at $q = 3$ and 5. Previous studies about DA procedures for multivariate repeated measures data have considered p ranging from three to ten, an increase in classification accuracies were quite significant when p increased from three to five. [35, 36] Total sample sizes of $N = 80$, 140 and 200 were investigated. This is consistent with previous simulation studies that examined the accuracy of DA for multivariate repeated measures data between 60 and 200. Moreover, consistent with previous studies that examined the impact of equal and unequal group sizes [35, 36,37, 38], we investigated conditions of $N = 80$, $(n_1, n_2) = (40, 40)$, and $(32, 48)$, which represent a group size ratio of 1:1 and 2:3, respectively. Similar equal and unequal group size ratios were investigated when $N = 140$ and $N = 200$. Furthermore, the accuracy of DA procedures is known to be influenced by both the magnitude and pattern of within- and multivariate-outcome correlations[39]. Therefore, we investigated the following within-outcome correlation structures: (a) Compound Symmetry with $\rho = 0.3$ and $\rho = 0.7$, (b) Autoregressive order 1 with $\rho = 0.3$ and $\rho = 0.7$[35, 36] for the within-outcome correlation $\mathbf{R}_p(\rho)$, and the between-outcomes correlation, $\mathbf{R}_q(\rho)$ was assumed to be unstructured (See Table 4.1 for more details). Hence, we assumed two Kronecker correlation structures $\mathbf{R}_q(\alpha) \otimes \mathbf{R}_p(\rho)$; UNAR = Unstructured between-outcomes and Autoregressive order-1 within-outcome correlation matrix, and UNCS=Unstructured between-outcomes and Compound symmetry within-outcome correlation matrix. For covariance heterogeneity, we assumed $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$ and $\boldsymbol{\Omega}_1 = 3\boldsymbol{\Omega}_2$.

In order to assess the performance of the discriminant function, we investigated multivariate correlated continuous outcome variables, count outcome variables and different types of correlated outcomes, namely Case 1, Case 2 and Case 3 respectively. Case 1: For the correlated continuous outcome variables, we assumed three normal variables jointly observed for $\boldsymbol{n}_j$ subjects,

where each observed at p time points. The true marginal mean outcome model $\boldsymbol{\mu}_{ipq}$ was assumed to take the following functional form that uses an identity link function:

$$\boldsymbol{\mu}_{ipq} = \boldsymbol{\beta}_{q1} x_{ip} + \boldsymbol{\beta}_{q2} t_{ip} \quad (4.2)$$

The number of covariates, $K = 2$, where x_{ip} was generated from an independent normal random variable $N(0,1)$ as a time-invariant covariate, and t_{ip} denoted the time of observation as a time-varying covariate. Details of the true parameters $\boldsymbol{\beta}$ for population 1 and population 2 can be found in Table 4.2. On the other hand, the marginal variance matrix of outcomes was assumed to have a common variance of 60. Case 2: For the multivariate count outcome variables, data were generated from a multivariate Poisson distribution using the log link function instead of identity link in Case 1 and log transformation of time of observation as a time-varying covariate.

$$\boldsymbol{log}(\boldsymbol{\mu}_{ipq}) = \boldsymbol{\beta}_{q1} x_{ip} + \boldsymbol{\beta}_{q2} \log(t_{ip}) \quad (4.3)$$

The true parameters $\boldsymbol{\beta}$ for population 1 and population 2 can be found in Table 4.2. Case 3: For generating different types of correlated outcomes, one of the outcomes generated from case 1 (multivariate normal distribution data) was converted to Bernoulli outcome using the NORmal-To-Anything (NORTA) algorithm[40] with probabilities from the logit function.

The LDA and QDA rules were developed using marginal mean and variance-covariance matrix estimated via GEE, and MLE for equal and unequal covariance matrix respectively. The classification performance of the procedures was evaluated using the overall mean classification accuracy and its corresponding standard errors.

$$\text{Overall classification accuracy} = \frac{\text{correct classifications}}{\text{Total sample size } (N)} \quad ((4.4)$$

All combinations of simulation generation conditions were investigated for each procedure and each method of estimation, resulting in a total of 194 combinations. There were 500 replications for each combination. All analyses were completed in R statistical software version 3.5.3.

4.4.1 Simulation Study Results

Tables 4.3 and 4.4 describe the mean classification accuracies and standard errors of repeated measures linear and quadratic discriminant analysis based on GEE, and MLE respectively by population distribution, number of repeated occasions and number of outcomes. The results of linear discriminant analysis showed 0.01-0.04 differences among all UNAR procedures when data were sampled from a multivariate normal distribution; however, repeated measures discriminant analysis based on GEE procedures were more accurate than repeated measures discriminant analysis based on MLE among UNCS procedures. For example: for the UNCS correlation matrix under GEE, the mean accuracy for $p = 3$ was 0.74 and $p = 5$, it was 0.89, whiles for the UNCS correlation matrix under MLE, the mean accuracy for $p = 3$ was 0.66 and $q = 5$, it was 0.63 when the number of outcomes was five (Table 4.3).

Moreover, repeated measures discriminant analysis based on GEE had the highest mean classification accuracy compared to repeated measures discriminant analysis procedures based on MLE when outcomes were sampled from a multivariate Poisson distribution and mixed type outcomes. For example, when p = 3 and q=5, the mean classification accuracies of repeated measures discriminant analysis procedures based on GEE and MLE were 0.97 and 0.84 when data were sampled from a multivariate Poisson distribution with outcome variables. Whereas the mean accuracy of the GEE and MLE procedures were 0.72 and 0.58, respectively, when mixed type outcomes, under UNAR correlation matrix (Table 4.3). In the quadratic discriminant analysis

procedures, repeated measures discriminant analysis procedures based on MLE were least accurate regardless of number of repeated occasions, number of outcomes, estimation method or multivariate distribution of outcome variables (Table 4.4). For example, when $q = 5$ and $p=3$ under UNAR correlation matrix, the mean classification accuracies of repeated measures discriminant analysis procedures based on GEE and MLE were 0.85 and 0.66 when data were sampled from a multivariate normal distribution with outcome variables.

Furthermore, the mean accuracy of each linear and quadratic discriminant analysis procedures increased as the number of repeated occasions and number of outcomes increased, regardless of the estimation method or multivariate distribution of outcome variables. For example, for $q = 3$ when data were sampled from a multivariate normal distribution, the increase in mean classification accuracy of the repeated measures discriminant analysis procedure based on GEE and MLE were about 0.11 and 0.05 respectively as p increased from 3 to 5, under UNCS correlation matrix (Table 4.3). Likewise, the increase in mean classification accuracy of the repeated measures discriminant analysis procedure based on GEE and MLE were about 0.10 and 0.01 respectively as q increased from 3 to 5, under UNCS correlation matrix and $p = 3$ (Table 4.3).

It is worth mentioning that, we observed little or no differences in classification accuracies for linear and quadratic discriminant procedures when repeated measures discriminant analysis procedures based on MLE were used, whereas the classification accuracies for quadratic discriminant procedures based on GEE increased compared to its corresponding linear discriminant procedures (Table 4.3 & Table 4.4).

For example: the mean accuracy for repeated measures discriminant analysis procedure based on GEE and MLE were 0.64 and 0.65 respectively for linear discriminant procedure (Table 4.3), whiles for quadratic discriminant procedure, the mean accuracy were 0.80 and 0.66 respectively

(Table 4.4) under the UNCS correlation matrix, when data were sampled from a multivariate normal distribution with outcome variables and p =3.

4.5 Application: Health-Related Quality of Life in Children with Epilepsy Study (HERQULES)

Multivariate repeated measures data were obtained from the Health-Related Quality of Life (HRQOL) in Children with Epilepsy Study (HERQULES), a two-year prospective cohort study assessing the course and characteristics potentially associated with HRQOL in children with new onset epilepsy across Canada[41, 42]. Details of HERQULES have been described elsewhere[41, 42]. Data were collected as soon as possible following the diagnosis of epilepsy at baseline (0 month), and approximately 6 months, 12 months, and 24 months later (p=4). Standardized questionnaires were used to collect parent-report of their children's HRQOL and a series of child and family characteristics, while a neurologist-report form collected information on clinical characteristics of the child's epilepsy.

Using this multivariate repeated measures data, we sought to identify patients who will not achieve remission from seizures within two years from disease onset. Early identification of patients who have refractory epilepsy can allow clinicians to explore alternative treatment options (e.g., surgery) to manage seizures and other aspects of the disease[2]. Data for this numeric example consists of outcome variables (q=3) such as severity of epilepsy, current number of anti-epileptic drugs (AEDs), and parent-reported quality of life in children using epilepsy-specific scale which were measured over four measurement occasions (p=4) and covariates such as time of observation, age at seizure onset, and sex. All repeated measures data were used for the classification. Repeated measures linear and quadratic discriminant analysis classification rules were developed based on multivariate GEE model using this data .

Of the 187 patients include in this analysis, 101 patients were in the remission group (n_1 =101) and 86 patients were in the refractory group (n_2 = 86) within two years. The sample included children ages 4 to 12 years. The mean age (standard deviation) in the remission group was 8.25(2.46) years and in the refractory group was 8.25(2.46) years. The patients included 45.54% and 41.86% females in the remission and refractory groups respectively. The QOLCE-55 ratings underwent a linear transformation such that domain scores can take values from 0 (low HRQOL) to 100 (high HRQOL). The ratings were treated as a continuous variable. The GASE scale is a 7-point Likert scale ranging from 1 (not severe at all) to 7 (extremely severe) was recoded as a binary variable, with $\geq$ 3 coded as severe thereby using the median severity 3 of the sample, corresponding to ‘‘somewhat severe” as a cut-off [43].

4.5.1 Results for HERQULES Data

Figure 4.1 describe the longitudinal changes in the levels of each of the outcome variables for all patients in each diagnostic group. For patients who achieved remission, severity of seizures appears to decrease over time whereas seizure severity remained high for the refractory group. The difference between the overall quality of life of the two groups is less noticeable. However, the overall quality of life appears constant over time in the refractory but as time increases the overall quality of life of the remission patients gradually increases. The number of AEDs increased over time for the refractory patients while those in the remission group had slightly reduced number of AEDs.

Table 4.5 gives the group-specific correlation parameter estimates of the joint modeling of the multiple repeated outcomes using multivariate GEE. We observed that in both remission and refractory groups, HRQOL was negatively associated with severity of seizures and the number of AEDs. However, there was little to no association between severity of seizures and the number of

AEDs. The accuracy of LDA and QDA classifiers based on GEE and maximum likelihood estimators are described in Table 4.6. Overall, repeated measures discriminant analysis procedures based on GEE exhibited higher overall classification accuracy than repeated measures discriminant analysis based on MLE in both LDA and QDA. Moreover, the classification accuracies observed using GEE estimators increased when QDA (accuracy, 0.79) was used for classification compared to its LDA (accuracy, 0.71) approach whilst the accuracy using MLE estimators for remain the same for both QDA and LDA (accuracy, 0.67). The classifiers were more accurate in correctly reclassifying patients in the remission group but less accurate for reclassifying those in the refraction group.

4.6 Discussion

This study investigates discriminant analysis procedures for multivariate repeated measures data using multivariate GEE for discriminating between population groups. The proposed approach allows the incorporation of repeated measures outcomes and covariates to improve the accuracy of the classifier. Our results showed that the repeated measures discriminant analysis based on multivariate GEE model resulted in better classification accuracy than the conventional repeated measures discriminant analysis based on maximum likelihood estimators especially in multivariate repeated measures data with discrete and/or mixed type of outcomes [44, 45]. This is because the GEE approach enables us to analyze multivariate repeated measures data all together regardless of the type of outcomes, without specifying of a full likelihood[20, 30, 45, 46].

Furthermore, our study revealed the impact of increasing repeated occasions and number of outcomes on the accuracy of the investigated procedures. The impact of increasing number of repeated occasions is consistent with literature on other repeated measures discriminant analysis

methods[22, 36] ; however, these studies did not investigate the impact of increasing number of outcomes. Specifically, the repeated measures discriminant analysis based on GEE was most accurate for increase in the number of repeated occasions and number of outcomes compared to repeated measures discriminant analysis based on MLE. Overall, the quadratic discriminant analysis was able to better classify individuals than the linear discriminant analysis in repeated measures discriminant analysis based on GEE. QDA provides a less restrictive procedure by allowing different covariance matrix for each population group, which minimizes misclassification. Even though, classification rules based on LDA can perform badly if the assumption of a common within-class covariance matrix is violated, classification rules based on QDA requires a larger sample size to overcome the singularity problem[13, 47, 48]. Also, the procedures developed in this study are based on two-group multivariate repeated designs, but our conclusions can be extended and generalized to multi-group designs[49, 50].

Despite the unique strengths of this class of repeated measures discriminant analysis models, they are not without their own limitations. First, the repeated measures discriminant analysis based on multivariate GEE relies on correctly specified link function and parsimonious covariance structures, which might not be realistic in typical multivariate repeated measures data. It is well known that GEEs yield asymptotically consistent parameter and variance estimates even under incorrect specification of the correlation structure but correctly specified link function [44, 46, 51]. This means that a crucial step in the GEE approach is to select a correct link function linking the mean response to the covariates[52]. With regards to parsimonious covariance structures, even though several authors have observed many advantages of using Kronecker product structure for analyzing multivariate repeated measures data[22, 24, 36, 53, 54], one could use the usual unstructured variance covariance matrix when there is sufficient data. Moreover, some work has been done on

the testing of hypotheses of Kronecker product structure[22, 24, 26]. It is also not clear whether the misspecification of the working c correlation structures for these procedures could influence their classification accuracy[55]. However, one does not know a priori which correlation structure is correct. Future research will examine the impact of misspecification of correlation structure on the accuracy of these classifiers. In addition, to help in choosing a working correlation matrix that is close to the true correlation matrix, a quasi-likelihood under the independence model criterion (QIC) which is a modified Akaike information criterion (AIC) has recommended for GEE model[56, 57]. Secondly, the assumption of complete multivariate repeated measures data in which there is no missing data on all outcomes and at all measurement occasions might not be realistic in multivariate repeated measures data often encountered in applied research. Even in a well-controlled repeated measures study, missing data may frequently occur due to missed visits, withdrawal from the study, or loss to follow-up[20]. Some studies have been done to drop-out problems in repeated measures studies via weighted generalized estimating equations[58] and imputations. Further research could extend the discriminant analysis procedures based on GEE by implementing some of the multiple imputation techniques[20, 59-61].

In summary, this study proposes a new class of discriminant analysis procedures based on multivariate GEE, which can be used for distinguishing between population groups in multivariate repeated measures data characterized by multivariate non-normal distributions with mixed types of outcome variables. An advantage of these procedures is their ability to accommodate both time-invariant and time-varying covariate to improve the accuracy of model classifiers.

Acknowledgments

This research was supported by a Natural Sciences & Engineering Research Council Discovery Grant to TTS. The authors are grateful to Dr Kathy Speechley for sharing data from the HERQULES study which was used as a numeric example in this study

References

1. Fieuws S, Verbeke G, Maes B, et al. Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics* 2007; 9: 419-431.
2. Hughes DM, Komárek A, Czanner G, et al. Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types. *Statistical methods in medical research* 2018; 27: 2060-2080.
3. Inoue LYT, Etzioni R, Morrell C, et al. Modeling disease progression with longitudinal markers. *Journal of the American Statistical Association* 2008; 103: 259-270.
4. Li Y, Wang Y, Wu G, et al. Discriminant analysis of longitudinal cortical thickness changes in alzheimer's disease using dynamic and network features. *Neurobiology of aging* 2012; 33: 427. e415-427. e430.
5. Marshall G and Barón AE. Linear discriminant models for unbalanced longitudinal data. *Statistics in medicine* 2000; 19: 1969-1981.
6. Morrell CH, Brant LJ, Sheng S, et al. Screening for prostate cancer using multivariate mixed-effects models. *Journal of applied statistics* 2012; 39: 1151-1175.
7. Verbeke G, Fieuws S, Molenberghs G, et al. The analysis of multivariate longitudinal data: A review. *Statistical methods in medical research* 2014; 23: 42-59.
8. Galecki AT. General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communications in Statistics-Theory and Methods* 1994; 23: 3105-3119.
9. Marshall G, De la Cruz-Mesía R, Barón AE, et al. Non-linear random effects model for multivariate responses with missing data. *Statistics in Medicine* 2006; 25: 2817-2830.
10. Lachenbruch PA and Goldstein M. Discriminant analysis. *Biometrics* 1979: 69-85.
11. Marks S and Dunn OJ. Discriminant functions when covariance matrices are unequal. *Journal of the American Statistical Association* 1974; 69: 555-559.
12. Flury BW and Schmid MJ. Quadratic discriminant functions with constraints on the covariance matrices: Some asymptotic results. *Journal of multivariate analysis* 1992; 40: 244-261.
13. Wahl PW and Kronmal RA. Discriminant functions when covariances are unequal and sample sizes are moderate. *Biometrics* 1977: 479-484.
14. Komárek A, Hansen BE, Kuiper EM, et al. Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statistics in Medicine* 2010; 29: 3267-3283.
15. Marshall G, De la Cruz-Mesía R, Quintana FA, et al. Discriminant analysis for longitudinal data with multiple continuous responses and possibly missing data. *Biometrics* 2009; 65: 69-80.
16. Roy A. A new classification rule for incomplete doubly multivariate data using mixed effects model with performance comparisons on the imputed data. *Statistics in medicine* 2006; 25: 1715-1728.
17. Fieuws S, Verbeke G and Molenberghs G. Random-effects models for multivariate repeated measures. *Statistical methods in medical research* 2007; 16: 387-397.
18. Liang K-Y and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73: 13-22.
19. Lipsitz SR, Fitzmaurice GM, Ibrahim JG, et al. Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: An application to acquired immune

deficiency syndrome data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009; 172: 3-20.
20. Inan G and Yucel R. Joint gees for multivariate correlated data with incomplete binary outcomes. *Journal of Applied Statistics* 2017; 44: 1920-1937.
21. Fong Y, Rue H and Wakefield J. Bayesian inference for generalized linear mixed models. *Biostatistics* 2010; 11: 397-412.
22. Roy A and Khattree R. On implementation of a test for kronecker product covariance structure for multivariate repeated measures data. *Statistical Methodology* 2005; 2: 297-306.
23. Srivastava MS, von Rosen T and Von Rosen D. Models with a kronecker product covariance structure: Estimation and testing. *Mathematical Methods of Statistics* 2008; 17: 357-370.
24. Lu N and Zimmerman DL. The likelihood ratio test for a separable covariance matrix. *Statistics & probability letters* 2005; 73: 449-457.
25. Roy A. A note on testing of kronecker product covariance structures for doubly multivariate data. In: *Proceedings of the American Statistical Association, Statistical Computing Section* 2007, pp.2157-2162.
26. Roy A and Khattree R. Tests for mean and covariance structures relevant in repeated measures based discriminant analysis. *Journal of Applied Statistical Science* 2003; 12: 91-104.
27. Werner K, Jansson M and Stoica P. On estimation of covariance matrices with kronecker product structure. *IEEE Transactions on Signal Processing* 2008; 56: 478-491.
28. Filipiak K, Klein D and Roy A. Score test for a separable covariance structure with the first component as compound symmetric correlation matrix. *Journal of Multivariate Analysis* 2016; 150: 105-124.
29. Roy A and Khattree R. Testing the hypothesis of a kronecker product covariance matrix in multivariate repeated measures data. *SAS Users Group International, Proceedings of the Statistics and Data Analysis Section* 2005: 199-130.
30. Cho H. The analysis of multivariate longitudinal data using multivariate marginal models. *Journal of Multivariate Analysis* 2016; 143: 481-491.
31. Chao EC. Generalized estimating equations. Taylor & Francis, 2003.
32. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* 2001; 96: 1348-1360.
33. Kauermann G and Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* 2001; 96: 1387-1396.
34. Wang M. Generalized estimating equations in longitudinal data analysis: A review and recent developments. *Advances in Statistics* 2014; 2014.
35. Roy A and Khattree R. Discrimination and classification with repeated measures data under different covariance structures. *Communications in Statistics—Simulation and Computation®* 2005; 34: 167-178.
36. Roy A and Khattree R. On discrimination and classification with multivariate repeated measures data. *Journal of Statistical Planning and Inference* 2005; 134: 462-485.
37. Barön AE. Misclassification among methods used for multiple group discrimination-the effects of distributional properties. *Statistics in medicine* 1991; 10: 757-766.
38. He X and Fung WK. High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis* 2000; 72: 151-162.

39. Thomas DR and Zumbo BD. Using a measure of variable importance to investigate the standardization of discriminant coefficients. *Journal of Educational and Behavioral Statistics* 1996; 21: 110-130.
40. Cario MC and Nelson BL. *Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix*. 1997. Citeseer.
41. Ferro MA, Camfield CS, Levin SD, et al. Trajectories of health-related quality of life in children with epilepsy: A cohort study. *Epilepsia* 2013; 54: 1889-1897.
42. Speechley KN, Ferro MA, Camfield CS, et al. Quality of life in children with new-onset epilepsy: A 2-year prospective cohort study. *Neurology* 2012; 79: 1548-1555.
43. Speechley KN, Sang X, Levin S, et al. Assessing severity of epilepsy in children: Preliminary evidence of validity and reliability of a single-item scale. *Epilepsy & Behavior* 2008; 13: 337-342.
44. Qu A, Lindsay BG and Li B. Improving generalised estimating equations using quadratic inference functions. *Biometrika* 2000; 87: 823-836.
45. Wang X and Qu A. Efficient classification for longitudinal data. *Computational Statistics & Data Analysis* 2014; 78: 119-134.
46. Asar Ö and İlk Ö. Mmm: An r package for analyzing multivariate longitudinal data with multivariate marginal models. *Computer Methods and Programs in Biomedicine* 2013; 112: 649-654.
47. Lu J, Plataniotis KN and Venetsanopoulos AN. Regularized discriminant analysis for the small sample size problem in face recognition. *Pattern recognition letters* 2003; 24: 3079-3087.
48. Pang H, Tong T and Ng M. Block-diagonal discriminant analysis and its bias-corrected rules. *Statistical applications in genetics and molecular biology* 2013; 12: 347-359.
49. Filzmoser P, Joossens K and Croux C. Multiple group linear discriminant analysis: Robustness and error rate. *Compstat 2006-proceedings in computational statistics*. Springer, 2006, pp.521-532.
50. Croux C, Filzmoser P and Joossens K. Classification efficiencies for robust linear discriminant analysis. *Statistica Sinica* 2008: 581-599.
51. Liang K-Y and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986: 13-22.
52. Molefe AC and Hosmane B. Test for link misspecification in dependent binary regression using generalized estimating equations. *Journal of Statistical Computation and Simulation* 2007; 77: 95-107.
53. Naik DN and Rao SS. Analysis of multivariate repeated measures data with a kronecker product structured covariance matrix. *Journal of Applied Statistics* 2001; 28: 91-105.
54. Krzyśko M and Skorzybut M. Discriminant analysis of multivariate repeated measures data with a kronecker product structured covariance matrices. *Statistical papers* 2009; 50: 817-835.
55. Zhou J and Qu A. Informative estimation and selection of correlation structure for longitudinal data. *Journal of the American Statistical Association* 2012; 107: 701-710.
56. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics* 2001; 57: 120-125.
57. Akaike H. Information theory and an extension of the maximum likelihood principle. *Selected papers of hirotugu akaike*. Springer, 1998, pp.199-213.
58. Beunckens C, Sotto C and Molenberghs G. A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational statistics & data analysis* 2008; 52: 1533-1548.

59. Satty A, Mwambi H and Molenberghs G. Different methods for handling incomplete longitudinal binary outcome due to missing at random dropout. *Statistical Methodology* 2015; 24: 12-27.
60. Yucel RM, He Y and Zaslavsky AM. Using calibration to improve rounding in imputation. *The American Statistician* 2008; 62: 125-129.
61. Yucel RM, He Y and Zaslavsky AM. Imputation of categorical variables using gaussian-based routines. *Statistics in Medicine* 2011; 30: 3447-3460.

Table 4.1: Configuration of unstructured between-outcomes correlation matrix given within-outcome correlation coefficient for the Monte Carlo Study

within-outcome correlation coefficient (ρ)	0.3	0.7
$q = 3$	$\begin{bmatrix} 1 & 0.15 & 0.30 \\ 0.15 & 1 & 0.45 \\ 0.30 & 0.45 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.65 & 0.66 \\ 0.65 & 1 & 0.70 \\ 0.66 & 0.70 & 1 \end{bmatrix}$
$q = 5$	$\begin{bmatrix} 1 & 0.28 & 0.25 & 0.28 & 0.28 \\ 0.28 & 1 & 0.30 & 0.40 & 0.23 \\ 0.25 & 0.30 & 1 & 0.24 & 0.24 \\ 0.28 & 0.40 & 0.24 & 1 & 0.37 \\ 0.28 & 0.23 & 0.24 & 0.37 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.70 & 0.79 & 0.64 & 0.70 \\ 0.70 & 1 & 0.73 & 0.65 & 0.74 \\ 0.79 & 0.73 & 1 & 0.63 & 0.62 \\ 0.64 & 0.65 & 0.63 & 1 & 0.62 \\ 0.70 & 0.74 & 0.62 & 0.62 & 1 \end{bmatrix}$

q=Number of outcomes

Table 4.2: True parameters ($\boldsymbol{\beta}$) for population 1 and population 2 simulated data

Population Distribution	Number of outcomes	population 1	population 2
Normal/ Mixed-type	3	(0.3,1,2,0.1,1,1.5)	(0.6,2,4,0.2,2,3)
	5	(0.2,1,2,1.5,1,0.4,0.7,3,1.2,0.8)	(0.4,2,4,3,2,0.8,1.4,6,2.4,1.6)
Poisson	3	(0.3,0.1,0.2,0.1,0.3,0.5)	(0.9, 0.3, 0.6, 0.3 ,0.9, 1.5)
	5	(0.3,0.1,0.4,0.1,0.45,0.6,0.2,0.15,0.3,0.4)	(0.9, 0.3,1.2,0.3,1.35,1.8,0.6,0.45,0.9,1.2)

Table 4.3: Overall Mean Accuracy (standard error) for repeated measures LDA procedures based on GEE, and MLE by population distribution, number of outcomes, and number of measurements occasions

Population Distribution	Number of outcomes	Number of measurements occasions	GEE		MLE	
			UNAR	UNCS	UNAR	UNCS
Normal	3	3	0.62(0.04)	0.64(0.04)	0.63(0.04)	0.65(0.04)
		5	0.73(0.04)	0.75(0.04)	0.69(0.04)	0.70(0.04)
	5	3	0.68(0.04)	0.74(0.04)	0.66(0.04)	0.66(0.04)
		5	0.83(0.03)	0.89(0.03)	0.82(0.03)	0.63(0.03)
Poisson	3	3	0.88(0.04)	0.90(0.03)	0.79(0.04)	0.81(0.04)
		5	0.97(0.02)	0.97(0.03)	0.84(0.05)	0.85(0.05)
	5	3	0.99(0.01)	0.99(0.01)	0.89(0.04)	0.90(0.04)
		5	0.99(0.01)	0.99(0.01)	0.95(0.02)	0.95(0.02)
Mixed-type	3	3	0.62(0.04)	0.63(0.04)	0.55(0.04)	0.55(0.04)
		5	0.72(0.04)	0.74(0.04)	0.58(0.04)	0.58(0.04)
	5	3	0.68(0.04)	0.72(0.04)	0.67(0.04)	0.57(0.04)
		5	0.81(0.03)	0.87(0.03)	0.62(0.04)	0.62(0.04)

Note: UNAR = Unstructured between-outcomes and autoregressive order 1 within-outcome correlation matrix; UNCS = Unstructured between-outcomes and compound symmetry within-outcome correlation matrix; GEE – Generalized estimating equation; MLE=Maximum likelihood estimation

Table 4.4: Overall Mean Accuracy (standard error) for repeated measures QDA procedures based on GEE, and MLE by population distribution, number of outcomes, and number of measurements occasions

Population Distribution	Number of outcomes	Number of measurements occasions	GEE		MLE	
			UNAR	UNCS	UNAR	UNCS
Normal	3	3	0.77(0.04)	0.80(0.04)	0.65(0.04)	0.66(0.04)
		5	0.85(0.04)	0.88(0.04)	0.71(0.04)	0.71(0.04)
	5	3	0.85(0.04)	0.89(0.04)	0.66(0.04)	0.66(0.04)
		5	0.90(0.03)	0.94(0.03)	0.85(0.03)	0.90(0.02)
Poisson	3	3	0.93(0.03)	0.94(0.03)	0.78(0.04)	0.79(0.04)
		5	0.99(0.01)	0.98(0.03)	0.85(0.05)	0.85(0.05)
	5	3	0.99(0.01)	0.99(0.01)	0.90(0.04)	0.92(0.03)
		5	0.99(0.01)	0.99(0.01)	0.95(0.02)	0.95(0.02)
Mixed-type	3	3	0.74(0.04)	0.75(0.04)	0.56(0.04)	0.55(0.04)
		5	0.84(0.04)	0.85(0.04)	0.58(0.04)	0.58(0.06)
	5	3	0.83(0.04)	0.86(0.04)	0.58(0.04)	0.58(0.04)
		5	0.91(0.03)	0.94(0.03)	0.63(0.04)	0.63(0.04)

Note: UNAR = Unstructured between-outcomes and autoregressive order 1 within-outcome correlation matrix; UNCS = Unstructured between-outcomes and compound symmetry within-outcome correlation matrix; GEE – Generalized estimating equation; MLE=Maximum likelihood estimation

Table 4.5: GEE Group-specific correlation parameter estimates for HERQULES data by the assumed correlation structure

	Remission		Refractory	
	UNAR	UNCS	UNAR	UNCS
ρ	0.812	0.749	0.744	0.726
Corr($\mathbf{Y_2Y_1}$)	-0.025		-0.023	
Corr($\mathbf{Y_3Y_1}$)	0.003		0.001	
Corr($\mathbf{Y_3Y_2}$)	-0.042		-0.038	

UNAR: Unstructured between-outcomes and Autoregressive order 1 within- outcome correlation matrix; UNCS: Unstructured between- outcomes and Compound symmetry within-outcome correlation matrix; Number of (AEDs) (Y_1), HRQOL (Y_2), Severe Seizure(Y_3)

Table 4.6: Classification accuracy for the generalized estimating equation (GEE) and maximum likelihood estimation (MLE) methods for repeated measures LDA and QDA by the assumed correlation structure

		GEE		MLE	
		UNAR	UNCS	UNAR	UNCS
LDA	Remission	0.772	0.770	0.762	0.76
	Refractory	0.651	0.640	0.570	0.558
	Overall	**0.711**	**0.705**	**0.665**	**0.660**
QDA	Remission	0.871	0.880	0.752	0.750
	Refractory	0.709	0.698	0.581	0.570
	Overall	**0.790**	**0.789**	**0.667**	**0.660**

LDA: Linear discriminant analysis; QDA: quadratic discriminant analysis; GEE: Generalized estimating equation; MLE: Maximum likelihood estimation; UNAR: Unstructured between outcomes and Autoregressive order 1 within outcome correlation matrix; UNCS: Unstructured between outcomes and Compound symmetry within outcome correlation matrix

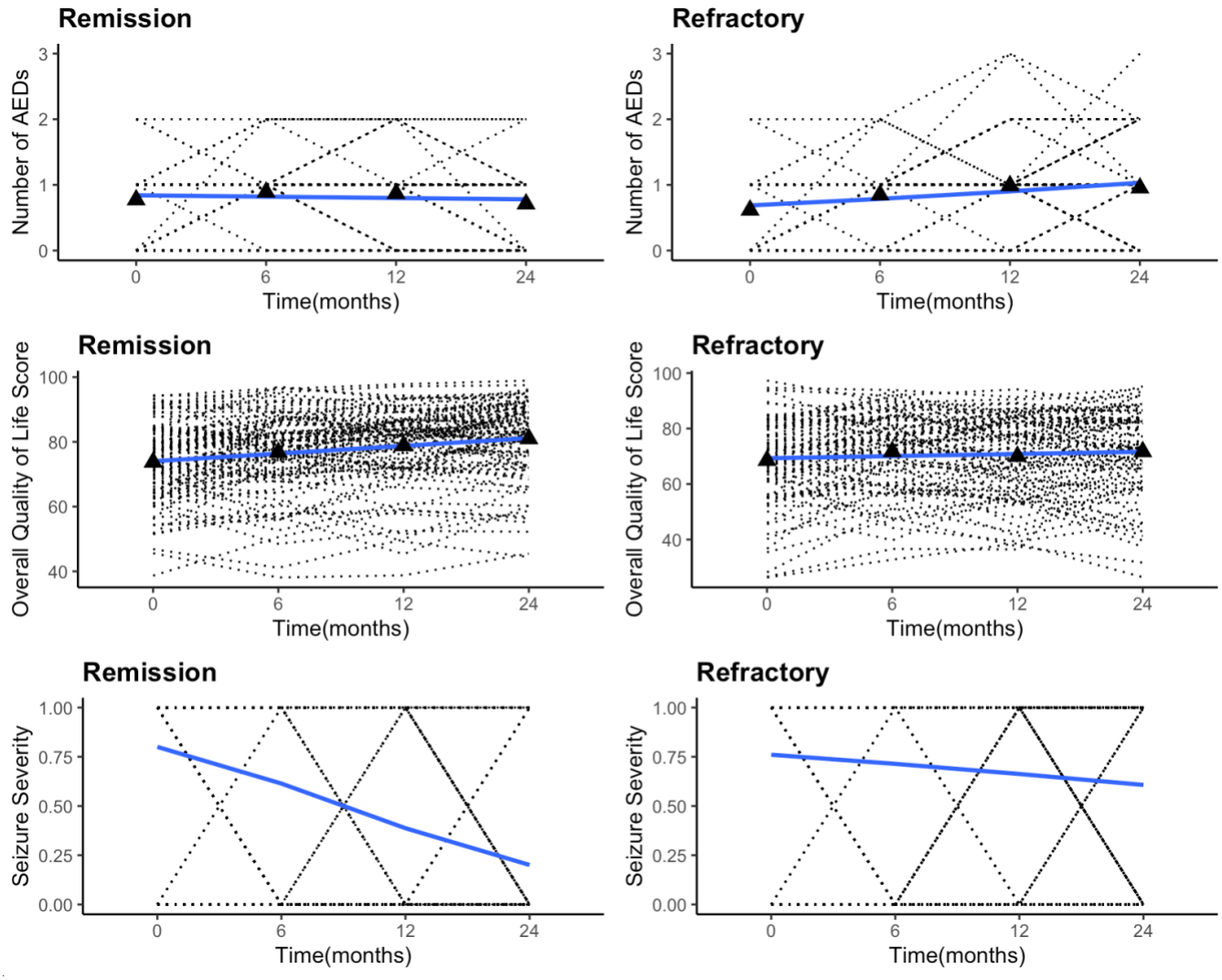

Figure 4.1: Observed longitudinal profiles of number of anti-epileptic drugs (AEDs), quality of life and seizure severity from the Remission group (left column) and the Refractory group (right column). Solid lines show LOESS smoothed profiles for Poisson, normal and binomial models calculated using data from all patients. Baseline (0 month), and 6 months, 12 months, and 24 months.

Chapter 5

Effects of Correlation Mis-specification in Generalized Estimating Equations Discriminant Function for Multivariate Repeated Measures Data: A simulation study

Brobbey A,, Lix LM., Nettel-Aguirre A., Tyler Williamson T., Samuel Wiebe, Sajobi T., Effects of correlation mis-specification in generalized estimating equations discriminant function for Multivariate Repeated Measures Data: A simulation study. *Communications in Statistics* (under review)

The simulation study design, implementation of methods, and manuscript preparation was done by AB. All co-authors provided supervisory support, reviewed the results and critically revised the manuscript for important intellectual content. AB assumes responsibility for the integrity of the manuscript. This manuscript in its entirety is included in Chapter 5.

Abstract

Repeated measures discriminant analyses are generally developed based on the assumption of parsimonious correlation structures in multivariate repeated measures data which are characterized by complex correlation structures such as within- and between-outcome variable correlations. The assumption of parsimony in discriminant analysis ensures that the increasing complexity of parameter estimation in multivariate repeated measures data can be handled. This study evaluates the impact of correlation structure mis-specification on the classification accuracy of discriminant analysis based on multivariate generalized estimation equations in multivariate normal and non-normal repeated measures data. A computer simulation indicated a clear impact of correlation structure on the performance of the classification rules under diverse simulation generation conditions, which included population distribution, number of outcomes, and repeated occasions. The classification accuracy of linear discriminant analysis and quadratic discriminant analysis of the multivariate GEE procedures decreased when the correlation structures were mis-specified in most cases. We observed higher impact of correlation mis-specification in multivariate repeated measures binary outcomes for parsimonious covariance estimation procedures than in multivariate count or continuous outcomes. In addition, the classification accuracy increased with increase in number of outcomes and repeated occasions even under correlation mis-specification.

Keywords*:* mis-specification, discriminant analysis, multivariate repeated measures data, generalized estimating equation, classification

5.1 Introduction

Multivariate repeated measures data in which multiple correlated outcomes are repeatedly measured over time occur often in environmental and ecological studies. Valid analysis of multivariate repeated measures data requires accurately modeling of the correlation structure, while failure to account for the complex correlation structures may lead to biased regression parameters[1, 2]. Modeling the unstructured correlation structure in multivariate repeated measures data is ideal option, might result in convergence problems especially when the sample size is smaller or equal to the dimension of the data[3]. The increasing complexity of multivariate repeated measures data analysis can be reduced by imposing additional restrictions on the correlation structure.

Repeated measures discriminant analysis procedures that assume Kronecker (separable) product structure on the correlation matrix have been used in covariance pattern models [4-8] , mixed-effects models and multivariate generalized estimating equations (GEE) models [9-11]. Although parsimonious correlation structure like Kronecker product structure models can result in efficient classification rules especially in small-sample data, Roy showed that repeated measures discriminant analysis based on covariance pattern models may result in lower classification accuracy when the correlation structure is incorrectly specified in multivariate normal outcome distributions[7, 12] , but their study was not extended to non-normal outcome distributions.

To enable the analysis of all types of outcomes and covariates simultaneously for discriminant analysis, Brobbey et al. developed repeated measures discriminant analysis based on multivariate GEE model[13]. The multivariate GEE's regression parameters and their variances are robust with respect to mis-specification of the correlation matrix of the outcomes provided the sample size is sufficiently large[14]. However, the consistency and asymptotic normality of the

regression estimates depend on the use of consistent correlation parameter estimates, which are not guaranteed to exist or to be feasible if the correlation structure is mis-specified[15, 16]. To our knowledge, there is no investigation of the impact of correlation structure mis-specification on the accuracy of repeated measures discriminant analysis based on multivariate GEE [13,14,17, 18]. Previous research of repeated measures discriminant analysis based on multivariate GEE model by Brobbey et al. assumed true parsimonious correlation matrices[13] and further research is needed to investigate the impact of the assumed structures on the accuracy of the models.

Therefore, this study aimed to investigate the effects of correlation mis-specification on the classification accuracy of repeated measures discriminant analysis based on multivariate GEE. Monte Carlo methods were used to compare the accuracy of repeated measures discriminant analysis based on multivariate GEE when the correlation structure is correctly and incorrectly specified under different types of simulation generation conditions

5.2 GEE Discriminant Analysis for Multivariate Repeated Measures Data

Suppose we have a random sample of n individuals. For each individual $i = 1, \ldots, n$, let $\mathbf{y}_i = (\mathbf{y}'_{i1}, \mathbf{y}'_{i2}, \ldots, \mathbf{y}'_{iq})'$ be a pq x 1 vector of q correlated outcomes that are each measured at p occasions, and $\mathbf{X}_i = \mathbf{X}_{i*} \otimes \mathbf{I}_q$ is a corresponding pq x Kq block diagonal covariate matrix, where $\mathbf{X}_{i*} = (\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots \mathbf{X}_{ik}, \ldots, \mathbf{X}_{iK})$ is a p x K matrix of covariates and $\mathbf{X}_{ik} = (\mathbf{X}_{i1k}, \ldots\ldots, \mathbf{X}_{ipk})$, $\mathbf{I}_q$ is an q x q identity matrix, and $\otimes$ is the Kronecker product sign. For the analysis of multivariate correlated data, the marginal mean vector $\boldsymbol{\mu}_i = (\boldsymbol{\mu}'_{i1}, \boldsymbol{\mu}'_{i2}, \ldots, \ldots, \boldsymbol{\mu}'_{iq})'$ is associated with K covariates through a generalized linear model (GLM) as follows:

$$\boldsymbol{\mu}_i = \mathbf{f}_l(\mathbf{X}_i\boldsymbol{\beta}), \tag{5.5}$$

where, $\mathbf{f}_l(\cdot)$, $l = 1,2, \dots, q$ is the inverse outcome-specific link function,

$\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \dots, \dots, \boldsymbol{\beta}_q')'$, where $\boldsymbol{\beta}_q = (\boldsymbol{\beta}_{q1}, \boldsymbol{\beta}_{q2}, \dots \boldsymbol{\beta}_{qk}, \dots , \boldsymbol{\beta}_{qK})'$ is the $pq \times 1$ dimensional vector of the qth outcome regression coefficients with population-averaged interpretations. The $pq \times pq$ marginal covariance matrix is:

$$\boldsymbol{\Omega}_i = \phi\boldsymbol{\Sigma}_i, \tag{5.6}$$

where ϕ is a scale parameter that can be known or estimated and $\boldsymbol{\Sigma}_i$ is an $pq \times pq$ working covariance matrix, which results in a total of pq (pq + 1)/2 unknown parameters to be estimated for any statistical inference [19, 20] which may not always be feasible especially when $pq \approx N$. Inferences of interest are easily influenced by the correlation structure's assumptions and unstructured correlation structure might cause convergence problems as the number of parameters to be estimated grows rapidly[3]. In the quasi-likelihood framework with repeated measures outcomes, the regression coefficients $\boldsymbol{\beta}$ can be estimated by solving the generalized estimating equations (GEEs)

$$\mathrm{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{D}_i' \boldsymbol{\Omega}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \tag{5.7}$$

where $\mathbf{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$ is the block diagonal matrix of derivatives mean with respect to the regression parameters , $\boldsymbol{\mu}_i$ is the marginal mean vector, and $\boldsymbol{\Omega}_i = \phi\mathbf{A}_i^{1/2}\mathbf{R}_{pq}\mathbf{A}_i^{1/2}$ is the working covariance matrix, where $\mathbf{A}_i$ is an $pq \times pq$ block diagonal matrix, which contains the marginal variance of outcomes on the main diagonals and $\mathbf{R}_{pq}$ is a $pq \times pq$ working correlation matrix. Under mild regularity conditions, the parameter estimates $\widehat{\boldsymbol{\beta}}$ are consistent and asymptotically normally distributed even when the "working" correlation structure of outcomes is misspecified, and the variance-covariance matrix can be estimated using a robust "sandwich" variance estimator[21].

Given $\mathbf{y}_{ij} = (\mathbf{y}'_{ij1}, \mathbf{y}'_{ij2}, \ldots, \mathbf{y}'_{ijq})'$ is a pq x 1 random vector corresponding to the *i*th individual in the *j*th population, estimations of the marginal mean $\boldsymbol{\mu}_j$, and covariance $\boldsymbol{\Omega}_j$ are obtained using a pre-defined structure from the multivariate GEE model in population j (j =1,2). The linear discriminant analysis (LDA) model is obtained when the variance components are homogeneous, that is, $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2 = \boldsymbol{\Omega}$, the pooled covariance matrix. The LDA implies that an individual with multiple outcome vector $\mathbf{y}_i$ is classified in the first population, if and only if

$$\left(\mathbf{y}_i - \frac{\widehat{\boldsymbol{\mu}}_1 + \widehat{\boldsymbol{\mu}}_2}{2}\right)' \widehat{\boldsymbol{\Omega}}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2) > \log \frac{\hat{\pi}_2}{\hat{\pi}_1} \tag{5.8}$$

, and for quadratic discriminant analysis (QDA) classification when $\boldsymbol{\Omega}_1 \neq \boldsymbol{\Omega}_2$, as

$$(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_2)'\widehat{\boldsymbol{\Omega}}_2^{-1}(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_2) - (\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_1)'\widehat{\boldsymbol{\Omega}}_1^{-1}(\mathbf{y}_i - \widehat{\boldsymbol{\mu}}_1) > \log \left|\frac{\widehat{\Omega}_1}{\widehat{\Omega}_2}\right| + 2\log \frac{\hat{\pi}_2}{\hat{\pi}_1}, \tag{5.9}$$

otherwise, it is classified into the second population, where $\widehat{\boldsymbol{\mu}}_j$ and $\widehat{\boldsymbol{\Omega}}_j^{-1}$ are the GEE estimates from (5.1) and (5.2), $\hat{\pi}_1$ and $\hat{\pi}_2$ are the *a priori* probabilities that observations belong to populations 1 and 2.

5.3 Simulation Study

A Monte Carlo simulation study was conducted to examine the impact of correlation misspecification on the classification accuracy of linear and quadratic GEE repeated measures discriminant analysis. The linear and quadratic GEE discriminant analysis procedures that assume Kronecker product structured working correlation with: (a) Unstructured between-outcomes correlation and within-outcome first-order autoregressive correlation (UNAR) (b) unstructured between-outcomes correlation and within-outcome compound symmetric correlation (UNCS) structures, and (c) an unstructured working correlation (UN). The following simulation generation conditions were investigated in the study: (a) number of repeated measurements (p), (b) number of outcomes (q), (c) total sample size (N), (d) group sizes (n_1, n_2), (e) pattern and magnitude of

correlation among the repeated measurements (ρ), (f) covariance heterogeneity, and (g) population distribution. All procedures were investigated for two independent groups. The number of repeated measurements points was set at $p = 3$ and 5, while the number of outcomes was set at $q = 3$ and 5. Previous studies about discriminant analysis procedures for multivariate repeated measures data have considered values of p ranging from three to ten, an increase in classification accuracies were quite significant when p increased from three to five. [6, 12] Total sample sizes of $N = 120$, 250 and 500 were investigated. Consistent with previous simulation studies that examined the accuracy of discriminant accuracy for multivariate repeated measures data between 60 and 500 and examined the impact of equal and unequal group sizes [6, 12,22, 23], we investigated conditions of $N = 120$, $(n_1, n_2) = (60, 60)$, and $(48, 72)$, which represent a group size ratio of 1:1 and 2:3, respectively. Similar equal and unequal group size ratios were investigated when $N = 250$ and $N = 500$.

Furthermore, the accuracy of discriminant analysis procedures is known to be influenced by both the magnitude and pattern of within- and multivariate-outcome correlations[24]. Therefore, we investigated the following within-outcome correlation structures for Kronecker product structure $\mathbf{R}_{pq} = \mathbf{R}_p(\rho) \otimes \mathbf{R}_q(\alpha)$: (a) Compound Symmetry with $\rho = 0.3$ and $\rho = 0.7$, (b) Autoregressive order 1 with $\rho = 0.3$ and $\rho = 0.7$[6, 12] for within-outcome correlation $\mathbf{R}_p(\rho)$**,** and the between-outcomes correlation, $\mathbf{R}_q(\alpha)$ was assumed to be unstructured (See Table 5.1) (c) unstructured correlation matrix from an independent uniform random variable on (0.2, 0.4) and (0.6, 0.8). For covariance heterogeneity, we assumed $\mathbf{\Omega}_1 = \mathbf{\Omega}_2$ and $\mathbf{\Omega}_1 = 3\mathbf{\Omega}_2$. All combinations of simulation generation conditions were investigated for each procedure and each method of estimation, resulting in a total of 144 combinations. There were 500 replications for each combination. Linear and quadratic discriminant analysis rules were developed using marginal

mean and variance-covariance matrix estimated via GEE for equal and unequal covariance matrix respectively. All analyses were completed in R statistical software version 3.5.3. In order to assess the performance of the discriminant function, we investigated multivariate correlated continuous and discrete outcome variables

For the correlated continuous outcome variables, we assumed three normal variables jointly observed for n_j subjects, where each observed at p time points. The true marginal mean outcome model $\boldsymbol{\mu}_{ipq}$ was assumed to take the following functional form that uses an identity link function:

$$\boldsymbol{\mu}_{ipq} = \boldsymbol{\beta}_{q1}x_{ip1} + \boldsymbol{\beta}_{q2}x_{ip2} + \boldsymbol{\beta}_{q3}t_{ip} \tag{5.6}$$

The number of covariates, $K = 3$, where x_{ip1} was generated from an independent normal random variable $N(0,1)$ and x_{ip2} from an independent binomial random variables $B(n_1, 0.6)$ and $B(n_2, 0.4)$ as time-invariant covariates for population 1 and population 2 respectively, and t_{ip} denoted the time of observation as a time-varying covariate. Details of the true parameters $\boldsymbol{\beta}$ for population 1 and population 2 can be found in Table 5.2. On the other hand, the marginal variance matrix of outcomes was assumed to have a common variance of 60. R package mvrnorm() function from the MASS R package[25] was used to generate the multivariate normal data.

For the multivariate binary and count (skewed) outcome variables, simulation data were generated from a multivariate binomial distribution using the logit link function and Poisson distribution using the log link function with the three covariates respectively.

$$\log\left(\frac{\boldsymbol{\mu}_{ipq}}{1-\boldsymbol{\mu}_{ipq}}\right) = \boldsymbol{\beta}_{q1}x_{ip1} + \boldsymbol{\beta}_{q2}x_{ip2} + \boldsymbol{\beta}_{q3}t_{ip} \tag{5.7}$$

$$\log\left(\boldsymbol{\mu}_{ipq}\right) = \boldsymbol{\beta}_{q1}x_{ip1} + \boldsymbol{\beta}_{q2}x_{ip2} + \boldsymbol{\beta}_{q3}t_{ip} \tag{5.8}$$

The true parameters $\boldsymbol{\beta}$ for population 1 and population 2 can be found in Table 5.2. R package "bindata" function[26] was used to generate the multivariate binary data and "PoisNor" function[27] was used to generate the multivariate count data. A total of 432 combination of simulation factors were investigated with 1000 replications for each combination. The Monte Carlo study was conducted using R version 3.6.3. The classification performance of the procedures was evaluated using the mean overall classification accuracy on a scale of 0 to 1 and its corresponding standard error were reported for each combination of simulation generation conditions.

5.4 Simulation Results

The mean classification accuracies and standard errors of linear and quadratic discriminant analysis based on multivariate GEE by number of repeated measurements and number of outcomes for multivariate normal outcomes were presented in Tables 5.3. The results of LDA and QDA showed that mis-specification of correlation structure resulted in decreased overall mean classification accuracy for all procedures when data were sampled from a multivariate normal distribution. For example, when the true correlation structure was UNCS, the mean accuracies for $q = 3$ and $p = 5$ were 0.64, 0.65 and 0.60 for UNAR, UNCS, and UN estimations respectively under LDA. But, QDA based on GEE procedures were more accurate than LDA based on GEE among all the conditions investigated. Under QDA, the mean accuracies were 0.77, 0.78 and 0.70 for UNAR, UNCS, and UN estimations respectively (Table 5.3). The mean classification accuracies under parsimonious correlation were higher than its corresponding unstructured covariance under true correlation structure. That is, misspecification of correlation affected mean classification accuracies under unstructured estimation more than parsimonious correlation estimation. For example, when the true correlation structure was UNCS, $q = 3$ and $p = 5$, the mean accuracies for the QDA based on multivariate GEE were 0.78, and 0.70 for UNCS and UN

correlation structures estimation, respectively. In contrast, when the true population correlation was UN, the mean accuracies for the QDA based on multivariate GEE were 0.70 and 0.73 when the UNCS and UN correlation structures were used for parameter estimation, respectively (Table 5.3). Furthermore, the mean classification accuracy of all the correlation estimation methods increased with increasing number of outcomes, regardless of the assumed true correlation structure for multivariate correlated normal outcomes. For example, for p =3, the mean increase in classification accuracy of the LDA and QDA procedure were about 0.08 and 0.07 respectively as q increased from 3 to 5, under UNAR true correlation structure and UNAR estimation (Table 5.3). Likewise, the mean increase in classification accuracy of the LDA and QDA procedure were about 0.07 and 0.04 respectively as q increased from 3 to 5, under UNAR true correlation structure and UN estimation (misspecification) and p =5 (Table 5.3).

Tables 5.4 describes the mean classification accuracies and standard errors by population distribution, number of repeated occasions and number of outcomes for multivariate correlated binary outcomes. The results of LDA and QDA based on GEE showed that misspecification of correlation structure did not affect overall mean classification accuracy under unstructured correlation estimation for multivariate binary outcomes regardless of the true correlation structure. For example, when the true correlation structure was UNAR, the mean accuracies for $q = 3$ and $p = 3$ were 0.63, 0.63 and 0.64 for UNAR, UNCS, and UN estimations respectively under LDA. Whiles under QDA, the mean accuracies for $q = 3$ and $p = 3$ were 0.68, 0.68 and 0.72 for UNAR, UNCS, and UN estimations respectively (Table 5.4). Similar to results presented in Table 5.3, QDA based on GEE procedures were more accurate than LDA based on GEE among all procedures. Furthermore, we observed little or no differences in classification accuracies for UNAR and UNCS correlation estimations regardless of number of outcomes and number of

repeated occasions under mis-specified and true correlation. Under multivariate correlated binary outcomes, the mean classification accuracy of all correlation estimation methods increased with increasing number of outcomes and repeated occasions, regardless of the assumed true correlation. For example, for $p = 3$, the mean increase in classification accuracy of the LDA procedure were about 0.03 and 0.07 as q increased from 3 to 5, under UNAR and UNCS estimation respectively for true UNAR correlation structure (Table 5.4). Whiles for $q = 3$, the mean increase in classification accuracy of the LDA procedure were about 0.03 as q increased from 3 to 5 for both UNAR and UNCS estimation respectively for true UNAR correlation structure (Table 5.4).

Tables 5.5 describes the mean classification accuracies and standard errors by population distribution, number of repeated occasions and number of outcomes for multivariate correlated Poisson distribution (count outcomes). Under multivariate correlated count outcomes, we observed a decrease in mean classification accuracies under misspecification of the true correlation structure especially when the number of outcomes was small ($q = 3$) regardless of estimation method. For example, when the true correlation structure was UN, the mean accuracies for $q = 3$ and $p = 3$ were 0.84, 0.83 and 0.87 for UNAR, UNCS, and UN estimations respectively under LDA. However, we observed similar mean classification accuracies for most cases even under misspecification of the true correlation structure when the number of outcomes increased ($q = 5$) for persimmons correlation estimation but not mis-specification under unstructured correlation estimation. For example, when the true correlation structure was UNAR, the mean accuracies for $q = 5$ and $p = 5$ were 0.99, 0.99 and 0.90 for UNAR, UNCS, and UN estimations respectively under LDA. Whiles under QDA, the mean accuracies for $q = 5$ and $p = 5$ were 0.99, 0.99 and 0.89 for UNAR, UNCS, and UN estimations respectively (Table 5.5). Similarly to outcomes generated from other population distributions, the mean classification accuracy of each LDA and

QDA procedures increased as the number of outcomes and repeated occasions increased, regardless of the estimation method for multivariate correlated count outcomes.

Overall, the impact of mis-specification of correlation on classification accuracy of discriminant analysis based on multivariate GEE depended on the population distributions. Mis-specification in multivariate non-normal outcomes under unstructured correlation estimation was accurate than parsimonious correlation estimation, or less decrease in classification accuracy for mis-specified parsimonious correlation in most scenarios. For example, when the true correlation structure was UNAR, the mean accuracies for $q = 3$ and $p = 5$ were 0.69, 0.69 and 0.81 for UNAR, UNCS, and UN estimations respectively under QDA (Tables 5.4). Whiles under count outcomes, the mean accuracies for $q = 3$ and $p = 5$ were 0.90, 0.90 and 0.87 for UNAR, UNCS, and UN estimations respectively (Table 5.5). However, we observed the opposite in multivariate normal outcomes, parsimonious correlation estimation was accurate in dealing with mis-specified correlation than unstructured correlation estimation.

5.5 Discussion

This study investigated the effect of correlation structure mis-specification on classification accuracy of LDA and QDA based on multivariate GEE when the data were sampled from multivariate normal and non-normal distributions. First, the results showed that mis-specification of correlation structure resulted in decreased overall mean classification accuracy for all procedures. The decrease was higher in multivariate non-normal distribution than normal distribution. Mis-specification of correlation structure severely affected multivariate binary outcomes followed by multivariate count and continuous outcomes in the repeated measures discriminant analysis based on multivariate GEE.

Second, the impact of mis-specification varied depending on the true and adopted correlation structure. More specifically, mis-specification of correlation structure had negligible impact on the mean classification accuracy under unstructured correlation estimation regardless of the true correlation structure in multivariate non-normal data but decreased mean classification accuracy was observed under parsimonious correlation estimation. However, parsimonious correlation estimation had less effect on the mean classification accuracy in multivariate normal data regardless of the true correlation structure compared to unstructured correlation estimation.

Based on the study findings, adopting a discriminant analysis procedure based on unstructured correlation matrices when the researcher has prior knowledge of the multivariate repeated measures correlation form for each group is recommended in general for correlated non-normal data. Specifically, either UNAR or UNCS parsimonious correlation are recommended for correlated count and continuous data as these parsimonious correlation structure result in efficient classification rules when data have less deviations from normal distributions and in small-sample data[7, 12].

Whilst it is well known that GEEs yield asymptotically consistent parameter and variance estimates under incorrect specification of the correlation structure [28-30], the findings of this study suggest that mis-specification of the true correlation structures for these discriminant analysis procedures could influence accuracy of these classifiers. To mitigate this problem, goodness of fit tests such as quasi-likelihood under the independence model criterion (QIC) which is a modified Akaike information criterion (AIC) for GEE model[31, 32] could be used to guide the choice of a working correlation matrix. Like the AIC, the QIC is a trade-off between a good fit to the model, as measured by the quasi-likelihood, and a penalty for over-complexity as measured by the trace. The use of independence assumption when computing the quasi-likelihood makes QIC easy to be

implemented but can lead to a considerable loss of efficiency in estimating the regression parameters[33] and not effective if the correlation structure of the data is far from independence[34].

Besides the strength of this study, there are some limitations. First, the assumption of complete multivariate repeated measures data in which there is no missing data on all outcomes and at all measurement occasions might not be satisfied in multivariate repeated measures data often encountered in applied research[18]. The GEE is valid under the strong assumption of missing completely at random (MCAR)[35] but not missing at random assumption (MAR). Some studies have been done to drop-out problems in repeated measures studies via weighted generalized estimating equations[36] and multiple imputations to ensure validity of the inference under MAR[35, 37]. Further research will be conducted of these misspecification procedures in discriminant analysis based on GEE by implementing some of the multiple imputation techniques[18, 38-40] using simulation techniques. Alternatively, mixed-effects models are efficient for dealing with incomplete data[41-44] and valid inferences can be obtained even with incomplete information under MAR. However, one disadvantage of mixed-effects models is that the dimension of random effects quickly increases as more outcomes and random effects are added to the model, increasing the computational burden[41, 45, 46].

The repeated measures discriminant analysis based on multivariate GEE relies on correctly specified link function. It is crucial in the GEE approach to select a correct link function linking the mean outcome to the covariates[47] as the consistency of GEE parameter estimates depends on correctly specified link function [28-30].

Our study conclusions might not be generalizable to very large number of outcomes, repeated measures, certain population distribution, correlation structures and mixed-type outcomes. Previous studies about discriminant analysis procedures for multivariate repeated

measures data observed an increase in classification accuracies were quite significant when number of repeated measures increased from three to five. [6, 12] but not six to ten. Therefore, future research will examine the impact of correlation structures on the robustness of these models in a large number of outcomes and to different simulation generation conditions using Monte Carlo methods to increase generalizability of our findings.

In summary, this study investigated the effect of misspecification of correlation structure of discriminant analysis procedures based on multivariate GEE in normal and non-normal multivariate repeated measures outcome variables for distinguishing between population groups. The adoption of discriminant analysis procedure based on a parsimonious correlation structure can reduce the number of parameters to estimate and provide efficient classification accuracy when sample size is small and in data with less deviations from normal distributions[7]. However, unstructured correlation structure is recommended if the researcher has prior knowledge of the correlation form for multivariate repeated measures data.

Acknowledgments

This research was supported by a Natural Sciences & Engineering Research Council Discovery Grant to TTS.

References

1. Shults J and Morrow AL. Use of quasi–least squares to adjust for two levels of correlation. *Biometrics* 2002; 58: 521-530.
2. Gurka MJ, Edwards LJ and Muller KE. Avoiding bias in mixed model inference for fixed effects. *Statistics in medicine* 2011; 30: 2696-2707.
3. Cho H. The analysis of multivariate longitudinal data using multivariate marginal models. *Journal of Multivariate Analysis* 2016; 143: 481-491.
4. Krzyśko M and Skorzybut M. Discriminant analysis of multivariate repeated measures data with a kronecker product structured covariance matrices. *Statistical papers* 2009; 50: 817-835.
5. Naik DN and Rao SS. Analysis of multivariate repeated measures data with a kronecker product structured covariance matrix. *Journal of Applied Statistics* 2001; 28: 91-105.
6. Roy A and Khattree R. On discrimination and classification with multivariate repeated measures data. *Journal of Statistical Planning and Inference* 2005; 134: 462-485.
7. Roy A and Khattree R. Classification of multivariate repeated measures data with temporal autocorrelation. *J Appl Stat Sci* 2007; 15: 283-294.
8. Roy A and Leiva R. Discrimination with jointly equicorrelated multi-level multivariate data. *Advances in Data Analysis and Classification* 2007; 1: 175-199.
9. Chaganty NR. An alternative approach to the analysis of longitudinal data via generalized estimating equations. *Journal of Statistical Planning and Inference* 1997; 63: 39-54.
10. Chaganty NR and Naik DN. Analysis of multivariate longitudinal data using quasi-least squares. *Journal of Statistical Planning and Inference* 2002; 103: 421-436.
11. Shults J and Chaganty NR. Analysis of serially correlated data using quasi-least squares. *Biometrics* 1998: 1622-1630.
12. Roy A and Khattree R. Discrimination and classification with repeated measures data under different covariance structures. *Communications in Statistics—Simulation and Computation®* 2005; 34: 167-178.
13. Brobbey A, Wiebe S, Aguirre AN, et al. Repeated measures discriminant analysis using multivariate generalized estimation equations. *University of Calgary* 2020. Unpublished doctoral dissertation.
14. Liang K-Y and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; 73: 13-22.
15. Crowder M. On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* 1995; 82: 407-410.
16. Crowder M. On repeated measures analysis with misspecified covariance structure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2001; 63: 55-62.
17. Lipsitz SR, Fitzmaurice GM, Ibrahim JG, et al. Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: An application to acquired immune deficiency syndrome data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009; 172: 3-20.
18. Inan G and Yucel R. Joint gees for multivariate correlated data with incomplete binary outcomes. *Journal of Applied Statistics* 2017; 44: 1920-1937.
19. Roy A and Khattree R. On implementation of a test for kronecker product covariance structure for multivariate repeated measures data. *Statistical Methodology* 2005; 2: 297-306.

20. Srivastava MS, von Rosen T and Von Rosen D. Models with a kronecker product covariance structure: Estimation and testing. *Mathematical Methods of Statistics* 2008; 17: 357-370.
21. Chao EC. Generalized estimating equations. Taylor & Francis, 2003.
22. Barön AE. Misclassification among methods used for multiple group discrimination-the effects of distributional properties. *Statistics in medicine* 1991; 10: 757-766.
23. He X and Fung WK. High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis* 2000; 72: 151-162.
24. Thomas DR and Zumbo BD. Using a measure of variable importance to investigate the standardization of discriminant coefficients. *Journal of Educational and Behavioral Statistics* 1996; 21: 110-130.
25. Ripley B. Ebooks corporation. Stochastic simulation. Wiley Online Library, 1987.
26. Leisch F, Weingessel A and Hornik K. Bindata: Generation of artificial binary data. *R package version 09-12* 2005.
27. Amatya A and Demirtas H. Poisnor: An r package for generation of multivariate data with poisson and normal marginals. *Communications in Statistics-Simulation and Computation* 2017; 46: 2241-2253.
28. Liang K-Y and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986: 13-22.
29. Qu A, Lindsay BG and Li B. Improving generalised estimating equations using quadratic inference functions. *Biometrika* 2000; 87: 823-836.
30. Asar Ö and İlk Ö. Mmm: An r package for analyzing multivariate longitudinal data with multivariate marginal models. *Computer Methods and Programs in Biomedicine* 2013; 112: 649-654.
31. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics* 2001; 57: 120-125.
32. Akaike H. Information theory and an extension of the maximum likelihood principle. *Selected papers of hirotugu akaike*. Springer, 1998, pp.199-213.
33. Fitzmaurice GM. A caveat concerning independence estimating equations with multivariate binary data. *Biometrics* 1995: 309-317.
34. Barnett AG, Koper N, Dobson AJ, et al. Using information criteria to select the correct variance–covariance structure for longitudinal data in ecology. *Methods in Ecology and Evolution* 2010; 1: 15-24.
35. Robins JM, Rotnitzky A and Zhao LP. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association* 1995; 90: 106-121.
36. Beunckens C, Sotto C and Molenberghs G. A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational statistics & data analysis* 2008; 52: 1533-1548.
37. Little RJ and Rubin DB. *Statistical analysis with missing data*. John Wiley & Sons, 2019.
38. Satty A, Mwambi H and Molenberghs G. Different methods for handling incomplete longitudinal binary outcome due to missing at random dropout. *Statistical Methodology* 2015; 24: 12-27.
39. Yucel RM, He Y and Zaslavsky AM. Using calibration to improve rounding in imputation. *The American Statistician* 2008; 62: 125-129.

40. Yucel RM, He Y and Zaslavsky AM. Imputation of categorical variables using gaussian-based routines. *Statistics in Medicine* 2011; 30: 3447-3460.
41. Fieuws S, Verbeke G, Maes B, et al. Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics* 2007; 9: 419-431.
42. Fieuws S, Verbeke G and Molenberghs G. Random-effects models for multivariate repeated measures. *Statistical methods in medical research* 2007; 16: 387-397.
43. Hughes DM, Komárek A, Bonnett LJ, et al. Dynamic classification using credible intervals in longitudinal discriminant analysis. *Statistics in medicine* 2017; 36: 3858-3874.
44. Hughes DM, Komárek A, Czanner G, et al. Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types. *Statistical methods in medical research* 2018; 27: 2060-2080.
45. Verbeke G, Fieuws S, Molenberghs G, et al. The analysis of multivariate longitudinal data: A review. *Statistical methods in medical research* 2014; 23: 42-59.
46. Komárek A, Hansen BE, Kuiper EM, et al. Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statistics in Medicine* 2010; 29: 3267-3283.
47. Molefe AC and Hosmane B. Test for link misspecification in dependent binary regression using generalized estimating equations. *Journal of Statistical Computation and Simulation* 2007; 77: 95-107.

Table 5.1: Configuration of unstructured between-outcomes correlation matrix given within-outcome correlation coefficient for the Monte Carlo Study

within-outcome correlation coefficient (ρ)	0.3	0.7
$q = 3$	$\begin{bmatrix} 1 & 0.15 & 0.30 \\ 0.15 & 1 & 0.45 \\ 0.30 & 0.45 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.65 & 0.66 \\ 0.65 & 1 & 0.70 \\ 0.66 & 0.70 & 1 \end{bmatrix}$
$q = 5$	$\begin{bmatrix} 1 & 0.28 & 0.25 & 0.28 & 0.28 \\ 0.28 & 1 & 0.30 & 0.40 & 0.23 \\ 0.25 & 0.30 & 1 & 0.24 & 0.24 \\ 0.28 & 0.40 & 0.24 & 1 & 0.37 \\ 0.28 & 0.23 & 0.24 & 0.37 & 1 \end{bmatrix}$	$\begin{bmatrix} 1 & 0.70 & 0.79 & 0.64 & 0.70 \\ 0.70 & 1 & 0.73 & 0.65 & 0.74 \\ 0.79 & 0.73 & 1 & 0.63 & 0.62 \\ 0.64 & 0.65 & 0.63 & 1 & 0.62 \\ 0.70 & 0.74 & 0.62 & 0.62 & 1 \end{bmatrix}$

q=Number of outcomes

Table 5.2: True parameters ($\boldsymbol{\beta}$) for population 1 and population 2 simulated data

Population Distribution	Number of outcomes	population 1	population 2
Normal	3	(0.3,1,2,0.1,1,1.5,1,1.5,2)	(0.6,2,4,0.2,2,3,2,3,4)
	5	(0.2,1,2,1.5,1,0.4,0.7,3,1.2,0.8,0.1,0.2,0.7, 0.9,1.2)	(0.4,2,4,3,2,0.8,1.4,6,2.4,1.6,0.2,0.4,1.4,1 . 8,2.4)
Binomial	3	(0.15,0.1,0.2,0.1,0.1,0.25,0.1,0.2,0.3)	(0.37,0.25,0.5,0.25,0.25,0.62,0.25,0.5,0.75)
	5	(0.15,0.1,0.2,0.1,0.1,0.25,0.1,0.2,0.3,0.25,0.1, 0.2,0.2,0.14,0.3)	(0.37,0.25,0.5,0.25,0.25,0.62,0.25,0.5,0.75, 0.625,0.25,0.5,0.5,0.35,0.75)
Poisson	3	(0.3,0.1,0.2,0.1,0.3,0.5)	(0.09, 0.03, 0.06, 0.03 ,0.09, 1.5)
	5	(0.3,0.1,0.4,0.1,0.45,0.6,0.2,0.15,0.3,0.4)	(0.09, 0.03,0.12,0.03,0.13,0.18,0.06,0.04, 0.09,0.12)

Table 5.3: Overall Mean Accuracy (standard error) for repeated measures LDA and QDA procedures based on GEE by number of outcomes, number of repeated occasions and correlation structure for multivariate correlated normal outcomes

Number of outcomes (*q*)	Number of repeated occasions (*p*)		LDA			QDA		
		WR→	UNAR	UNCS	UN	UNAR	UNCS	UN
		TR ↓						
3	3	UNAR	0.64(0.03)	0.65(0.03)	0.63(0.04)	0.76(0.03)	0.76(0.03)	0.71(0.05)
	3	UNCS	0.65(0.03)	0.66(0.03)	0.64(0.04)	0.76(0.03)	0.77(0.03)	0.72(0.05)
	3	UN	0.64(0.03)	0.65(0.03)	0.65(0.03)	0.70(0.03)	0.71(0.03)	0.70(0.05)
	5	UNAR	0.64(0.03)	0.64(0.03)	0.61(0.03)	0.78(0.04)	0.78(0.04)	0.71(0.06)
	5	UNCS	0.64(0.03)	0.65(0.03)	0.60(0.05)	0.77(0.04)	0.78(0.04)	0.70(0.06)
	5	UN	0.63(0.03)	0.63(0.03)	0.65(0.03)	0.70(0.03)	0.70(0.03)	0.73(0.05)
5	3	UNAR	0.72(0.03)	0.73(0.03)	0.70(0.05)	0.83(0.03)	0.83(0.03)	0.75(0.06)
	3	UNCS	0.71(0.03)	0.77(0.03)	0.72(0.06)	0.83(0.03)	0.87(0.02)	0.77(0.07)
	3	UN	0.71(0.03)	0.75(0.03)	0.77(0.04)	0.75(0.03)	0.77(0.03)	0.77(0.06)
	5	UNAR	0.69(0.03)	0.70(0.03)	0.64(0.06)	0.84(0.04)	0.83(0.04)	0.72(0.07)
	5	UNCS	0.69(0.03)	0.74(0.03)	0.63(0.08)	0.82(0.04)	0.87(0.04)	0.72(0.08)
	5	UN	0.69(0.03)	0.70(0.03)	0.72(0.06)	0.71(0.03)	0.72(0.03)	0.73(0.06)

Note:TR=true correlation structure; WR=working correlation structure; LDA= Linear Discriminant Analysis; QDA= Quadratic Discriminant Analysis ; *p*= number of repeated occasions; *q* = number of outcomes; UNAR = Unstructured between-outcomes and Autoregressive order 1 within-outcome correlation matrix; UNCS = Unstructured between-outcomes and Compound symmetry within-outcome correlation matrix; UN = unstructured correlation

Table 5.4: Overall Mean Accuracy (standard error) for repeated measures LDA and QDA procedures based on GEE by number of outcomes, number of repeated occasions and correlation structure for multivariate correlated binary outcomes

Number of outcomes(q)	Number of repeated occasions (p)		LDA			QDA		
		WR→	UNAR	UNCS	UN	UNAR	UNCS	UN
		TR ↓						
3	3	UNAR	0.63(0.03)	0.63(0.04)	0.64(0.06)	0.68(0.03)	0.68(0.03)	0.72(0.03)
	3	UNCS	0.63(0.04)	0.63(0.04)	0.64(0.05)	0.68(0.03)	0.68(0.03)	0.72(0.03)
	3	UN	0.62(0.03)	0.62(0.04)	0.64(0.07)	0.66(0.03)	0.66(0.03)	0.71(0.03)
	5	UNAR	0.66(0.03)	0.66(0.03)	0.68(0.06)	0.69(0.03)	0.69(0.03)	0.81(0.03)
	5	UNCS	0.65(0.03)	0.66(0.03)	0.69(0.06)	0.69(0.03)	0.69(0.03)	0.81(0.03)
	5	UN	0.64(0.03)	0.64(0.04)	0.69(0.08)	0.68(0.03)	0.68(0.03)	0.78(0.03)
5	3	UNAR	0.69(0.03)	0.70(0.03)	0.77(0.10)	0.70(0.03)	0.70(0.03)	0.78(0.05)
	3	UNCS	0.68(0.03)	0.69(0.03)	0.77(0.10)	0.68(0.03)	0.68(0.03)	0.77(0.05)
	3	UN	0.66(0.03)	0.67(0.04)	0.76(0.11)	0.67(0.03)	0.68(0.03)	0.75(0.06)
	5	UNAR	0.72(0.03)	0.73(0.04)	0.81(0.07)	0.75(0.03)	0.76(0.03)	0.87(0.04)
	5	UNCS	0.70(0.03)	0.71(0.04)	0.79(0.09)	0.74(0.03)	0.76(0.03)	0.87(0.04)
	5	UN	0.67(0.03)	0.68(0.03)	0.82(0.10)	0.70(0.03)	0.71(0.03)	0.85(0.05)

Note:TR=true correlation structure; WR=working correlation structure; LDA= Linear Discriminant Analysis; QDA= Quadratic Discriminant Analysis ; p= number of repeated occasions; q = number of outcomes; UNAR = Unstructured between-outcomes and Autoregressive order 1 within-outcome correlation matrix; UNCS = Unstructured between-outcomes and Compound symmetry within-outcome correlation matrix; UN = unstructured correlation

Table 5.5: Overall Mean Accuracy (standard error) for repeated measures LDA and QDA procedures based on GEE by number of outcomes, number of repeated occasions and correlation structure for multivariate correlated Poisson outcomes

Number of outcomes (*q*)	Number of repeated occasions (*p*)		LDA			QDA		
		WR→	UNAR	UNCS	UN	UNAR	UNCS	UN
		TR ↓						
3	3	UNAR	0.83(0.02)	0.84(0.03)	0.81(0.03)	0.86(0.04)	0.87(0.04)	0.80(0.06)
	3	UNCS	0.86(0.04)	0.86(0.04)	0.83(0.03)	0.87(0.04)	0.87(0.05)	0.82(0.04)
	3	UN	0.84(0.02)	0.83(0.01)	0.87(0.06)	0.81(0.01)	0.82(0.01)	0.86(0.05)
	5	UNAR	0.91(0.01)	0.91(0.01)	0.93(0.03)	0.90(0.01)	0.90(0.01)	0.87(0.08)
	5	UNCS	0.90(0.01)	0.90(0.01)	0.92(0.05)	0.89(0.01)	0.89(0.01)	0.88(0.08)
	5	UN	0.90(0.02)	0.90(0.01)	0.92(0.09)	0.87(0.02)	0.87(0.03)	0.91(0.06)
5	3	UNAR	0.97(0.01)	0.98(0.01)	0.91(0.10)	0.98(0.01)	0.99(0.01)	0.92(0.08)
	3	UNCS	0.97(0.01)	0.98(0.01)	0.90(0.11)	0.98(0.01)	0.99(0.01)	0.91(0.09)
	3	UN	0.98(0.01)	0.98(0.01)	0.99(0.01)	0.99(0.01)	0.99(0.01)	0.98(0.05)
	5	UNAR	0.99(0.01)	0.99(0.01)	0.90(0.10)	0.99(0.01)	0.99(0.01)	0.89(0.13)
	5	UNCS	0.99(0.01)	0.99(0.01)	0.90(0.01)	0.99(0.01)	0.99(0.01)	0.90(0.04)
	5	UN	0.99(0.01)	0.99(0.01)	0.99(0.01)	0.99(0.01)	0.99(0.01)	0.98(0.02)

Note:TR=true correlation structure; WR=working correlation structure; LDA= Linear Discriminant Analysis; QDA= Quadratic Discriminant Analysis; *p*= number of repeated occasions; *q*= number of outcomes; UNAR = Unstructured between-outcomes and Autoregressive order 1 within-outcome correlation matrix; UNCS = Unstructured between-outcomes and Compound symmetry within-outcome correlation matrix; UN = unstructured correlation

Chapter 6

Discussion and Conclusions

6.1 Summary of Study Findings

This dissertation studied and developed discriminant analysis procedures based on parsimonious covariance or correlation structures in multivariate non-normal repeated measures data to discriminate between two populations. This research investigated three major research questions. (1) How accurate are existing repeated measures discriminant analysis classifiers when applied to discriminate between study samples of multivariate non-normal repeated measures distributions? (2) Can we develop more accurate classification models that overcome the restriction of multivariate normality for classification in multivariate non-normal distributions, in comparison to the conventional discriminant analysis models based on MLE? (3) What is the impact of mis-specification of correlation structures on the accuracy of repeated measures discriminant analysis based on GEE when used for classification in multivariate repeated measures data? How is the impact of mis-specification influenced by outcome variable distributions?

The key findings of this research revealed that the mean classification accuracy of repeated measures discriminant analysis procedures proposed in this study is influenced by a number of data characteristics including population distribution, mean configuration, number of outcome variables, number of repeated occasions, and the correlation among the multivariate repeated measures data. Also, the impact of correlation misspecification on the mean classification accuracy is largely influenced by population distribution and therefore preliminary examination of the appropriateness of parsimonious correlation structures in repeated measures discriminant analysis (such as LRT and QIC) is always recommended.

In chapter 3, we developed repeated measures discriminant analysis models based on maximum trimmed likelihood estimators for different assumptions of parsimonious covariance structures. Our simulation study showed that repeated measures discriminant analysis procedures based on MTLE were more accurate than the conventional repeated measures discriminant analysis based on MLEs when data were sampled from multivariate heavy-tailed distributions but not multivariate skewed distributions[1, 2]. The MTLE approach adopted for the repeated measures discriminant analysis procedures has good theoretical properties (affine equivariant estimators with bounded influence function properties and high breakdown points) that have been demonstrated in previous research for multivariate data[3]. These models were found to be more accurate when there are outlying observations. However, our discriminant analysis procedures based on MTLE could not be used in multivariate repeated measures data with binary outcomes because of the underlying assumptions.

In Chapter 4, we developed repeated measures discriminant analysis models based on multivariate generalized estimating equations and examined its accuracy in comparison to repeated measures discriminant analysis based on MLEs under Kronecker product structured correlations in multivariate repeated measures data with discrete and/or mixed type of outcomes[4, 5]. The significant advantages of discriminant analysis based on multivariate GEE include its computational simplicity and its flexibility for classification of multivariate repeated measures data with different types of outcomes (continuous, categorical, or ordinal), [5-8]. This class of discriminant analysis based on multivariate GEE showed better classification accuracy than discriminant analysis based on MLEs especially in multivariate repeated measures binary, count, and/or skewed data. Specifically, quadratic discriminant analysis based on multivariate GEE was more accurate than the linear discriminant analysis based on multivariate GEE considering

population distribution and covariance heterogeneity conditions. However, the accuracy of these discriminant analysis based on multivariate GEE depends on the correct specification of the multivariate link function for each outcome.

In Chapter 5, we examined the impact of mis-specification of correlation structures on the accuracy of the discriminant analysis models based on multivariate GEE using unstructured and parsimonious Kronecker correlation structures. The findings of this study using Monte Carlo methods reveal that mis-specification of the true multivariate repeated measures correlation structure for these classification models result in decreased mean classification accuracy. The decrease in accuracy also varied depending on the correlation structure adopted for estimation and the population distribution of the outcome variables. In addition, increase in classification accuracy for increased numbers of outcomes and repeated occasions was not influenced by correlation mis-specification.

6.2 Implications of Study Findings

This study contributes to the statistical literature on methods for analyzing multivariate non-normal repeated measures data and classifying individuals in populations. The major contribution of the present research is the development of much needed repeated measures discriminant analysis procedures that can be used for developing classification or prediction models for multivariate non-normal repeated measures data.

The findings from our simulation studies showed that the mean classification accuracy of repeated measures discriminant analysis procedures proposed in this study were found to be influenced by a number of data characteristics including population distribution, mean configuration, number of outcome variables, number of repeated occasions, and the correlation among the multivariate repeated measures data.

More specifically, the research revealed that repeated measures discriminant analysis based on MLE should not be used when the data violates the assumptions of multivariate normality. Instead, the procedures developed in this study should be used. Tests of multivariate normality can be explored by looking at graphs (such as box plots, QQ plots, multi-dimensional graphs). Alternatively, the Mardia's test can be used to check whether the multivariate skewness and kurtosis in the multivariate repeated measures are consistent with a multivariate normal distribution[9].

The observed impact of mis-specification of correlation structure on the accuracy of these developed models call for preliminary examination of the appropriateness of parsimonious correlation structures before choosing among the repeated measures discriminant analysis procedures proposed in this study[10, 11,12]. Most authors have used likelihood ratio test (LRT) statistic for testing separability of a covariance structure in multivariate repeated measures data[13] to avoid invalid inferences. However, the LRT statistic is reliable with very large samples, which may be limited in the real-life applications because we have only finite samples. Rao's score test (RST) has been proposed as an alternative to LRT approach for small sample data[14, 15]. In general, goodness of fit test such as LRT should be used to determine the appropriateness of correlation structure in the data before deciding the choice of repeated measures discriminant analysis.

Furthermore, our study revealed the positive impact of increasing both repeated occasions and number of outcomes on the classification accuracy of the proposed repeated measures discriminant analysis procedures even under mis-specified correlation[16, 17]. However, the increasing of the number of repeated occasions and number of outcomes improves classification accuracy provided the *n/pq* is satisfied for the estimation of covariance matrix even with the assumption of parsimony. This is often violated when *p* is large. In addition, if the within-variable

correlation increases as *pq* increases, the repeated measurements becomes less and less informative. Hence, the classification accuracy might not necessarily increase because more repeated measurements are added.

The choice between proposed repeated measures discriminant analysis procedures in this research (MTLE and GEE) should also be determined by the type of outcomes and sample size. For example, discriminant analysis models based on MTLE might be more useful for continuous outcomes with heavy-tailed distributions. However, discriminant analysis models based on GEE should be used in multivariate repeated measures data with discrete and/or mixed type outcomes (e.g. binary and count data). In addition, the appropriate link function for the multivariate GEE model needs to be explored and correctly determined for the outcome distributions before using repeated measures discriminant analysis based on GEE. Even though, any suitable link function can be used to relate the mean response to the covariates, the choice of a canonical link function produces many of the most widely used regression models[18].

These proposed procedures have a number of uses in clinical and population settings where multiple outcomes are repeatedly collected to inform clinical decisions such as diagnosis or treatment decisions. For example, in chapter 4, we demonstrated the potential use of these models for predicting children with new onset epilepsy who are likely to have treatment resistant epilepsy based on repeated measures data collected on some clinical outcomes over one-year period. In addition, discriminant analysis have been used for repeated measures data in dementia and other neuromuscular diseases where repeated measurement of severe clinical data and biomarkers are needed to arrive at a diagnosis[19-21].

There are few or no formal software packages developed to implement these methods in clinical settings. Therefore, the development of these new repeated measures discriminant analysis

methods in multivariate repeated measures data calls for the development of open-source packages to promote its use and implementation in applied research settings.

6.3 Strengths and Limitations

This dissertation study have several areas of strength of this research. First, the proposed repeated measures discriminant analysis procedures for discriminating between populations in multivariate repeated measures data are flexible models for modeling different types of outcomes, and are advantageous for discrimination in small-sampled multivariate repeated measures data. To our knowledge, the repeated measures discriminant analysis procedures developed in this study have not been previously studied for developing classification models for multivariate non-normal repeated measures data. This is an important contribution to the statistical literature on methods for classification in multivariate non-normal repeated measures data.

These proposed repeated measures discriminant analysis procedures account for the complex correlation structures that are inherent in multivariate repeated measures models to improve the accuracy of the classifiers[22]. In addition, the procedures developed in this research are based on parsimonious covariance structures for discriminating between populations in multivariate repeated measures data, which is beneficial to studies with small sample sizes. Also, quadratic discriminant analysis is recommended over linear discriminant analysis if there is evidence of covariance heterogeneity among population groups, to help minimize misclassification.

Most existing discriminant analysis methodologies in multivariate repeated measures data are based on mixed effects models and covariance pattern models. Therefore, our repeated measures discriminant analysis based of multivariate GEE offers an alternative flexible algorithm that can be used to simultaneously analyse different types of outcomes (continuous, counts and binary) for researchers in the area of multivariate repeated measures classification studies. In addition, the proposed GEE discriminant analysis approach allows the incorporation of covariates

to improve the accuracy of the classifier. Our simulation results for the repeated measures discriminant analysis based on multivariate GEE model outperforms the conventional repeated measures discriminant analysis based on MLEs [4, 5]. GEE packages and procedures are available in common statistical software such as R and SAS, thus these procedures can easily be applied in clinical research[23-26].

Another strength of this study is that, two population-based longitudinal registries were used to demonstrate the application of these models. These demonstrations encourage the use and show the flexibility of proposed models for different types of outcomes in applied settings.

Despite the unique strengths of this research, the limitations of this study should also be noted. First, the models developed in this study rely on the assumption that the group covariances have parsimonious covariance structures. Even though several authors have observed many advantages of using Kronecker product structure in addressing sample size and computational issues in multivariate repeated measures data[16, 17, 27-29], this assumption may not always be satisfied in typical datasets obtained from in medical studies. The LRT, QIC and CIC have been recommended to guide the selection of a well-fitted model with an appropriate covariance structure

Second, the investigated repeated measures discriminant analysis procedures assumed complete data on all observations and across repeated measurements which might not be realistic in multivariate repeated measures data often encountered in applied research. Deletion of data may result in biased estimates of discriminant function coefficients and loss of statistical power due to smaller sample size. Alternative approaches of models for classification include extension of mixed-effects models to these robust trimmed methods as they are useful in handling MAR assumption and extension to weighted/penalized generalized estimating equations.

Lastly, the simulation generation conditions were designed a priori and conclusions from this study may not be generalizability because the limited conditions investigated. For example, the study results may not be generalizable to all distributions. This might affect the generalizability of the study conclusions.

6.4 Future Directions

This dissertation focused on the development of repeated measures discriminant analysis procedures for discriminating between populations in multivariate non-normal repeated measures data. In addition to the strengths of this research, the limitations have raised a number of opportunities for future research in multivariate repeated measures studies. First, the assumption of complete multivariate repeated measures data in which there is no missing data on all outcomes and at all measurement occasions might not be realistic in multivariate repeated measures data often encountered in applied research. Multivariate repeated measures discriminant analysis based on mixed-effects models have been proposed for incomplete repeated measures data and have been shown to result in better classification accuracy when the missing data are assumed to be missing at random (MAR) but not on missing not at random (MNAR) or nonignorable missing data[21, 22, 30, 31]. Pattern mixture and selection models have been proposed to adjust for potential bias in models when it cannot be assumed that the mechanism of missingness is ignorable[32, 33] in multivariate repeated measures data. Therefore, future research will investigate the development of repeated measures discriminant analysis procedures based on these models with imputation models and developments in which mixed-effects models can be extended to these robust trimmed methods for classification. With regards to repeated measures discriminant analysis based GEE, further

research could be conducted to address drop-out problems in repeated measures by implementing weighted generalized estimating equations[34].

While repeated measures discriminant analysis procedures based on MTLE had higher classification accuracy even under extreme departures from non-normality for correctly specified covariance structures, mis-specification covariance structure on the repeated measurements in the classification rule could increase misclassification error in these models[35]. However, one does not know a priori which correlation structure is correct in multivariate repeated measures data analysis and therefore further research is warranted to investigate misspecification of covariance structures on these robust trimmed estimation methods.

Also, covariance pattern models and GEE models used for the development of repeated measures discriminant analysis procedures in this research can be fit to multivariate repeated measures data using packages and procedures available in common statistical software such as R (JGEE, multgee, geepack)[24-26] and SAS (proc MIXED, proc GENMOD)[23]. Future research will focus on the development of R packages that implement these repeated measures discriminant analysis procedures with example datasets. Such packages could promote the use of these procedures by clinical researchers for prognostic tools in clinical practice.

Finally, Copula models have been employed as an alternative class of robust procedures that jointly models mixed discrete and continuous longitudinal outcomes to develop classification models[36, 37]. Future research will focus on the comparison of robust repeated measures discriminant analysis procedures investigated in this study to copula approaches for repeated measures models for classification.

6.5 Conclusion

Our major findings show that repeated measures discriminant analysis based on trimmed estimators and multivariate GEE are more accurate in comparisons to the conventional discriminant analysis models based on MLE for classification in multivariate non-normal repeated measures. Also, our results reveal negative impact of correlation structure mis-specification on classification accuracy. We recommend that the choice between these classes of repeated measures models should be guided by a preliminary examination of the distribution of the data and the nature of correlation between multiple outcomes.

References

1. Hadi AS and Luceño A. Maximum trimmed likelihood estimators: A unified approach, examples, and algorithms. *Computational Statistics & Data Analysis* 1997; 25: 251-272.
2. Sajobi TT, Lix LM, Dansu BM, et al. Robust descriptive discriminant analysis for repeated measures data. *Computational Statistics & Data Analysis* 2012; 56: 2782-2794.
3. Rousseuw PJ and Leroy AM. Robust regression and outlier detection. Wiley, New York, 1987.
4. Qu A, Lindsay BG and Li B. Improving generalised estimating equations using quadratic inference functions. *Biometrika* 2000; 87: 823-836.
5. Wang X and Qu A. Efficient classification for longitudinal data. *Computational Statistics & Data Analysis* 2014; 78: 119-134.
6. Asar Ö and İlk Ö. Mmm: An r package for analyzing multivariate longitudinal data with multivariate marginal models. *Computer Methods and Programs in Biomedicine* 2013; 112: 649-654.
7. Cho H. The analysis of multivariate longitudinal data using multivariate marginal models. *Journal of Multivariate Analysis* 2016; 143: 481-491.
8. Inan G and Yucel R. Joint gees for multivariate correlated data with incomplete binary outcomes. *Journal of Applied Statistics* 2017; 44: 1920-1937.
9. Mardia KV. Measures of multivariate skewness and kurtosis with applications. *Biometrika* 1970; 57: 519-530.
10. Roy A. A new classification rule for incomplete doubly multivariate data using mixed effects model with performance comparisons on the imputed data. *Statistics in medicine* 2006; 25: 1715-1728.
11. Roy A. A note on testing of kronecker product covariance structures for doubly multivariate data. In: *Proceedings of the American Statistical Association, statistical computing section* 2007, pp.2157-2162.
12. Zhou J and Qu A. Informative estimation and selection of correlation structure for longitudinal data. *Journal of the American Statistical Association* 2012; 107: 701-710.
13. Khattree R and Naik DN. *Applied multivariate statistics with sas software*. SAS Institute Inc., 2018.
14. Filipiak K, Klein D and Roy A. Score test for a separable covariance structure with the first component as compound symmetric correlation matrix. *Journal of Multivariate Analysis* 2016; 150: 105-124.
15. Filipiak K, Klein D and Roy A. A comparison of likelihood ratio tests and rao's score test for three separable covariance matrix structures. *Biometrical Journal* 2017; 59: 192-215.
16. Roy A and Khattree R. On discrimination and classification with multivariate repeated measures data. *Journal of Statistical Planning and Inference* 2005; 134: 462-485.
17. Roy A and Khattree R. On implementation of a test for kronecker product covariance structure for multivariate repeated measures data. *Statistical Methodology* 2005; 2: 297-306.
18. Fitzmaurice GM, Laird NM and Ware JH. *Applied longitudinal analysis*. John Wiley & Sons, 2012.
19. Nasiri M, Faghihzadeh S, Majd HA, et al. Longitudinal discriminant analysis of hemoglobin level for predicting preeclampsia. *Iranian Red Crescent Medical Journal* 2015; 17.

20. Li Y, Wang Y, Wu G, et al. Discriminant analysis of longitudinal cortical thickness changes in alzheimer's disease using dynamic and network features. *Neurobiology of aging* 2012; 33: 427. e415-427. e430.
21. Hughes DM, Komárek A, Czanner G, et al. Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types. *Statistical methods in medical research* 2018; 27: 2060-2080.
22. Marshall G, De la Cruz-Mesía R, Barón AE, et al. Non-linear random effects model for multivariate responses with missing data. *Statistics in Medicine* 2006; 25: 2817-2830.
23. Institute S. *Sas user's guide: Statistics*. Sas Inst, 1985.
24. Højsgaard S, Halekoh U, Yan J, et al. Package 'geepack'. *R package version* 2016: 1.2-0.2015.
25. Inan G. Jgee: Joint generalized estimating equation solver. R package version, 2015.
26. Touloumis A. R package multgee: A generalized estimating equations solver for multinomial responses. *arXiv preprint arXiv:14105232* 2014.
27. Naik DN and Rao SS. Analysis of multivariate repeated measures data with a kronecker product structured covariance matrix. *Journal of Applied Statistics* 2001; 28: 91-105.
28. Krzyśko M and Skorzybut M. Discriminant analysis of multivariate repeated measures data with a kronecker product structured covariance matrices. *Statistical papers* 2009; 50: 817-835.
29. Lu N and Zimmerman DL. The likelihood ratio test for a separable covariance matrix. *Statistics & probability letters* 2005; 73: 449-457.
30. Fieuws S, Verbeke G, Maes B, et al. Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics* 2007; 9: 419-431.
31. Komárek A, Hansen BE, Kuiper EM, et al. Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statistics in medicine* 2010; 29: 3267-3283.
32. Little RJ. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association* 1993; 88: 125-134.
33. Little RJ. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the american statistical association* 1995; 90: 1112-1121.
34. Beunckens C, Sotto C and Molenberghs G. A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal binary data. *Computational statistics & data analysis* 2008; 52: 1533-1548.
35. Roy A and Khattree R. Classification rules for repeated measures data from biomedical research. *Computational Methods in Biomedical Research* 2007: 323-370.
36. de Leon AR and Wu B. Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine* 2011; 30: 175-185.
37. Ahn JY, Fuchs S and Oh R. A copula transformation in multivariate mixed discrete-continuous models. *arXiv preprint arXiv:200812411* 2020.