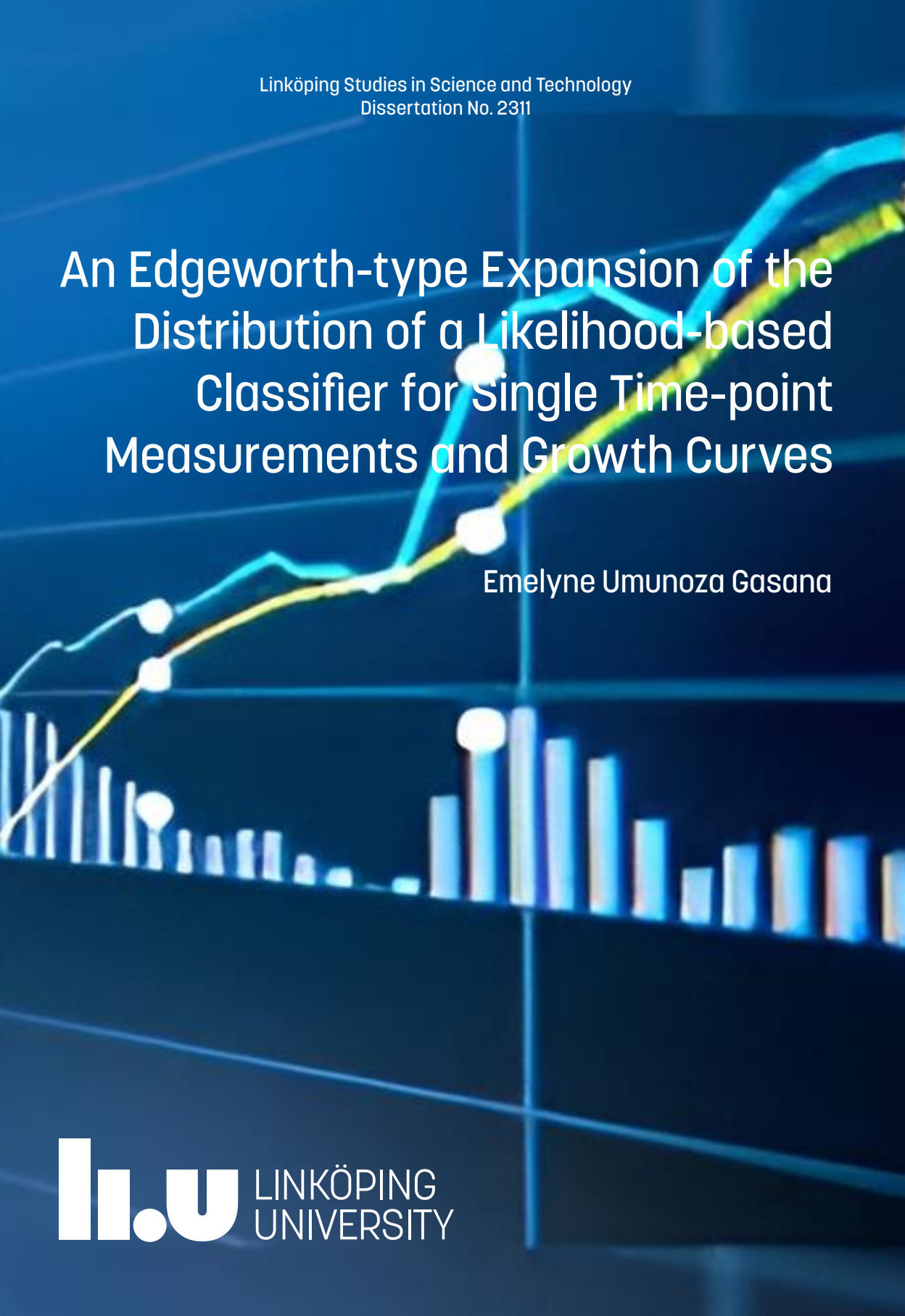
Linköping Studies in Science and Technology
Dissertation No. 2311

An Edgeworth-type Expansion of the Distribution of a Likelihood-based Classifier for Single Time-point Measurements and Growth Curves

Emelyne Umunoza Gasana



LINKÖPING UNIVERSITY

Linköping Studies in Science and Technology.
Dissertations.
No. 2311

An Edgeworth-type Expansion of the Distribution of a Likelihood-based Classifier for Single Time-point Measurements and Growth Curves

Emelyne Umunoza Gasana

Department of Mathematics
Mathematical Statistics
Linköping University
SE–581 83 Linköping, Sweden

Linköping 2023

Linköping Studies in Science and Technology.
Dissertations.
No. 2311

An Edgeworth-type Expansion of the Distribution of a Likelihood-based Classifier for Single Time-point Measurements and Growth Curves

Emelyne Umunoza Gasana

emelyne.umunoza.gasana@liu.se
https://doi.org/10.3384/9789180751537
Mathematical Statistics
Department of Mathematics
Linköping University
SE–581 83 Linköping, Sweden

ISBN 978-91-8075-152-0 (Print) 978-91-8075-153-7 (PDF) ISSN 0345-7524

Printed by LiU-Tryck, Linköping, Sweden 2023

To my amazing support system and to the women who defy the odds

Abstract

Emelyne Umunoza Gasana (2021). An Edgeworth-type Expansion of the Distribution of a Likelihood-based Classifier for Single Time-point Measurements and Growth Curves.
Doctoral dissertation No. 2311. ISBN 978-91-8075-152-0 (Print) 978-91-8075-153-7 (PDF). ISSN 0345-7524.

This thesis focuses on approximating misclassification errors of likelihood-based classifiers considering two cases. The first case assumes the allocation of a new observation into two normal populations. The second case classifies repeated measurements using the growth curve model, considering the fact that the new observation might not belong to any of the two predetermined populations but to an unknown population.

In this thesis, likelihood-based approaches were used to derive classification rules used to allocate a new observation in any of the two predefined normally distributed populations. Moreover, a two-step likelihood-based classification of growth curves is studied from which the distribution of a new observation is either drawn from any of the two predetermined populations or from an unknown population. Furthermore, moments of the classifiers were calculated and utilized to approximate the distribution of the proposed classifiers through an Edgeworth-type expansion. In addition, probabilities of misclassifications for the above-mentioned classifiers were estimated.

Keywords

Cumulants of discriminant function, Edgeworth-type expansion, growth curves classification, likelihood-based discriminant analysis, misclassification errors.

Populärvetenskaplig sammanfattning

Den här avhandlingen studeras approximativa felklassificeringssannolikheter hos likelihood-baserade klassificerare för två olika modeller. Den första modellen bygger på klassisk normalfördelning samt två populationer. I det andra fallet klassificeras upprepade mätningar, också med två populationer, med hjälp av så kallad tillväxtkurvmodellen. I fallet med upprepade mätningar ger klassificeraren att den nya observationen kanske inte tillhör någon av de två förutbestämda populationerna utan till en okänd population. I många tillämpningar är detta naturligt. Till exempel om man vill bestämma om en patient är frisk eller har en specifik sjukdom, så kan fallet vara att patienten har en annan sjukdom med snarlika symptom.

För att approximera fördelningarna för de likelihood-baserade klassificerarna, härleds första och andra momenten (väntevärde och varians), för att sedan nyttja en sorts Edgeworth-utveckling. Med hjälp av de approximerade fördelningarna, kan sedan de approximativa sannolikheterna för felklassificeringarna beräknas.

Acknowledgments

My sincere gratitude goes to my supervisors Prof. Dietrich von Rosen and Prof. Martin Singull for their supervision, technical support and wise advice, insightful review, encouragement, and guidance that led this work to a successful end.

My sincere acknowledgments are addressed to the Linköping University (LiU) for its appreciable policy of promoting education at all levels and providing an excellent environment, to Sida bilateral program for their financial support that allowed me to pursue my Ph.D. degree studies fruitfully. To Lindha Nilsson, Mathilda Kåhlin, Karin Johansson and the entire administrative staff of the Mathematics department who generously gave their time and support, and to the whole academic staff who helped me to sharpen my knowledge during this period I have been training in LiU and contributed significantly to the finalization of these studies.

To the entire Mathematics community at the University of Rwanda and the whole UR-Sida coordination team, I am grateful for your support.

I would like to extend my deepest and most heartfelt gratitude to my friend and colleague Denise for her moral support, fruitful exchanges, and shared experiences that made this journey a lot easier, to Cigdem, Jennifer, Roghi, and Vincent, for the great friendships and all fellow Ph.D. students and alumni, who contributed to these studies in terms of ideas, technical knowledge, "fika" and other social fun activities.

Last but not least, to my entire support system; my family and friends for their prayers and encouragement. In a special way, my wonderful mother, there are no profound words to express my gratitude for the love and support that you have given me since day one, to my husband, Juru, without whom all this would not be possible, and to my daughter, Iriza, who puts a smile to my face after a long and stressful day. Your endless love, sacrifice and continued support made this happen. Muri indashyikirwa. Ndabakunda.
To my beloved late father, "Promise fulfilled!". God is good!

Linköping, May 2, 2023
Emelyne Umunoza Gasana

Contents

Part I

Theoretical background and main results

1

Introduction

SEVERAL researchers in multivariate statistics have over the years shown interest in classification and discriminant analysis (Fisher, 1936, 1938; Wallace and Travers, 1938; Wald, 1944; Rao, 1948; Anderson, 1951; Srivastava and Khatri, 1979; Muirhead, 1982; McLachlan, 2004). Discrimination and classification applications are very common. In many cases, there are more than two populations to consider. However, in this thesis, we will mainly study cases involving two populations except when we consider the presence of an unknown population when classifying growth curves.

Suppose we are given a set of multivariate observations of size p and each observation comes from one of q predetermined populations with the same characteristics. These populations can, for instance, be medical healthy/non-healthy populations, plant species, levels of a customer's satisfaction with a new product, whether a client is eligible to receive a bank loan or not, or whether an e-mail is spam or not. In each of these situations, there are two purposes and ways to distinguish the distinct sets from each other; *discrimination* and *classification*.

Discriminant analysis uses information from predefined populations to determine a significance test or to define a distance measure that best emulates the separation between the populations to define a classification rule. Classification techniques use a classifier to allocate a new observation, or predict its belonging, in one of the well-defined populations, given a set of measurements derived from the predefined populations. Though the two approaches are different, classification techniques are often derived using discriminant analysis.

For example, a medical doctor receives a patient with weakness in the arms, one side of the face bending down, and difficult to speak. The patient's illness may either be a stroke or a seizure. The treatment of these two diseases is different since a stroke requires a more acute treatment than a simple seizure. Therefore, an incorrect diagnosis could lead to a

lethal outcome. A classification rule is constructed from past experience, say, based on the test from other patients treated previously, in order to discriminate between the two illnesses. On the other hand, consider a different scenario where a doctor aims to predict that another patient is at high or low risk for developing a stroke. The classifier allocates the patient into high- or low-risk groups based on their personal attributes (e.g., cholesterol level, body mass, family history, ...), lifestyle behavior (exercises, food diets, ...), etc.

Nowadays, machine learning arouses a lot of curiosity among researchers. In the machine learning literature, classification and discrimination are defined as a subdivision of supervised learning. The machine learning setting, classification techniques assume that observations are factual. They are described as labeled observations (Izenman, 2008). Classification algorithms in machine learning usually predict the probability that an unlabeled observation lies in one of the predetermined classes (populations) on the basis of the labeled observations (James et al., 2021). In statistics, however, a classification rule usually constitutes of a random model with unknown parameters.

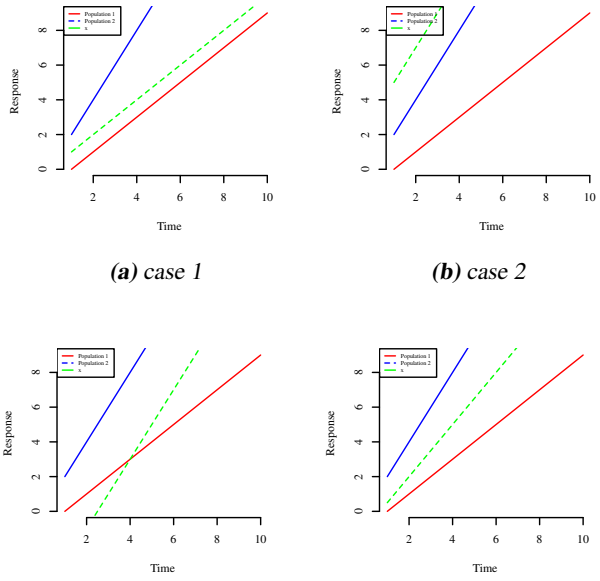
There exist different approaches for deriving a classifier. The most commonly used is the plug-in approach, that is, unknown parameters are replaced by their estimates, which usually results in a linear discriminant function. The likelihood approach is an alternative, which under normality results in a quadratic discriminant function. In this thesis, we derive likelihood-based classifiers that serve as alternatives to existing classifiers. However, as mentioned by Wald (1944), Sitgreaves (1952, 1961), these classifiers commonly reveal complex distributions.

There are many ways to approximate the distribution of a statistic. Asymptotic expansions are very common when it comes to the approximation of the distribution of the classification rule. However, there are a number of other effective approximations such as the Gram-Charlier series and Edgeworth expansions. Kollo and von Rosen (2005) proposed an Edgeworth-type expansion based on the normal distribution and is appropriate for approximating a likelihood estimator. The Edgeworth-type expansion is expressed in terms of moments and a standard normal density to approximate the densities connected to the classifiers.

Discriminant analysis techniques are not only used on cross-sectional data but also on repeated measurements when data is collected at several points in time. These data are often modeled by the generalized multivariate analysis of variance (GMANOVA) model, also called the Growth Curve model or the bilinear model, initiated by Potthoff and Roy (1964). Thereafter, several researchers such as Rao (1965), Khatri (1966) made contributions to the estimation of growth curves (von Rosen, 1991; Srivastava and von Rosen, 1999; von Rosen, 2018). The classification of growth curves is not well-studied in the literature, and therefore one of the aims of this thesis is to contribute to this area.

The classification of growth curves can be challenging. Consider Figure 1.1 and assume we aim to allocate a new observation $\boldsymbol{x}$ to any of the two populations; Population 1 and Population 2. In case 1, it is clear that the profile of the observation $\boldsymbol{x}$ follows the growth of Population 1 and hence belongs to it. Moreover, case 2 shows that the profile of $\boldsymbol{x}$ follows the growth of Population 2 but deviates from it. Furthermore, in case 3, $\boldsymbol{x}$ is close to Population 1 but its mean structure follows the growth of Population 2. On the

other hand, case 4 shows that $\boldsymbol{x}$ does not follow the growth of any of the two populations. Therefore, on all the 4 cases, only in case 1 the new observation $\boldsymbol{x}$ can be allocated to a predefined population with accuracy. In the remaining cases, the new observation $\boldsymbol{x}$ is thus drawn from an unknown population, that is, there is not enough information in the data to classify the observation into one of the groups. Except for Rao (1948) who briefly mentioned the possibility of classifying a new observation into an unknown population; this concept remains untouched in the literature. In this thesis, a two-step classification of growth curves is considered and it takes into account the fact that a new observation does not follow the mean structure of any of the two populations.



***(a)** case 1* ***(b)** case 2*

***(c)** case 3* ***(d)** case 4*

***Figure 1.1:** Example of growth curves which fit/no fit one of two populations.*

There is always a chance that a classification rule allocates an observation to the wrong group. In statistics, it is of interest to estimate the probability that a classifier makes such an error. It is called the probability of misclassification or misclassification error. The misclassification error measures the goodness of a classifier (Fujikoshi et al., 2011). However, it is often very complicated to express misclassification probabilities. In this thesis, the distribution of the classifiers is approximated using an Edgeworth-type expansion.

1.1 Objectives

The objective of the present thesis is to contribute to the development of new approaches for approximating probabilities of misclassification for the likelihood-based discriminant

function for discriminating between two normally distributed populations and for the classification of growth curves which serve as good alternatives to existing methods. The specific objectives are the following:

(i) to derive likelihood-based classification rules and calculate the first two moments of the classifiers;

(ii) to propose an approximation of the probabilities of misclassification via Edgeworth-type expansions;

(iii) calculate the first two moments of the two-step classification rule for classifying repeated measurements using the Growth Curve model;

(iv) to derive an approximation for the probabilities of misclassification for growth curves through Edgeworth-type expansions.

1.2 Thesis outline

This thesis consists of two parts. The first part provides the background and summary and presents the necessary concepts behind the four papers presented in the second part of the present thesis.

1.2.1 Outline of Part I

Chapter 1 is the introduction which comprises the objectives, notations used throughout the thesis, a summary of the papers, and contributions. Chapter 2 consists of a literature review that provides the historical aspects of the concepts supporting this thesis. Chapter 3 produces some results for the multivariate statistics applied in this thesis. Chapter 4, gives a general introduction to the classification analysis, focusing on the likelihood approach and the derivation of misclassification errors. Part I ends with the fifth chapter, which presents a summary of contributions, concluding remarks, and suggestions for future research.

1.2.2 Outline of Part II

Part II consists of four papers. Below follows a short summary of each of the papers.

Paper A: Moments of the likelihood-based discriminant function.

Umunoza Gasana, E., von Rosen, D., and Singull, M. (2022). Moments of the likelihood-based discriminant function. *Communications in Statistics - Theory and Methods, p. 1–13.*

In Paper A (Umunoza Gasana et al., 2022b), we propose two classification rules using a maximum likelihood procedure to estimate the unknown parameters for the purpose of classifying a new observation into two known multivariate normal populations with a known and unknown covariance matrix. The two classifiers which occur as quadratic functions of the new observation $\boldsymbol{x}$ serve as good alternatives to existing linear and quadratic discriminant functions. The two classifiers are dependent on the sample size

and the Mahalanobis distance. We calculated the first two moments of each of the two proposed discriminant functions which will be used to approximate their distributions in Paper B (Umunoza Gasana et al., 2022a).

Paper B: Approximated misclassification errors for the likelihood based discriminant function via Edgeworth-type expansion.

Umunoza Gasana, E., von Rosen, D., and Singull, M. (2022). Approximated misclassification errors for the likelihood based discriminant function via Edgeworth-type expansion. *Linköping University Electronic Press*, LiTH-MAT-R–2021/08–SE.

In Paper B (Umunoza Gasana et al., 2022a), we have approximated the probability distributions of the two classification rules proposed in Paper A using an Edgeworth-type expansion in terms of the moments derived in Paper A (Umunoza Gasana et al., 2022b). The moments are based on moment relations for the normal, Wishart, and inverse Wishart distributions. We further derived the misclassification errors and assess the results via a simulation study comparing them with the results of the well-known $W-$ and $Z-$ rules.

Paper C: Moments of the likelihood-based classification function using growth curves.

Umunoza Gasana, E., von Rosen, D., and Singull, M. (2023). Moments of the likelihood-based classification function using growth curves. *Linköping University Electronic Press*, LiTH-MAT-R–2023/01–SE.

In Paper C (Umunoza Gasana et al., 2023b), we considered a discriminant function for growth curves taking into account the possibility for a new observation to be classified to an unknown population. Such a classifier developed by von Rosen and Singull (2022) consists of two criteria. We establish the expected value and variance of these two criteria to express the basic characteristics of their respective distribution.

Paper D: Edgeworth-type expansion of the density of the classifier when growth curves are classified via the likelihood.

Umunoza Gasana, E., von Rosen, D., and Singull, M. (2023). Edgeworth-type expansion of the density of the classifier when growth curves are classified via the likelihood. *Linköping University Electronic Press*, LiTH-MAT-R-2023/02.

Paper D (Umunoza Gasana et al., 2023a) is a continuation of Paper C (Umunoza Gasana et al., 2023b). We determined the approximations for the distribution of each of the two criteria of the two-step classifier derived by von Rosen and Singull (2022) by an Edgeworth-type expansion using the two moments calculated in Paper C. We have derived the misclassification errors of the two-step classifier using Edgeworth-type expansions for the distributions of both criteria separately.

1.3 Notations and Definitions

Here is a list of symbols, operators and acronyms used throughout the present thesis. Generally, bold lowercase letters are used to denote vector-valued and bold uppercase letters are used for matrix-valued variables. However, there might be exceptions to these general rules. Note that all vectors are column vectors. Any deviations are explained in the text.

Symbols and Operators

$\mathbf{0}_n$	Null vector of size n, sometimes just $\mathbf{0}$ is used
$\mathbf{1}_n$	Vector of n ones, sometimes just $\mathbf{1}$ is used
$\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	p-dimensional normal (Gaussian) distribution with mean value $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{N}_{p,n}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Psi})$	Matrix normal distribution
$W_p(\boldsymbol{\Sigma}, n)$	p-dimensional central Wishart distribution with n degrees of freedom and scale matrix $\boldsymbol{\Sigma}$
$\Phi(\boldsymbol{x})$	Normal (Gaussian) distribution function
$f_{\boldsymbol{x}}(\boldsymbol{x})$	Probability density function of $\boldsymbol{x}$
$L(\cdot)$	Likelihood function
$l(\cdot)$	Log-likelihood function
$\Gamma(\cdot)$	Gamma function
$\frac{d(\cdot)}{d\boldsymbol{x}}$	Derivative of a function with respect to $\boldsymbol{x}$
$\otimes$	Kronecker product
$\sim$	Denotes "is distributed according to"
$\in$	Belongs to
$\xrightarrow{p}$	Convergence in probability
π_i	i^{th} population
$P(x \leq k)$	Probability that the random variable x is less than or equal to k
$P(\boldsymbol{x} \rightarrow \pi_i)$	Probability that an observation $\boldsymbol{x}$ is classified into population π_i
$\boldsymbol{I}_n$	Identity matrix of size n
$\boldsymbol{K}_{p,q}$	Commutation matrix
$\lvert\boldsymbol{A}\rvert$	Determinant of matrix $\boldsymbol{A}$
$\mathrm{tr}\boldsymbol{A}$	Trace of matrix $\boldsymbol{A}$
$\boldsymbol{A}^\intercal$	Transpose of matrix $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	Inverse of matrix $\boldsymbol{A}$
$\mathcal{C}(\boldsymbol{A})$	Column space
$\mathcal{C}(\boldsymbol{A})^\perp$	Orthogonal complement to $\mathcal{C}(\boldsymbol{A})$
$\boldsymbol{A}^o$	Matrix $\boldsymbol{A}$ with columns generating $\mathcal{C}(\boldsymbol{A})^\perp$
$\boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}}$	Projection on $\mathcal{C}(\boldsymbol{A})$, $\boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}} = \boldsymbol{A}(\boldsymbol{A}^\intercal\boldsymbol{S}^{-1}\boldsymbol{A})^-\boldsymbol{A}^\intercal\boldsymbol{S}^{-1}$
$E[\boldsymbol{x}]$	Expectation of the random variable $\boldsymbol{x}$
$E[\boldsymbol{x}\lvert\boldsymbol{y}]$	Conditional expectation of the random variable $\boldsymbol{x}$, given the random variable $\boldsymbol{y}$

$\mathrm{Var}[\boldsymbol{x}]$	Variance-covariance matrix of the random variable $\boldsymbol{x}$, i.e. $\mathrm{Var}[\boldsymbol{x}] = E[\boldsymbol{x}\boldsymbol{x}^\intercal] - E[\boldsymbol{x}]E[\boldsymbol{x}^\intercal]$
$\min$	Minimum
$\max$	Maximum
$\mathrm{vec}()$	Vectorization of a matrix
$(\boldsymbol{X})()^\intercal$	$(\boldsymbol{X})(\boldsymbol{X})^\intercal$
$\boldsymbol{\Sigma} > 0$	$\boldsymbol{\Sigma}$ is positive definite
$\boldsymbol{\Sigma} \geq 0$	$\boldsymbol{\Sigma}$ is positive semi-definite.

Abbreviations and Acronyms

LDA	Linear Discriminant Analysis
QDA	Quadratic Discriminant Analysis
MANOVA	Multivariate Analysis of Variance
GMANOVA	Generalized Multivariate Analysis of Variance
MLE	Maximum Likelihood Estimator (Estimate)
pdf	Probability density function
s.t.	Subject to
w.r.t.	With respect to
p.d.	Positive definite

1.4 Author's contributions

All aforementioned works are co-authored with Dietrich von Rosen and Martin Singull. I have contributed by carrying out all detailed calculations, deriving the results, and writing the articles. The idea of research problems was mainly suggested by Dietrich von Rosen and Martin Singull and they both proofread, and reviewed the thesis. Moreover, together with Martin Singull, I have done the coding and simulations.

1.5 Awards from conference presentations

During my doctoral studies, I have attended and presented my research at a number of conferences, from which on one occasion, I was awarded.

- Prize Winner of Young Scientists Awards – LinStat2022 (The International Conference on Trends and Perspectives in Linear Statistical Inference), 4–8 July 2022, Tomar, Portugal. I presented Paper B (Umunoza Gasana et al., 2022a).

2

Literature review

CURRENTLY, there is an enormous literature on discriminant analysis and classification techniques which makes it difficult to provide a complete review of the topic. Therefore, in this chapter, we try to give historical perspectives on discriminant analysis and some preliminary concepts of discriminant analysis used throughout the papers included in the present thesis.

Since the early twentieth-century several studies took interest in the idea of separating between distinct sets, mostly starting by discriminating between two populations (Pearson, 1915, 1926; Mahalanobis, 1925, 1930; Barnard, 1935; Fisher, 1936, 1938; Bose, 1936; Rao, 1966) and more generally, allocating observations into several groups (Anderson, 1951). Even though Fisher (1936, 1938) is to be considered the pioneer of discriminant analysis, according to Kendall (1957), classification into multiple populations was initiated earlier by Karl Pearson. However, Pearson (1911) and most research studies before Fisher (1936) studied the differences between groups solely based on the sample characteristics such as moments or frequency tables, ignoring the correlations between the observed variables (McLachlan, 1992). On the other hand, Fisher (1936) used a set of multiple measurements from the observed groups to define an equation that maximizes the distance between the groups, where the population densities and prior probabilities are unknown. This resulted in a classification rule famously known as Fisher's linear discriminant analysis (LDA). Wallace and Travers (1938), Smith (1936) and Rao (1948) applied Fisher's technique in different areas. Moreover, one cannot talk about discriminant analysis and neglect to mention the major contributions made by Hotelling (1931) with his famous T^2 statistic and by Mahalanobis (1936) for his conventional Mahalanobis distance, Δ^2 and other distance measures. These distance measures are referred to by Fisher (1938) and many of the recent studies on discriminant analysis. Hodges (1955), Huberty (1975), Srivastava and Khatri (1979), Muirhead (1982), McLachlan (1992), Anderson (2003), and Huberty and Olejnik (2006) provide more references on classification and discriminant analysis.

There exist several techniques for deriving a classifier, considering the same known covariance matrix across groups (Fisher, 1936, 1938; Wald, 1944; Anderson, 1951) or unknown parameters (Rao, 1948; Kudo, 1959, 1960; Anderson, 2003). Fisher (1936) determined a linear classifier when parameters are known and Anderson (1951) used a plug-in approach to compute a linear classifier whereas Kudo (1959) and Srivastava and Khatri (1979), for example, used a likelihood approach to derive quadratic classification functions. McLachlan (1992), Chapter 3 is dedicated to different approaches to discriminate under normality assumptions.

In the late 1950s, studies in multivariate statistics expanded to discriminating between repeated measurements. Modeling multivariate repeated measurements data was first introduced by Potthoff and Roy (1964) who generalized the existing multivariate analysis of variance (MANOVA). For more references to the MANOVA model, see Srivastava and Khatri (1979), Muirhead (1982), Anderson (2003). Potthoff and Roy (1964)'s Growth Curve model is also known as the generalized MANOVA (GMANOVA) model or the bilinear regression model. Earlier Wishart (1938) analyzed growth curves in a repeated measurement setting but did not assume any dependency between the repeated measurements: see Gleser and Olkin (1970), Woolson and Leeper (1980), Srivastava and von Rosen (1999), and von Rosen (1991, 2018) for more reviews of the Growth Curve model.

Though the development of discriminant analysis blossomed over the years, the classification of growth curves provides limited literature. Burnaby (1966), who did not refer to Potthoff and Roy (1964)'s Growth Curve model, is one of the first researchers to examine classification techniques for multiple measurements drawn on the same subject at several time points but considered growth curves as a nuisance parameter that needs to be removed. However, growth curves have been briefly discussed before by Rao (1958, 1959). Afterward, it is worth mentioning Lee (1977, 1982) who applied Bayesian and non-Bayesian processes for the allocation of growth curves. McLachlan (1992), Section 3.7.6, introduces the classification of growth curves.

In Statistics, we are interested in misclassification errors. McLachlan (1992) defines those errors as a measure of the global performance of the classifier. However, the distribution of normal-based classifiers is often highly complex (Sitgreaves, 1952, 1961; Wald, 1944; Anderson, 2003) and has been investigated by many researchers over the years (Wald, 1944; Anderson, 1951; Sitgreaves, 1952; Okamoto, 1963; Siotani, 1982). To obtain an exact distribution of the discriminant rule with unknown parameters is difficult and several authors have applied asymptotic expansions (Sitgreaves, 1961; Bowker and Sitgreaves, 1961; Memon and Okamoto, 1971; Siotani and Wang, 1975; Siotani, 1977; Anderson, 2003) and (Fujikoshi, 1987; Kollo et al., 2007).

The history of approximating the distribution of a statistic can be traced back to the early 1800s. Laplace (1811) developed an idea of estimating a frequency function using a series containing a normal density, Hermite polynomials, and their expectations. Using Laplace's results, a number of authors developed density approximations through quantities such as cumulants (Edgeworth, 1907; Thiele, 1873, 1889; Cornish and Fisher, 1938) and moments (Gram, 1879). Gram-Charlier expansion, Edgeworth expansion, saddle point approximation, and the Cornish-Fisher expansion are the most popular density approximations, see Kollo and von Rosen (1998), Hald (2000), Hald and Steffensen

(2002), Lauritzen et al. (2002), Kollo and von Rosen (2005), von Rosen (2018) for detailed literature and Gupta and Panchapakesan (1982) provides a brief review on Edgeworth expansions and their use in statistics. However, it is worth noting that expansions may not be densities. Chapter 3 assesses distribution expansions, including an Edgeworth-type expansion which is of interest in the present thesis.

3

Multivariate distributions

MULTIVARIATE statistics encompass the concurrent observation and analysis of more than one response variable. In this chapter, some definitions, notions, and important results needed to derive the main results of this thesis are given.

3.1 Useful definitions

Definition 3.1. A symmetric $p \times p$ matrix $\boldsymbol{M}$ is said to be positive definite (p.d) if $\boldsymbol{x}^\intercal \boldsymbol{M} \boldsymbol{x} > \boldsymbol{0}$ and negative definite if $\boldsymbol{x}^\intercal \boldsymbol{M} x < \boldsymbol{0}$ for any vector $\boldsymbol{x} \neq \boldsymbol{0}$.

Definition 3.2. A $p \times q$ matrix $\boldsymbol{M}$ is said to be a partitioned matrix if it consists of mn submatrices, each of size $p_i \times q_j$, such that

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{M}_{11} & \boldsymbol{M}_{12} & \cdots & \boldsymbol{M}_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{M}_{m1} & \boldsymbol{M}_{m2} & \cdots & \boldsymbol{M}_{mn} \end{pmatrix}, \tag{3.1}$$

where $\sum_{i=1}^{m} p_i = p$ and $\sum_{j=1}^{n} q_j = q$. It is usually denoted $\boldsymbol{M} = [\boldsymbol{M}_{ij}]$, $i \in \{1, \ldots, m\}$, $j \in \{1, \ldots, v\}$.

Definition 3.3. The partitioned matrix $\boldsymbol{K}_{p,q} : pq \times pq$ consisting of $q \times p$-blocks is called commutation matrix, if

$$(\boldsymbol{K}_{p,q})_{(i,j)(g,h)} = \begin{cases} 1; & g = j, h = i, \quad i, h \in \{1, \ldots, p\}; \quad k, g \in \{1, \ldots, q\}, \\ 0; & \text{otherwise.} \end{cases}$$

Definition 3.4. The vectorization of a $p \times q$ matrix $\boldsymbol{M}$, is the $pq \times 1$ column vector

$$\text{vec}\boldsymbol{M} = [m_{11}, \ldots, m_{p1}, m_{12}, \ldots, m_{p2}, \ldots, m_{1q}, \ldots, m_{pq}]^\intercal.$$

Definition 3.5. The Kronecker product of $\boldsymbol{M} = (m_{ij})$ and $\boldsymbol{N} = (n_{kl})$ is defined as $\boldsymbol{M} \otimes \boldsymbol{N} = (m_{ij}\boldsymbol{N})$.

Definition 3.6. For matrices $\boldsymbol{M}$ and $\boldsymbol{W}$, the projection on $\mathcal{C}(\boldsymbol{M})$ is given by

$$\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{W}} = \boldsymbol{M}(\boldsymbol{M}^\intercal \boldsymbol{W}^{-1}\boldsymbol{M})^- \boldsymbol{M}^\intercal \boldsymbol{W}^{-1}, \quad \boldsymbol{W} > 0, \tag{3.2}$$

where "$^-$" denotes an arbitrary g-inverse. If $\boldsymbol{W} = \boldsymbol{I}$, the orthogonal projector equals $\boldsymbol{P}_{\boldsymbol{M}} = \boldsymbol{M}(\boldsymbol{M}^\intercal \boldsymbol{M})^- \boldsymbol{M}^\intercal$

3.2 Normal distributions

Univariate and multivariate normal distributions have made an important contribution to statistics. Here, we will present the simplest form of the normal distribution for which a density function exists.

Definition 3.7 (Univariate normal distribution). Let x be univariate normally distributed with mean $\mu \in \mathcal{R}$ and variance $\sigma^2 > 0$, which is denoted, $x \sim \mathcal{N}(\mu, \sigma^2)$. Its probability density function is given by

$$f(x) = (2\pi)^{-\frac{1}{2}}(\sigma^2)^{-\frac{1}{2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (-\infty < x < \infty).$$

Example 3.1: Standard normal distribution

Assume x is a univariate normal distribution variable with mean $\mu = 0$ and variance $\sigma = 1$, denoted $x \sim \mathcal{N}(0, 1)$. Then we have the standard univariate normal distribution with density given by

$$f(x) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2}. \tag{3.3}$$

Definition 3.8 (Multivariate normal distribution). Let $\boldsymbol{x}$ be a $p \times 1$ random vector, $\boldsymbol{x} = (x_1, \ldots, x_p)^\intercal$, that is distributed according to a multivariate normal distribution with $p \times 1$ mean vector $\boldsymbol{\mu}$ and $p \times p$ covariance matrix $\boldsymbol{\Sigma}$, denoted as $\boldsymbol{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If $\boldsymbol{\Sigma} > 0$, then its probability density function is given by

$$f_x(\boldsymbol{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^\intercal\} \right\}. \tag{3.4}$$

The unknown parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are usually estimated using the maximum likelihood method. Let $\boldsymbol{x}_i \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $i \in \{1, 2, \ldots, n\}$ be independent random observation vectors drawn from a multivariate normal distribution. Hence, the joint likelihood function

equals

$$L(\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$= (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^\intercal\right\}\right\}.$$

Let $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$. Then, the MLEs of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are respectively given by

$$\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i = \frac{1}{n}\boldsymbol{X}\mathbf{1}_n, \quad \widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\boldsymbol{S},$$

where

$$\boldsymbol{S} = \sum_{i=1}^{n}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}})^\intercal = \boldsymbol{X}(\boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\intercal)\boldsymbol{X}^\intercal.$$

Moments and cumulants for the normal distribution are widely applied in statistics. On occasions, moments of quadratic forms are needed. Since the discriminant rules studied in this paper can all be written as a quadratic form, we find it useful to express the first and second moments displayed in the theorems below.

Theorem 3.1 (Mathai and Provost (1992))
Let $\boldsymbol{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and let $\boldsymbol{M}$ be a $p \times p$ constant matrix such that $\boldsymbol{M} = \boldsymbol{M}^\intercal$. The h^{th} cumulant of the quadratic form $\boldsymbol{x}^\intercal\boldsymbol{M}\boldsymbol{x}$ equals

$$\psi_h(\boldsymbol{x}^\intercal\boldsymbol{M}\boldsymbol{x}) = 2^{h-1}h!\left\{\frac{\mathrm{tr}\{(\boldsymbol{M}\boldsymbol{\Sigma})^h\}}{h} + \boldsymbol{\mu}^\intercal(\boldsymbol{M}\boldsymbol{\Sigma})^{h-1}\boldsymbol{M}\boldsymbol{\mu}\right\}, \quad h \geq 1.$$

The first two cumulants of the quadratic form can directly be deduced from Theorem 3.1 above. The results are used in Umunoza Gasana et al. (2022b, 2023b).

Example 3.2

Let $\boldsymbol{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{M}$ be a constant matrix, then

(i) $E[\boldsymbol{x}^\intercal\boldsymbol{M}\boldsymbol{x}] = \mathrm{tr}\{\boldsymbol{M}\boldsymbol{\Sigma}\} + \boldsymbol{\mu}^\intercal\boldsymbol{M}\boldsymbol{\mu}$;

(ii) $Var[\boldsymbol{x}^\intercal\boldsymbol{M}\boldsymbol{x}] = 2[\mathrm{tr}\{(\boldsymbol{M}\boldsymbol{\Sigma})^2\} + 2\boldsymbol{\mu}^\intercal\boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}\boldsymbol{\mu}]$.

Multivariate Hermite polynomials are commonly applied when one wishes to approximate multivariate distribution functions through a multivariate normal distribution.

Definition 3.9. The matrix $H_i(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, called the i^{th}-degree multivariate Hermite polynomial, for the vector mean $\boldsymbol{\mu}$ and the covariance $\boldsymbol{\Sigma} > 0$, equals

$$H_i(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{(-1)^i}{f_{\boldsymbol{x}}(\boldsymbol{x})}\frac{d^i f_{\boldsymbol{x}}(\boldsymbol{x})}{d\boldsymbol{x}^i}, \quad i \in \mathbb{N}_0, \tag{3.5}$$

where $\mathbb{N}_0$ is the set of positive integers and $f_{\boldsymbol{x}}(\boldsymbol{x})$ is the density function of the normal distribution, $\boldsymbol{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, given by (3.4) and $\frac{d^i}{d\boldsymbol{x}^i}$ is the i^{th} matrix derivative defined as in Kollo and von Rosen (2005).

Usually, Hermite polynomials are defined for a centered normal distribution (Kollo and von Rosen, 2005). When $\boldsymbol{x}$ follows a standard normal distribution, the first four Hermite polynomials can be calculated directly and they are used when proving results in Paper B.

Example 3.3

Assume that $\boldsymbol{x}$ follows a standard normal distribution, $\boldsymbol{x} \sim \mathcal{N}(0, 1)$. The i^{th} Hermite polynomial $H_i(x, 1)$, $i \in \{0, 1, 2, 3\}$ equal

$$H_0(x, 1) = 1; \quad H_1(x, 1) = x; \quad H_2(x, 1) = x^2 - 1; \quad H_3(x, 1) = x^3 - 3x.$$

3.3 The Wishart distribution

The Wishart distribution is named after John Wishart (1928) who first derived it. It is often considered as a multivariate extension of the χ^2-distribution (Kollo and von Rosen, 2005; Fujikoshi et al., 2011).

3.3.1 The central and non-central Wishart distribution

Definition 3.10.

(i) The random $p \times p$ matrix $\boldsymbol{W}$ is Wishart-distributed with n degrees of freedom if and only if $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{X}^\intercal$, for some $\boldsymbol{X} \sim \mathcal{N}_{p,n}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{I})$, $\boldsymbol{\Sigma} \geq 0$.

(ii) If $\boldsymbol{\mu} = \boldsymbol{0}$, we get the central Wishart distribution denoted $\boldsymbol{W} \sim W_p(\boldsymbol{\Sigma}, n)$, and if $\boldsymbol{\mu} \neq \boldsymbol{0}$, we get the non-central Wishart distribution, denoted $\boldsymbol{W} \sim W_p(\boldsymbol{\Sigma}, n, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}^\intercal\boldsymbol{\mu}$ is the non-centrality parameter.

The covariance matrix $\boldsymbol{\Sigma}$ is usually assumed to be an unknown parameter. When $\boldsymbol{W}$ is a scalar with $\boldsymbol{X} \sim \mathcal{N}_n(\boldsymbol{0}, \boldsymbol{I}_n)$, then the Wishart matrix is identical to a central χ^2-distributed variable with n degrees of freedom. Furthermore, if $\boldsymbol{W}$ is a scalar with $\boldsymbol{\Sigma} = 1$ but with nonzero mean μ, then we get a non-central χ^2-distribution with n degrees of freedom and a non-centrality parameter equal to μ^2. Note that working with the Wishart matrix gets complicated when $\boldsymbol{\Psi} \neq \boldsymbol{0}$.

Theorem 3.2
Let the $p \times p$ matrix $\boldsymbol{W} \sim W_p(\boldsymbol{\Sigma}, n)$, where $\boldsymbol{\Sigma} > 0$ and $n \geq p$. Then its density function is given by

$$f_{\boldsymbol{W}}(\boldsymbol{W}) = \begin{cases} \frac{|\boldsymbol{W}|^{\frac{n-p-1}{2}} e^{-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{W}\right\}}}{2^{\frac{np}{2}} \pi^{\frac{p(p-1)}{4}} \prod_{i=1}^{p} \Gamma[\frac{1}{2}(n+1-i)] |\boldsymbol{\Sigma}|^{\frac{n}{2}}}, & if \quad \boldsymbol{W} > 0, \\ 0, & otherwise. \end{cases}$$

The proof of Theorem 3.2 is for example derived in Anderson (2003, pp. 252-255).

3.3.2 The inverted Wishart distribution and its moments

If the inverse exists, the random matrix $\boldsymbol{W}^{-1}$ is said to follow an inverted Wishart distribution where $\boldsymbol{W} \sim W_p(\boldsymbol{\Sigma}, n)$. The results of this section are mostly applied in Paper A and Paper C.

Theorem 3.3
Let $\boldsymbol{W} \sim W_p(\boldsymbol{\Sigma}, n)$ and let $\boldsymbol{M}$ be a $p \times p$ constant matrix. Then

(i) $E[\boldsymbol{W}^{-1}] = k_0\boldsymbol{\Sigma}^{-1}$, $n-p-1>0$;

(ii) $E[\boldsymbol{W}^{-1}\boldsymbol{M}\boldsymbol{W}^{-1}] = k_1\boldsymbol{\Sigma}^{-1}\boldsymbol{M}\boldsymbol{\Sigma}^{-1} + k_2[\boldsymbol{\Sigma}^{-1}\boldsymbol{M}^{\intercal}\boldsymbol{\Sigma}^{-1} + \mathrm{tr}\{\boldsymbol{M}\boldsymbol{\Sigma}^{-1}\}\boldsymbol{\Sigma}^{-1}]$, $n-p-3>0$;

(iii) $Var[\boldsymbol{W}^{-1}] = k_2(\boldsymbol{I}+\boldsymbol{K}_{p,p})(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1}) + (k_1-k_0^2)\mathrm{vec}\boldsymbol{\Sigma}^{-1}\mathrm{vec}^{\intercal}\boldsymbol{\Sigma}^{-1}$, $n-p-3>0$,

where

$$k_0 = \frac{1}{n-p-1},\ k_1 = \frac{n-p-2}{(n-p)(n-p-3)}k_0,\ k_2 = \frac{1}{n-p-2}k_1.$$

The proofs and technical expressions in Theorem 3.3 (i) and (ii) above can be found in von Rosen (1988), (iii) in Kollo and von Rosen (2005), Fujikoshi et al. (2011).

Lemma 3.1
Let $\boldsymbol{W} \sim W_p(\boldsymbol{\Sigma}, n-2)$. For $\boldsymbol{M} : p \times q$ of rank q and $n > p-3$,

$$E[\boldsymbol{W}^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{W}}] = \frac{q}{(n-p-3)(n-p+q-3)}\boldsymbol{\Sigma}^{-1} + \frac{1}{n-p+q-3}\boldsymbol{\Sigma}^{-1}\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{\Sigma}},$$

where $\boldsymbol{P}_{\boldsymbol{M},\boldsymbol{W}} = \boldsymbol{M}(\boldsymbol{M}^{\intercal}\boldsymbol{W}^{-1}\boldsymbol{M})^{-}\boldsymbol{M}^{\intercal}\boldsymbol{W}^{-1}$.

Lemma 3.1 is proved by von Rosen (2018), Theorem B.26 (i). Lemma 3.1 is specifically used in Paper C for calculating moments of the likelihood-based discriminant rule for classifying growth curves.

3.4 Edgeworth-type expansion

Since early 19^{th}, quantities such as cumulants and moments have been used to approximate a frequency function with a complex density (Hald and Steffensen, 2002). The most common density approximations are Gram-Charlier and Edgeworth expansions. Edgeworth (1907) introduced the so-called Edgeworth series that can be used to approximate a probability density with respect to cumulants. Several authors afterward applied this concept to express Edgeworth expansions (Ranga Rao, 1960; Davis, 1976; Skovgaard, 1986; Barndorff-Nielsen and Cox, 1989) and the Edgeworth-type expansions (Kollo and von Rosen, 2005). Edgeworth expansions, using the standard normal distribution, are usually applied to find approximations of the sample distributions. Edgeworth-type expansions are often obtained through the multivariate normal distribution $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$. In this section, we consider the density approximation proposed by Kollo and von Rosen (2005), where

a density is approximated in terms of its cumulants, Hermite polynomials, and a simpler density function. Note that a density expansion is often not a true density.

Theorem 3.4 (Kollo and von Rosen (2005))
Let $\boldsymbol{u}$ be a random p-vector with finite first four cumulants, then we can approximate its density $f_{\boldsymbol{u}}(\boldsymbol{x})$ through the density $f_N(\boldsymbol{x})$ of the normal distribution, $\mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, by the Edgeworth-type expansion

$$f_{\boldsymbol{u}}(\boldsymbol{x}) \approx f_N(\boldsymbol{x}) \Big\{ 1 + E[\boldsymbol{u}]^{\intercal} \boldsymbol{H}_1(\boldsymbol{x}, \boldsymbol{\Sigma}) + \frac{1}{2} \mathrm{vec}^{\intercal} \left(\mathrm{Var}[\boldsymbol{u}] - \boldsymbol{\Sigma} + (E[\boldsymbol{u}])(E[\boldsymbol{u}])^{\intercal} \right) \mathrm{vec} \boldsymbol{H}_2(\boldsymbol{x}, \boldsymbol{\Sigma}) + \frac{1}{6} \left(\mathrm{vec}^{\intercal} c_3[\boldsymbol{u}] + 3\mathrm{vec}^{\intercal} (\mathrm{Var}[\boldsymbol{u}] - \boldsymbol{\Sigma}) \otimes (E[\boldsymbol{u}])^{\intercal} + (E[\boldsymbol{u}])^{\intercal \otimes 3} \right) \mathrm{vec} \boldsymbol{H}_3(\boldsymbol{x}, \boldsymbol{\Sigma}) + \cdots \Big\}, \tag{3.6}$$

where $c_3[\boldsymbol{u}]$ is the third cumulant of $\boldsymbol{u}$, and the multivariate Hermite polynomials $\boldsymbol{H}_i(\boldsymbol{x}, \boldsymbol{\Sigma})$, $i \in \{1, 2, 3\}$, directly result from Definition 3.9.

To obtain the Edgeworth-type expansion for the distribution of a classifier, a less challenging way is to approximate through a standard normal distribution. In addition, since every classifier is a 1-dimensional variable, $p = 1$. In the following example, we present that case, which is applied in Umunoza Gasana et al. (2022a, 2023b).

Example 3.4: (Paper B)

The density function $f_u(x)$ of a random variable u can be approximated through the density $f_N(x)$ of the standard normal distribution, $\mathcal{N}(0, 1)$, by the Edgeworth-type expansion

$$f_u(x) \approx f_N(x) \Big\{ 1 + E[u] H_1(x, 1) + \frac{1}{2} \left(Var[u] - 1 + (E[u])^2 \right) H_2(x, 1) + \frac{1}{6} \left(c_3[u] + 3(Var[u] - 1) E[u] + (E[u])^3 \right) H_3(x, 1) \Big\}, \tag{3.7}$$

where $H_i(x, 1)$, $i \in \{1, 2, 3\}$, are given in Example 3.3 and the pdf of the standard normal distribution is given by (3.3).

3.5 The Growth Curve model

As discussed before, the Growth Curve model, also known as the GMANOVA or the bilinear model is an extension of the MANOVA model. Therefore, we start by defining the classical MANOVA model.

Definition 3.11 (MANOVA). Let $\boldsymbol{X}$ be a $p \times n$ random observation matrix. Assume that $\boldsymbol{B}$ is a $p \times m$ unknown parameter matrix and $\boldsymbol{C}$ an $m \times n$ design matrix such that $\mathrm{rank}(\boldsymbol{C}) + p \leq n$. Then the MANOVA model is defined as

$$\boldsymbol{X} = \boldsymbol{B}\boldsymbol{C} + \boldsymbol{E}, \tag{3.8}$$

where $\boldsymbol{E}$ is a $p \times n$ random errors matrix whose columns follow a multivariate normal distribution with mean $\mathbf{0}$ and p.d. covariance matrix $\boldsymbol{\Sigma}$, that is, $\boldsymbol{E} \sim \mathcal{N}_{p,n}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{I}_n)$.

The MLEs for the parameter matrices $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ are given by (if $\boldsymbol{C}$ is of full rank)

$$\widehat{\boldsymbol{B}} = \boldsymbol{X}\boldsymbol{C}^\intercal(\boldsymbol{C}\boldsymbol{C}^\intercal)^{-1}, \quad n\widehat{\boldsymbol{\Sigma}} = (\boldsymbol{X} - \widehat{\boldsymbol{B}}\boldsymbol{C})()^\intercal = \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{C}^\intercal(\boldsymbol{C}\boldsymbol{C}^\intercal)^{-1}\boldsymbol{C})\boldsymbol{X}^\intercal.$$

The Growth Curve model is an extension of (3.8) and can be used when dealing with repeated measurements and balanced data, that is, the dataset for instance consists of observations from the two populations all observed at the same time points. Its applications are often found in natural sciences, medicine, social sciences, etc.

Definition 3.12 (The Growth Curve model). Let $\boldsymbol{X}$ be a $p \times n$ observation matrix and $\boldsymbol{B}$ be the $q \times m$ unknown growth curve parameter matrix. Assume $\boldsymbol{A}$ is the $p \times q$ within-individuals design matrix and $\boldsymbol{C} : m \times n$ the between-individuals design matrix such that $\mathrm{rank}(\boldsymbol{C}) + p \leq n$. Then the growth curve model is stated as

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{B}\boldsymbol{C} + \boldsymbol{E}, \tag{3.9}$$

where $\boldsymbol{E} \sim \mathcal{N}_{p,n}(\boldsymbol{0}, \boldsymbol{\Sigma}, \boldsymbol{I}_n)$ and $\boldsymbol{\Sigma}$ is an unknown p.d. covariance matrix.

Assume the design matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ are of full rank. The MLEs of the unknown parameters $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ derived by Khatri (1966), are given by

$$\widehat{\boldsymbol{B}} = (\boldsymbol{A}^\intercal\boldsymbol{S}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^\intercal\boldsymbol{S}^{-1}\boldsymbol{X}\boldsymbol{C}^\intercal(\boldsymbol{C}\boldsymbol{C}^\intercal)^{-1}, \quad n\widehat{\boldsymbol{\Sigma}} = (\boldsymbol{X} - \boldsymbol{A}\widehat{\boldsymbol{B}}\boldsymbol{C})()^\intercal, \tag{3.10}$$

where $\boldsymbol{S}$ is the sum of squares matrix given by $\boldsymbol{S} = \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{C}^\intercal(\boldsymbol{C}\boldsymbol{C}^\intercal)^{-1}\boldsymbol{C})\boldsymbol{X}^\intercal$. Kollo and von Rosen (2005) and von Rosen (2018) provide many results for the Growth Curve model.

Example 3.5: Coronary sinus potassium data from Grizzle and Allen (1969)

The data in Table 3.1 consists of coronary sinus potassium measurements, from four groups of dogs. The first group contains 9 untreated dogs with coronary occlusion. The second group of 10 dogs was treated with extrinsic cardiac denervation three weeks ahead of coronary blockage whereas the third group contains 8 dogs given treated similarly to the second group but immediately before coronary occlusion. The final group of 9 dogs was bilateral thoracic sympathectomy and stellectomy was done three weeks before coronary occlusion. Assume these data follow the Growth Curve model (3.5). Suppose third-degree growth curves explain the growth profiles for the four groups. One could use time powers (t, t^2, t^3) but since the observations were taken at uniformly spaced time periods after occlusion, it is assumed that coefficients of orthogonal polynomial trend contrasts are more appropriate to use. There are $p = 7$ repeated measurements so that

$$\mu_i = b_{0i}a_{j0} + b_{1i}a_{j1} + b_{2i}a_{j2} + b_{3i}a_{j3}, \quad i \in \{1,2,3,4\}, \quad j \in \{1,2,3,4,5,6,7\},$$

where fixing j, a_{jk}, $k \in \{0,1,2,3\}$, is the k^{th} column of the within-individuals design matrix $\boldsymbol{A}$ and for fixed i, b_{ki} is the k^{th} row of the unknown parameter $\boldsymbol{B}$. Then, the design matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ equal

$$\boldsymbol{A} = \begin{pmatrix} 1 & -3 & 5 & -1 \\ 1 & -2 & 0 & 1 \\ 1 & -1 & -3 & 1 \\ 1 & 0 & -4 & 0 \\ 1 & 1 & -3 & -1 \\ 1 & 2 & 0 & -1 \\ 1 & 3 & 5 & 1 \end{pmatrix}, \; \boldsymbol{C} = \begin{pmatrix} \mathbf{1}_9^\intercal \otimes \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} : \mathbf{1}_{10}^\intercal \otimes \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} : \mathbf{1}_8^\intercal \otimes \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} : \mathbf{1}_9^\intercal \otimes \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix} \end{pmatrix}.$$

The MLEs of the unknown parameters are given by

$$\widehat{\boldsymbol{B}} = \begin{pmatrix} 4.693 & 3.575 & 4.021 & 3.835 \\ 0.151 & -0.027 & 0.136 & 0.076 \\ -0.059 & -0.020 & -0.031 & -0.002 \\ -0.185 & 0.001 & 0.062 & -0.031 \end{pmatrix}, \quad \widehat{\boldsymbol{\Sigma}} = \begin{pmatrix} 0.182 & 0.133 & \cdots & 0.168 \\ 0.133 & 0.130 & \cdots & 0.128 \\ \vdots & \vdots & \ddots & \vdots \\ 0.168 & 0.128 & \cdots & 0.465 \end{pmatrix}.$$

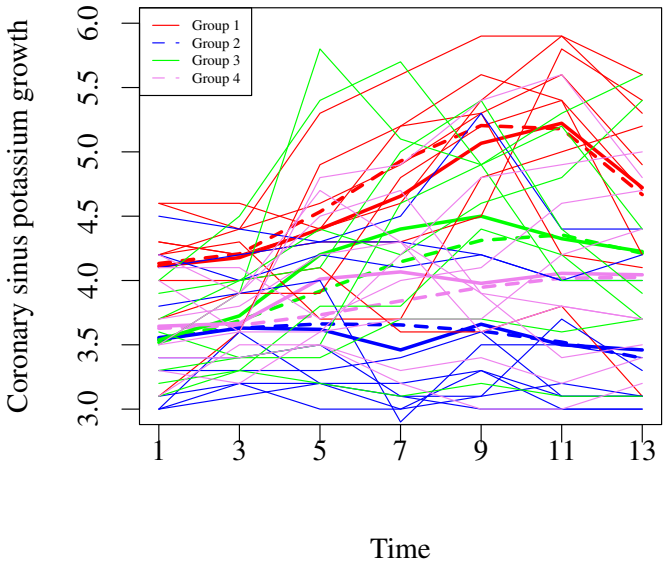


Figure 3.1: *Growth curves of coronary sinus potassium after occlusion from four types of treatments applied on 36 dogs. The dataset is presented in Table 3.1. Solid lines show group means and solid dashed lines show their estimates.*

Figure 3.1 shows the actual measurements of the coronary sinus potassium from the four different treatments groups of dogs 13 minutes after occlusions, plotted with their actual means (solid lines) against their estimated sample means (solid dashed lines) defined as $(j \in \{1, \ldots, 7\})$

$$\begin{aligned} \widehat{\mu}_1 &= 4.693 + 0.151a_{j1} - 0.059a_{j2} - 0.185a_{j3}, \\ \widehat{\mu}_2 &= 3.575 - 0.027a_{j1} - 0.020a_{j2} + 0.001a_{j3}, \\ \widehat{\mu}_3 &= 4.021 + 0.136a_{j1} - 0.031a_{j2} + 0.062a_{j3}, \\ \widehat{\mu}_4 &= 3.835 + 0.076a_{j1} - 0.002a_{j2} - 0.031a_{j3}. \end{aligned}$$

The coronary sinus potassium in the group of dogs without treatment as well as that in dogs treated with extrinsic denervation instantly prior to coronary occlusion seems to increase over time and starts decreasing at about 11 minutes after occlusion. On the other hand, the potassium constantly increases over time in dogs treated with bilateral thoracic sympactectomy and stellectomy 3 weeks prior to occlusion, whereas for the group that received extrinsic cardiac denervation 3 weeks prior to occlusion, it increases in the first minutes but starts decreasing over time.

Table 3.1: *Coronary sinus potassium measurements for dogs where the first group was untreated with coronary occlusion and the other 3 groups of dogs were provided different treatments before coronary occlusion.*

		Minutes after occlusion						
Group	Dog id	1	3	5	7	9	11	13
	1	4.0	4.0	4.1	3.6	3.6	3.8	3.1
	2	4.2	4.3	3.7	3.7	4.8	5.0	5.2
	3	4.3	4.2	4.3	4.3	4.5	5.8	5.4
	4	4.2	4.4	4.6	4.9	5.3	5.6	4.9
1	5	4.6	4.4	0.3	5.6	5.9	5.9	5.3
	6	3.1	3.6	4.9	5.2	5.3	4.2	4.1
	7	3.7	3.9	3.9	4.8	5.2	5.4	4.2
	8	4.3	4.2	4.4	5.2	5.6	5.4	4.7
	9	4.6	4.6	4.4	4.6	5.4	0.9	5.6
	10	3.4	3.4	3.5	3.1	3.1	3.7	3.3
	11	3.0	3.2	3.0	3.0	3.1	3.2	3.1
	12	3.0	3.1	3.2	3.0	3.3	3.0	3.0
	13	3.1	3.2	3.2	3.2	3.3	3.1	3.1
	14	3.8	3.9	4.0	2.9	3.5	3.5	3.4
2	15	3.0	3.6	3.2	3.1	3.0	3.0	3.0
	16	3.3	3.3	3.3	3.4	3.6	3.1	3.1
	17	4.2	4.0	4.2	4.1	4.2	4.0	4.0
	18	4.1	4.2	4.3	4.3	4.2	4.0	4.2
	19	4.5	4.4	4.3	4.5	5.3	4.4	4.4
	20	3.2	3.3	3.8	3.8	4.4	4.2	3.7
	21	3.3	3.4	3.4	3.7	3.7	3.6	3.7
	22	3.1	3.3	3.2	3.1	3.2	3.1	3.1
	23	3.6	3.4	3.5	4.6	4.9	5.2	4.4
3	24	4.0	4.5	5.4	5.7	4.9	4.0	4.0
	25	3.7	4.0	4.4	4.2	4.6	4.8	5.4
	26	3.5	3.9	0.8	5.1	4.9	5.3	5.6
	27	3.9	4.0	4.1	5.0	5.4	4.4	3.9
	28	3.1	3.5	3.5	3.2	3.0	3.0	3.2
	29	3.3	3.2	3.6	3.7	3.7	4.2	4.4
	30	3.5	3.9	4.7	4.3	3.9	3.4	3.5
	31	3.4	3.4	3.5	3.3	3.4	3.2	3.4
4	32	3.7	3.8	4.2	4.3	3.6	3.8	3.7
	33	4.0	3.6	4.8	4.9	5.4	5.6	4.8
	34	4.2	3.9	4.5	4.7	3.9	3.8	3.7
	35	4.1	4.1	3.7	4.0	4.1	4.6	4.7
	36	3.5	3.6	3.6	4.2	4.8	4.9	5.0

4

Discriminant analysis in relation to the main results of the thesis

DISCRIMINANT analysis and classification techniques belong to the well-known and classical multivariate statistical methods. The idea of discriminating between sets of populations and allocating new observations into predefined populations has been studied by many researchers since the late 1920s. This concept is widely applied in many areas such as psychology, biology, agriculture, medicine, and other areas for example pattern recognition and credit scoring. In this chapter, we provide some theoretical aspects, definitions, and some results on classification. We will mostly focus on classification via an underlying normal model including either known or unknown parameters.

4.1 Introduction

With the contribution of measurements such as Hotelling's T^2 statistic and the Mahalanobis distance, Δ^2, several authors since the early thirties have done important development to determine sets of values that best distinguish between populations. Since the early work by Fisher (1936), classification and discriminant analysis evolved either using a significance test or introducing a distance measure (Smith, 1936; Wallace and Travers, 1938; Wald, 1944; Rao, 1948; Anderson, 1951; Schaafsma and van Vark, 1979; Muirhead, 1982). The application of classification and discriminant analysis boosted with the introduction to the modern machine learning literature, especially for pattern recognition, which categorizes this method as supervised learning (McLachlan, 1992; Izenman, 2008; James et al., 2021; Matloff, 2017).

The term "discriminant analysis" is usually used to describe two main objectives: discrimination and classification. Discrimination is concerned with a set of rules for separating groups and classification consists of procedures for classifying observations into groups. In this thesis, groups are referred to as populations and we will consider the case of only

two populations, π_i, $i \in \{1, 2\}$, having a common covariance matrix $\boldsymbol{\Sigma}$. Assume that an observation is coming from either population π_1 or π_2. The classification methodology is based on the p-dimensional random vector $\boldsymbol{x} = (x_1, \ldots, x_p)^\intercal$ of measurements on that observation. The observation is allocated either to π_1 or to π_2. The aim is to classify into one of these populations by minimizing the error of misclassification.

A classification rule decides to which population the observation is allocated. In this thesis, the probability of classifying an observation $\boldsymbol{x}$ to population π_i is denoted $P(\boldsymbol{x} \to \pi_i)$ and the probability of wrongly allocating $\boldsymbol{x}$ to π_i when it belongs to π_j, $i \neq j$, is denoted $P(\boldsymbol{x} \to \pi_i | \boldsymbol{x} \in \pi_j)$. The latter probability is called a misclassification error or error rate. The more the sets of characteristics that best distinguish observations from populations π_i and π_j overlap the larger the misclassification error. Naturally, observations are more likely to be classified to a more frequent population, hence stronger evidence is required to allocate an observation to the population with a smaller prior probability. On the other hand, there exist other attributes of classification than the characteristics of observations; prior probability, and cost of misclassification. Naturally, observations are more likely to be classified to a larger population, hence stronger proof is required to allocate to the population with a smaller prior probability. Furthermore, the cost of classifying an unhealthy patient as healthy can be higher than classifying a healthy person as sick. Nevertheless, these classification aspects will not be considered in this thesis.

4.2 Classifiers

In discriminant analysis, we aim to classify an observation into predetermined populations or define a rule that separates the population. There exists several techniques to achieve this goal. In supervised learning classification, the most used approach is to identify an algorithm that predicts to which population an observation belongs based on labeled observations. See James et al. (2021) for alternative approaches.

In a general context, the discriminant rule is to determine a region $R_i \in \mathcal{R}^p$ that consists of a set of p characteristics of individuals from population π_i, $i \in \{1, 2, \ldots, q\}$, that allocates an observation $\boldsymbol{x}$ into one of the q different predefined populations with minimal risk for misclassification.

Assume an observation $\boldsymbol{x} = (x_1, \ldots, x_p)^\intercal$ is to be allocated to one of two known multivariate populations π_1 and π_2. Let $\boldsymbol{y} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) \in \pi_1$ and $\boldsymbol{z} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \in \pi_2$. The likelihood function based on $\boldsymbol{y}$, $\boldsymbol{z}$ and $\boldsymbol{x}$ is then given by

$$L_i(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{x} \in \pi_i) = (2\pi)^{-\frac{3}{2}p} |\boldsymbol{\Sigma}|^{-\frac{p}{2}} \exp\left\{ -\frac{1}{2} \left((\boldsymbol{y} - \boldsymbol{\mu}_1)^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_1) \right.\right.$$
$$\left.\left. + (\boldsymbol{z} - \boldsymbol{\mu}_2)^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_2) + (\boldsymbol{x} - \boldsymbol{\mu}_i)^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) \right) \right\}, \; i \in \{1, 2\}.$$

A likelihood ratio can be constructed which equals

$$\frac{L_1(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{x} \in \pi_1)}{L_2(\boldsymbol{y}, \boldsymbol{z} \mid \boldsymbol{x} \in \pi_2)} = e^{-\frac{1}{2}\left[(\boldsymbol{x} - \boldsymbol{\mu}_1)^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_1) - (\boldsymbol{x} - \boldsymbol{\mu}_2)^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_2) \right]}. \tag{4.1}$$

The observation $\boldsymbol{x}$ is said to belong to population π_1 if the ratio (4.1) is greater than or equal to 1 and otherwise it is allocated to π_2. Taking the logarithm of the ratio with known mean and covariance results in Fisher's linear discriminant function given below.

Example 4.1: Fisher (1936)'s LDA

Assume that the observation $\boldsymbol{x} = (x_1, \ldots, x_p)^\intercal$ is to be classified to either one of two normal populations with mean $\boldsymbol{\mu}_i$, $i \in \{1, 2\}$ and a common covariance matrix $\boldsymbol{\Sigma}$, $\mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $i \in \{1, 2\}$. Then Fisher's classifier is defined as

$$\begin{aligned} l(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) &= (\boldsymbol{\mu}_1 - \boldsymbol{x})^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{x}) - (\boldsymbol{\mu}_2 - \boldsymbol{x})^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{x}) \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\intercal \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2). \end{aligned} \tag{4.2}$$

The classification rule is to allocate $\boldsymbol{x}$ into population π_1 if $l(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) \geq 0$ and into π_2 if it is negative. Note that Fisher's LDA is normally distributed with mean $\pm \frac{1}{2}\Delta^2$, with "+" if $\boldsymbol{x}$ belongs to π_1 and "-" if $\boldsymbol{x}$ belongs to π_1 and a common variance of Δ^2, where

$$\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

is the squared Mahalanobis distance.

The population considered in discriminant problems can comprise known or unknown parameters. With unknown parameters, two other techniques for determining a classifier stand out; the plug-in and the likelihood approaches. The plug-in technique merely consists of replacing the unknown parameters in the discriminant rule with their estimators. Several researchers opted for this technique, including Anderson (1951) with his famous W-rule, which results in a linear function w.r.t. $\boldsymbol{x}$.

Example 4.2: Anderson (1951)'s W-rule via a plug-in approach

Assume we wish to classify the observation $\boldsymbol{x} = (x_1, \ldots, x_p)^\intercal$ in one of the two predetermined populations π_1 and π_2. Let $\boldsymbol{y}_i$, $i \in \{1, \ldots, n_1\}$, be a sample drawn from population π_1 in $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n_1})$ such that $\boldsymbol{y}_i \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, with $\overline{\boldsymbol{y}} = \frac{1}{n_1} \sum_{i=1}^{n_1} \boldsymbol{y}_i$ and $\boldsymbol{z}_j$, $j \in \{1, \ldots, n_2\}$, be collected from population π_2 in $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{n_2})$ such that $\boldsymbol{z}_j \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ with $\overline{\boldsymbol{z}} = \frac{1}{n_2} \sum_{j=1}^{n_2} \boldsymbol{z}_j$. The W- classification rule is given by

$$W = \boldsymbol{x}^\intercal \boldsymbol{S}^{-1} (\overline{\boldsymbol{y}} - \overline{\boldsymbol{z}}) - \frac{1}{2} (\overline{\boldsymbol{y}} + \overline{\boldsymbol{z}})^\intercal \boldsymbol{S}^{-1} (\overline{\boldsymbol{y}} - \overline{\boldsymbol{z}}), \tag{4.3}$$

where $\boldsymbol{S} \sim W_p(n, \boldsymbol{\Sigma})$ is the estimate of $\boldsymbol{\Sigma}$ based on the pooled sample. The rule is that $\boldsymbol{x}$ is allocated to π_1 if $W \geq 0$ and to π_2 otherwise.

4.2.1 A likelihood-based classifier

Consider the case where the two normal populations have an equal known covariance matrix but the mean is unknown, instead of (4.2) the classification problem results in a quadratic function of $\boldsymbol{x}$. The case was considered in Paper A.

Example 4.3: Classification rule with unknown μ and known Σ

Assume that an observation $\boldsymbol{x}$ is to be allocated to either π_1 or π_2. Let $\boldsymbol{y}_i \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $i \in \{1, \ldots, n_1\}$, be selected from π_1 and $\boldsymbol{z}_j \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, $j \in \{1, \ldots, n_2\}$ be selected from π_2 such that $\overline{\boldsymbol{y}} = \frac{1}{n_1}\sum_{i=1}^{n_1} \boldsymbol{y}_i$, and $\overline{\boldsymbol{z}} = \frac{1}{n_2}\sum_{j=1}^{n_2} \boldsymbol{z}_j$. Then a likelihood-based classifier is given by

$$D_1 = \frac{1}{2}\frac{n_2}{n_2+1}(\overline{\boldsymbol{z}} - \boldsymbol{x})^\intercal \boldsymbol{\Sigma}^{-1}(\overline{\boldsymbol{z}} - \boldsymbol{x}) - \frac{1}{2}\frac{n_1}{n_1+1}(\overline{\boldsymbol{y}} - \boldsymbol{x})^\intercal \boldsymbol{\Sigma}^{-1}(\overline{\boldsymbol{y}} - \boldsymbol{x}). \tag{4.4}$$

The rule states that $\boldsymbol{x}$ is classified to π_1 if $D_1 \geq 0$ and to π_2 otherwise (Umunoza Gasana et al., 2022b).

The classifier given by (4.4) in Example 4.3 provides a rational alternative to the existing known classifiers when two normal populations with a known common covariance are considered. It is important to note that the distribution of this classifier does not depend on the covariance since the distributions for $\boldsymbol{\Sigma}^{-\frac{1}{2}}(\overline{\boldsymbol{z}} - \boldsymbol{x})$ and $\boldsymbol{\Sigma}^{-\frac{1}{2}}(\overline{\boldsymbol{y}} - \boldsymbol{x})$ are independent of $\boldsymbol{\Sigma}$. Therefore, this classifier is suitable in this case because we supposed that the two populations have the same covariance, i.e. $\boldsymbol{\Sigma}$ should not be involved in allocating the observation $\boldsymbol{x}$ to either one of the two populations.

Moreover, if the classification problem regards two populations with unknown means and variance, a likelihood approach is often used to estimate the unknown parameters. See (Kudo, 1959, 1960). Since the sample covariance matrix follows a Wishart distribution, the expectation of the inverse of the covariance can be estimated using Theorem 3.3 (i) in this thesis. Let $\boldsymbol{x}_{ij} \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, $j \in \{1, \ldots, n_i\}$, $i \in \{1, 2\}$, be a sample from π_i, collected in $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})$, $i \in \{1, 2\}$. Then the following models emerge

$$\pi_i : \begin{cases} (\boldsymbol{X}_i : \boldsymbol{x}) &= \boldsymbol{\mu}_i \mathbf{1}_{n_i+1}^\intercal + \boldsymbol{\varepsilon}, & \boldsymbol{\varepsilon} \sim \mathcal{N}_{p,n_i+1}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{I}_{n_i+1}), \\ \boldsymbol{X}_k &= \boldsymbol{\mu}_k \mathbf{1}_{n_k}^\intercal + \boldsymbol{\varepsilon}, & \boldsymbol{\varepsilon} \sim \mathcal{N}_{p,n_k}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{I}_{n_k}), \end{cases} \tag{4.5}$$

where $i \in \{1, 2\}$, $k \in \{1, 2\}$, $i \neq k$. If $\boldsymbol{x} \in \pi_i$, $i \in \{1, 2\}$, the maximum likelihood estimators of the unknown parameters in (4.5) equal

$$\pi_i : \begin{cases} (\widehat{\boldsymbol{\mu}}_1, \widehat{\boldsymbol{\mu}}_2) &= \boldsymbol{X}\boldsymbol{C}_i^\intercal(\boldsymbol{C}_i\boldsymbol{C}_i^\intercal)^{-1}, & i \in \{1, 2\}, \\ (n+1)\widehat{\boldsymbol{\Sigma}}_i &= \boldsymbol{S}_i, & i \in \{1, 2\}, \end{cases} \tag{4.6}$$

where $\boldsymbol{X} = (\boldsymbol{X}_1 : \boldsymbol{x} : \boldsymbol{X}_2)$, $\boldsymbol{S}_i = \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{C}_i^\intercal(\boldsymbol{C}_i\boldsymbol{C}_i^\intercal)^{-1}\boldsymbol{C}_i)\boldsymbol{X}^\intercal$ are the sum of squares based on the sample drawn from populations π_i, $i \in \{1, 2\}$, and

$$\boldsymbol{C}_1 = \begin{pmatrix} \mathbf{1}_{n_1+1}^\intercal & \mathbf{0}_{n_2}^\intercal \\ \mathbf{0}_{n_1+1}^\intercal & \mathbf{1}_{n_2}^\intercal \end{pmatrix}, \quad \boldsymbol{C}_2 = \begin{pmatrix} \mathbf{1}_{n_1}^\intercal & \mathbf{0}_{n_2+1}^\intercal \\ \mathbf{0}_{n_1}^\intercal & \mathbf{1}_{n_2+1}^\intercal \end{pmatrix},$$

Details can be found in Paper A. Therefore, the likelihood ration for allocating $\boldsymbol{x}$ into π_1 or π_2 equal

$$\frac{(2\pi)^{-\frac{p}{2}(n+1)}|\widehat{\boldsymbol{\Sigma}}_1|^{-\frac{1}{2}(n+1)}\exp\left\{-\frac{1}{2}\mathrm{tr}\{\widehat{\boldsymbol{\Sigma}}_1^{-1}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_1)(\boldsymbol{x} - \boldsymbol{\mu}_1)^\intercal\}\right\}}{(2\pi)^{-\frac{p}{2}(n+1)}|\widehat{\boldsymbol{\Sigma}}_2|^{-\frac{1}{2}(n+1)}\exp\left\{-\frac{1}{2}\mathrm{tr}\{\widehat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{x} - \widehat{\boldsymbol{\mu}}_2)(\boldsymbol{x} - \boldsymbol{\mu}_2)^\intercal\}\right\}} = \left(\frac{\widehat{\boldsymbol{\Sigma}}_2}{\widehat{\boldsymbol{\Sigma}}_1}\right)^{\frac{n+1}{2}}, \tag{4.7}$$

where $n = n_1 + n_2$. The new observation $\boldsymbol{x}$ is allocated to population π_1 if the ratio (4.7) is greater than 1 and to π_2 if (4.7) is less than 1. Moreover, consider the sum of squares matrix $\boldsymbol{S} \sim W_p(\boldsymbol{\Sigma}, n_1 + n_2 - 2)$, based on the joint samples selected from populations π_i, $i \in \{1, 2\}$ without considering the fact that the new observation $\boldsymbol{x}$ is to be classified to any of the two populations. Then,

$$\boldsymbol{S}_i = \boldsymbol{S} + \frac{n_1}{n_1 + 1}(\overline{\boldsymbol{x}}_i - \boldsymbol{x})(\overline{\boldsymbol{x}}_i - \boldsymbol{x})^\intercal, \ i \in \{1, 2\}.$$

Therefore, for $n_1 + n_2 - 2 \geq p$, the likelihood ratio for allocating $\boldsymbol{x}$ into one of the two populations, when the unknown parameters are replaced with their estimators, is given by

$$\frac{|\boldsymbol{S}_2|}{|\boldsymbol{S}_1|} = \frac{1 + \frac{n_2}{n_2+1}(\overline{\boldsymbol{x}}_2 - \boldsymbol{x})^\intercal \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_2 - \boldsymbol{x})}{1 + \frac{n_1}{n_1+1}(\overline{\boldsymbol{x}}_1 - \boldsymbol{x})^\intercal \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_1 - \boldsymbol{x})}, \quad n_1 + n_2 - 2 \geq p. \tag{4.8}$$

Consequently, the new observation $\boldsymbol{x}$ belongs to π_1 when

$$\frac{n_2}{n_2 + 1}(\overline{\boldsymbol{x}}_2 - \boldsymbol{x})^\intercal \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_2 - \boldsymbol{x}) \geq \frac{n_1}{n_1 + 1}(\overline{\boldsymbol{x}}_1 - \boldsymbol{x})^\intercal \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}_1 - \boldsymbol{x}). \tag{4.9}$$

In Paper A, Section 3.2, a detailed MLE approach when both mean and covariance are unknown is given.

Example 4.4: Kudo (1959)'s Z-rule, with unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

Let $\boldsymbol{x}$, $\overline{\boldsymbol{y}}$ and $\overline{\boldsymbol{z}}$ be as in Example 4.3. The likelihood-based Z-rule is given by

$$Z = \frac{1}{2}\left[\frac{n_2}{n_2 + 1}(\boldsymbol{x} - \overline{\boldsymbol{z}})' \boldsymbol{S}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{z}}) - \frac{n_1}{n_1 + 1}(\boldsymbol{x} - \overline{\boldsymbol{y}})' \boldsymbol{S}^{-1}(\boldsymbol{x} - \overline{\boldsymbol{y}})\right]. \tag{4.10}$$

The rule is to classify $\boldsymbol{x}$ as coming from π_1 if $Z \geq 0$ and from π_2 otherwise. Note that in (4.9) above, $\overline{\boldsymbol{x}}_1$ is used instead of $\overline{\boldsymbol{y}}$ and $\overline{\boldsymbol{x}}_2$ instead of $\overline{\boldsymbol{z}}$.

The Z-rule given by (4.10) was provided by Kudo (1959) as an alternative to the Anderson (1951)'s W-rule. By Theorem 3.3 (i), $E[\boldsymbol{S}^{-1}] = \frac{1}{n_1+n_2-p-3}\boldsymbol{\Sigma}^{-1}$. Using this fact, Umunoza Gasana et al. (2022b) transformed the Z-rule by multiplying both sides by $\frac{1}{2}(n_1 + n_2 - p - 3)$ which results in the discriminant rule shown in the following example.

Example 4.5: An alternative classification rule with unknown $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

Consider $\boldsymbol{x}$, $\overline{\boldsymbol{y}}$ and $\overline{\boldsymbol{z}}$ defined as in Example 4.3.

$$D_2 = \frac{1}{2}(n_1+n_2-p-3)\left[\frac{n_2}{n_2 + 1}(\overline{\boldsymbol{z}} - \boldsymbol{x})^\intercal \boldsymbol{S}^{-1}(\overline{\boldsymbol{z}} - \boldsymbol{x}) - \frac{n_1}{n_1 + 1}(\overline{\boldsymbol{y}} - \boldsymbol{x})^\intercal \boldsymbol{S}^{-1}(\overline{\boldsymbol{y}} - \boldsymbol{x})\right]. \tag{4.11}$$

$\boldsymbol{x}$ is classified as coming from π_1 if $D_2 \geq 0$ and from π_2 if $D_2 < 0$.

The distributions of D_1 and D_2 stated in (4.4) and (4.11), respectively, proposed in Paper A are asymptotically equivalent and are alternatives to existing discriminant rules. Note that if $n_1 = n_2 = n$, $D_2 = -\frac{(2n-p-3)n}{n+1}W = (2n-p-3)Z$. Note that with $n = n_1 + n_2 - 2$, the inverse covariance, $\boldsymbol{S}^{-1}$, exists only if $n - p - 1 > 0$. Otherwise, a high-dimensional case has to be considered.

4.3 Classification of growth curves

In Section 4.2 we focused on results obtained from a study on data where several variables are collected at a single point in time. In this section, we look at a classification of growth curves including repeated measurement data. Classification of growth curves was first examined by Lee (1977, 1982) using Bayesian and non-Bayesian approaches. However, Rao (1958) briefly mentioned the classification of growth curves, and Burnaby (1966) considered the allocation of observations where the mean follows the Growth Curve model but regarded growth as a nuisance parameter to be excluded. Recently, Mentz and Kshirsagar (2005); Ngailo et al. (2021) classified growth curves but their approaches were not likelihood-based. von Rosen and Singull (2022) took into account the growth parameter using a likelihood technique. von Rosen (2018) provides many references about the Growth Curve model. In this thesis, we are interested in a likelihood approach for classifying observations with mean following the Potthoff and Roy (1964) model defined in Section 3.5.

Assume we wish to allocate an observation $\boldsymbol{x}$ with p repeated measurements into either population π_1 or π_2. Let $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ be random matrices of observations drawn from populations π_1 and π_2, respectively. Then the model can be written as

$$\pi_i: \ \boldsymbol{X}_i = \boldsymbol{A}\boldsymbol{\beta}_i\mathbf{1}_{n_i}^{\intercal} + \boldsymbol{\varepsilon}, \ \boldsymbol{\varepsilon} \sim \mathcal{N}_{p,n_i}(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{I}_{n_i}), \ i \in \{1, 2\},$$

where $\boldsymbol{\beta}_i$ and the dispersion matrix $\boldsymbol{\Sigma}$ are unknown parameters and the $p \times q$ matrix $\boldsymbol{A}$, $q < p$, contains the within-individual design matrix which expresses the growth curve structure. Note that there n_i columns in ε are independent and $\boldsymbol{\Sigma}$ is positive definite.

As mentioned in the introduction, most of the classification textbooks do not include the possibility that a new observation belongs to a population with a mean structure different from the two predetermined populations. Our focus in this thesis is on a likelihood-ratio-based discriminant rule derived by von Rosen and Singull (2022) which constitutes a two-step criterion that considers the prospect of a new observation coming from an unknown population, showcased in the following example.

Example 4.6: A two-step criterion for allocating growth curves

Assume $\boldsymbol{x}$, $\overline{\boldsymbol{y}}$ and $\overline{\boldsymbol{z}}$ are defined as in Example 4.3. von Rosen and Singull (2022) propose the two-step classification given below

$$D_{31} = \frac{n_2}{n_2+1}(\bar{\boldsymbol{z}} - \boldsymbol{x})^{\intercal}\boldsymbol{S}^{-1}(\bar{\boldsymbol{z}} - \boldsymbol{x}) - \frac{n_1}{n_1+1}(\bar{\boldsymbol{y}} - \boldsymbol{x})^{\intercal}\boldsymbol{S}^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{x}) \tag{4.12}$$

$$D_{32} = c(\boldsymbol{A}\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}}\boldsymbol{x})^{\intercal}\boldsymbol{S}^{-1}(\boldsymbol{A}\widehat{\boldsymbol{\beta}}_2 - \boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}}\boldsymbol{x}) - (\boldsymbol{A}\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}}\boldsymbol{x})^{\intercal}\boldsymbol{S}^{-1}(\boldsymbol{A}\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{P}_{\boldsymbol{A},\boldsymbol{S}}\boldsymbol{x}), \tag{4.13}$$

where

$$c = \frac{n_2(n_1+1)}{n_1(n_2+1)} \frac{1 + \frac{n_1}{n_1+1}(\bar{\boldsymbol{y}} - \boldsymbol{x})^{\intercal}\boldsymbol{S}^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{x})}{1 + \frac{n_2}{n_2+1}(\bar{\boldsymbol{z}} - \boldsymbol{x})^{\intercal}\boldsymbol{S}^{-1}(\bar{\boldsymbol{z}} - \boldsymbol{x})}, \tag{4.14}$$

and the sum of squares matrix $\boldsymbol{S} \sim W_p(\boldsymbol{\Sigma}, n-2)$, $n = n_1 + n_2$. The rule is to classify the observation $\boldsymbol{x}$ into population π_1 if both D_{31} and D_{32} are positive and into π_2 if both D_{31} and D_{32} are negative. Moreover, if only one of them is positive, that is D_{31} and D_{32} have opposite signs, then the new observation $\boldsymbol{x}$ is to be allocated to some unknown population.

One would logically assume that the within-group variance is fairly small. Consequently, in this thesis, c given by (4.14) is assumed to be fixed. The above two-step classifier is examined in Paper C where we determined the first two moments for D_{31} and D_{32} which were used in Paper D to compute misclassification errors via an Edgeworth-type expansion. A standard likelihood-based classifier for growth curves would rely on the allocation of the new observation based on whether $D_{32} \geq 0$ or $D_{32} < 0$. Therefore, the two-step setting produces more information. On the other hand, deriving error rates in this setting is quite challenging, and hence, an extra process is needed to approximate misclassification errors. In Paper D, we examine the error rates for the two-step criteria given in Example 4.6 separately. Though it is not obvious, misclassification errors through a joint distribution via Edgeworth-type expansion of the two criteria can be approximated using the Bonferroni inequality.

4.4 Misclassification errors via an Edgeworth-type expansion

Classifiers alone are not sufficient to decide about a decision rule. In statistics, we are interested in estimating probabilities that a discriminant rule wrongly classifies a new observation. Fujikoshi et al. (2011) defines the error rate as a measure of goodness of the given classifier.

Let $f_i(\boldsymbol{x})$ be the density function when the observation $\boldsymbol{x}$ is drawn from population π_i, $i \in \{1, 2\}$. The probability $P(\boldsymbol{x} \to \pi_2 | \boldsymbol{x} \in \pi_1)$ of allocating a new observation $\boldsymbol{x}$ as coming from π_2 whereas it belongs π_1 via the classifier D is given by

$$P(\boldsymbol{x} \to \pi_2 | \boldsymbol{x} \in \pi_1) = P(D \leq 0 | \boldsymbol{x} \in \pi_1) = \int_{R_2} f_1(\boldsymbol{x}) d\boldsymbol{x}, \tag{4.15}$$

where R_2 is the region composed by a set of characteristics of individuals that classifies $\boldsymbol{x}$ into π_2. The probability of classifying $\boldsymbol{x}$ as belonging to π_1 when it comes from π_2 equals

$$P(\boldsymbol{x} \to \pi_1 | \boldsymbol{x} \in \pi_2) = P(D > 0 | \boldsymbol{x} \in \pi_2) = \int_{R_1} f_2(\boldsymbol{x}) d\boldsymbol{x}, \tag{4.16}$$

where R_1 is the region comprising a set of values of $\boldsymbol{x}$ for which individuals are allocated to π_1. Probabilities of misclassification under normality assumption are usually monotone functions of the Mahalanobis distance Δ.

Example 4.7: Misclassification errors when applying the Fisher's LDA

Consider Fisher's linear classification function, given by (4.2). Let $\Phi(\,\cdot\,)$ be the standard

normal cumulative distribution function. Then, it is well known that

$$P(\boldsymbol{x} \to \pi_2 | \boldsymbol{x} \in \pi_1) = P(l \leq 0 | \boldsymbol{x} \in \pi_1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\frac{1}{2}\Delta} e^{-\frac{1}{2}\boldsymbol{x}^2} d\boldsymbol{x} = \Phi(-\frac{1}{2}\Delta), \quad (4.17)$$

$$P(\boldsymbol{x} \to \pi_1 | \boldsymbol{x} \in \pi_2) = P(l > 0 | \boldsymbol{x} \in \pi_1) = \frac{1}{\sqrt{2\pi}} \int_{\frac{1}{2}\Delta}^{\infty} e^{-\frac{1}{2}\boldsymbol{x}^2} d\boldsymbol{x} = 1 - \Phi(\frac{1}{2}\Delta). \quad (4.18)$$

The distribution of classifiers, such as the ones mentioned in Section 4.2 and Section 4.3, is often too complicated to allow for exact computations. Asymptotic expansions are therefore very common when it comes to estimating probabilities of misclassifications. Okamoto (1963) evaluated an asymptotic expansion for the distribution of the W linear discriminant rule. The Edgeworth-type expansion given by (3.7) in Section 3.4 is used in this thesis to approximate misclassification errors and has not been used before for such a matter (Umunoza Gasana et al., 2022a, 2023a).

Theorem 4.1
Let $\boldsymbol{\phi}(\cdot)$ and $\Phi(\cdot)$ be the density of $\mathcal{N}(0,1)$ and the cumulative distribution function, respectively. Consider the Edgeworth-type approximation of the form (3.7)*. Then misclassification errors about classifier D are approximated as*

$$P(\boldsymbol{x} \to \pi_2 | \boldsymbol{x} \in \pi_1) \approx \Phi(-\Delta) + \frac{1}{2}\boldsymbol{\phi}(\Delta)\left(c_1 + \Delta c_2 - (\Delta^2 + 2)c_3\right), \quad (4.19)$$

where Δ^2 is the squared Mahalanobis distance given by

$$\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\intercal \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \quad (4.20)$$

and

$$\begin{aligned} c_1 &= 5E[D] - 3Var[D]E[D] - (E[D])^3, \\ c_2 &= Var[D] - 1 + (E[D])^2, \\ c_3 &= (Var[D] - 1)E[D] + \frac{1}{3}(E[D])^3. \end{aligned} \quad (4.21)$$

The probability $P(\boldsymbol{x} \to \pi_1 | \boldsymbol{x} \in \pi_2)$ is the same as in (4.19) *where,*

$$\begin{aligned} c_1 &= -5E[D] + 3Var[D]E[D] + (E[D])^3, \\ c_2 &= Var[D] - 1 + (E[D])^2, \\ c_3 &= -(Var[D] - 1)E[D] - \frac{1}{3}(E[D])^3. \end{aligned} \quad (4.22)$$

The proof of Theorem 4.1 can be found in Paper B (Umunoza Gasana et al., 2022a).

Edgeworth-type expansions defined in Section 3.4 were used in Paper B to estimate error rates by the classifiers D_1 given by (4.4) and D_2 given by (4.11) using moments computed in Paper A. Similarly, the aforementioned Edgeworth-type expansions are useful for the case of repeated measurement data and were applied in Paper D using moments evaluated in Paper C.

5

Concluding observations

5.1 Summary of contributions

In this thesis, quadratic discriminant analysis has been examined using a likelihood approach when discriminating between two multivariate normal populations constituting single time points and repeated measurement data, namely growth curves. New methods have been developed mainly focusing on developing new likelihood-based classifiers and deriving the approximations of their probabilities of misclassifications. Principal contributions to achieving the objectives of the current thesis are:

In Paper A, two asymptotically equivalent discriminant rules have been derived using a likelihood process, for the classification into two normally distributed populations with distinct means and a common covariance matrix, considering two cases; when the same covariance is known and when it is unknown. The distribution of the classifier when $\mathbf{\Sigma}$ is known is similar to the difference between two non-central χ^2-distribution whereas the classifier, when $\mathbf{\Sigma}$, is an unknown has the same distribution as the difference between two non-central F-distributions. In addition, in order to assess the basic properties of the two classifiers with complicated distributions, the first two cumulants of the proposed classifiers have been calculated. These two classifiers turn out to have smaller variances than the widely known W-rule. Paper A fulfills the first objective.

In Paper B, the second objective was met by approximating misclassification errors for the two classifiers developed in Paper A. Since both of the two discriminant rules presented a too complicated distribution to allow for numerical calculations, an Edgeworth-type expansion of the classifiers has been applied to approximate their distributions. Most of the existing literature has considered an asymptotic expansion for the distribution of the dscriminant rule. The importance of an Edgeworth-type

expansion lies in the fact that asymptotic normality is widely applied in statistics. However, one disadvantage of the Edgeworth-type expansion worth mentioning is that it is not always a density, which may lead to negative probabilities of misclassifications. On the other hand, a negative value can be interpreted as a small error rate.

Paper A and Paper B focused on single time-point data. In Paper C we examined a two-step criteria for classification of growth curves, developed by von Rosen and Singull (2022). The main importance of the two-step classifier is that it takes into consideration the possibility that the new observation to be allocated might not belong to any of the two predetermined populations. Furthermore, the first two moments of each of the two conditions from the two-step classification rule were computed, in order to meet the third objective.

To achieve the fourth objective, in Paper D we extended the results of Paper C by evaluating misclassification errors through an Edgeworth-type expansion using the two moments for the two-step classifier computed in Paper C. Since the classifier constitutes two criteria, the distribution of each criterion is expanded separately and error rates of the two-step classifier were approximated.

5.2 Future research

During the course of this thesis, several related research problems arose. In all four papers presented in this thesis, we considered the discriminant analysis when only two multivariate normal populations are involved. This research can be extended by generalizing the cases of classification into one of several populations. This generalization can be possible for instance using pairwise conjectures that consist of a collection of all allocations of the new observation to any of each pair of populations.

In addition, the distribution properties of the likelihood-based quadratic discriminant rules derived in Paper A are developed under the assumption that the two populations are multivariate normally distributed with an equal covariance matrix and different means. This paper can be extended by deriving new results for the case of classifying into two populations with unequal covariance matrices. Anderson and Bahadur (1962) constitute a potential starting point where they examined linear procedures for classification under this assumption.

In Paper B and Paper D, an Edgeworth-type expansion was applied in order to approximate the distribution of the examined classifiers. As mentioned, an Edgeworth-type expansion has a weak spot in the fact that the expansion might not be a density which may lead to irrelevant results. Other distribution approximations such as Cornish-Fisher expansion can provide good alternative results.

In Paper C we studied the distribution characteristics of the two-step classifier of growth curves (see Example 4.6) that constitute of two conditions required to allocate a new observation to either one of the two predefined populations or an unknown population. In Paper D we approximated the misclassification errors for the classifier examined in Paper C. In both papers, we assumed that the within-group variance is relatively small to allow c

given by (4.14) to be fixed. The results in these two papers can be extended by estimating probabilities of misclassification for the two-step classifier when c in (4.14) is not fixed.

In addition, in Paper D, separate misclassification errors were approximated for the two criteria that constitute a two-step classifier. The joint distribution of the two criteria can be approximated using the Bonferroni inequality and applied to express the combined error rates of the classifier.

In all four papers enclosed in this thesis, we assumed that the number of observations, n, is larger than the number of observed variables, p. Since the covariance matrix is not invertible in a high-dimensional case, the discriminant rules proposed in this current thesis would fail. Therefore, other classifiers must be considered. Some modified LDA already exists in the literature using diagonalization and regularization of the covariance matrix or applying the Moore-Penrose inverse. One can develop a new approach to discriminant analysis that allows the approximation of misclassification errors when $p > n$ in a QDA.

At last, the results developed throughout this thesis can be applied to real-life data to convey their worthiness.

Bibliography

[1] Anderson, T. W. (1951). Classification by multivariate analysis. *Psychometrika*, 16(1):31–50.

[2] Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. John Wiley & Sons, Inc., Hoboken, third edition.

[3] Anderson, T. W. and Bahadur, R. R. (1962). Classification into two multivariate normal distributions with different covariance matrices. *The Annals of Mathematical Statistics*, 33(2):420–431.

[4] Barnard, M. M. (1935). The secular variations of skull characters in four series of egyptian skulls. *Annals of Eugenics*, 6(4):352–371.

[5] Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic Techniques for Use in Statistics*, volume 11. Springer, New York.

[6] Bose, R. C. (1936). On the exact distribution and moment-coefficients of the D^2-statistic. *Sankhyā: The Indian Journal of Statistics*, 2(2):143–154.

[7] Bowker, A. H. and Sitgreaves, R. (1961). An asymptotic expansion for the distribution of the W-classification statistic. *Chapter 19 in Solomon, H. (ed.), Studies in Item Analysis and Prediction*, pages 285–310.

[8] Burnaby, T. P. (1966). Growth-invariant discriminant functions and generalized distances. *Biometrics*, 22(1):96–110.

[9] Cornish, E. A. and Fisher, R. A. (1938). Moments and cumulants in the specification of distributions. *Revue de l'Institut International de Statistique*, 5(4):307–320.

[10] Davis, A. (1976). Statistical distributions in univariate and multivariate Edgeworth populations. *Biometrika*, 63:661–670.

[11] Edgeworth, F. (1907). On the representation of statistical frequency by a series. *Journal of the Royal Statistical Society*, 70(1):102–106.

[12] Fisher, R. A. (1936). The use of multiple measurements in 10 taxonomic problems. *Annals of Eugenics*, 7(2):179–188.

[13] Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Annals of eugenics*, 8(4):376–386.

[14] Fujikoshi, Y. (1987). Error bounds for asymptotic expansions of the distribution of the mle in a gmanova model. *Annals of the Institute of Statistical Mathematics*, 39(1):153–161.

[15] Fujikoshi, Y., Ulyanov, V. V., and Shimizu, R. (2011). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*, volume 760. John Wiley & Sons, Inc., Hoboken.

[16] Gleser, L. J. and Olkin, I. (1970). Linear models in multivariate analysis. *Essays in Probability and Statistics, (R.C. Bose et al., eds.)*, pages 267–292.

[17] Gram, J. P. (1879). *Om Rækkeudviklinger, bestemte ved Hjælp af de mindste Kvadraters Methode*. PhD thesis, Andr. Fred. Høst & Son, København.

[18] Grizzle, J. E. and Allen, D. M. (1969). Analysis of growth and dose response curves. *Biometrics*, 25(2):357–381.

[19] Gupta, S. S. and Panchapakesan, S. (1982). Edgeworth expansions in statistics: A brief review. *Technical report, Purdue university Lafayette in department of statistics*.

[20] Hald, A. (2000). The early history of the cumulants and the Gram-Charlier series. *International Statistical Review*, 68(2):137–153.

[21] Hald, A. and Steffensen, J. (2002). *On the History of Series Expansions of Frequency Functions and Sampling Distributions, 1873-1944*. Det Kongelige Danske Videnskabernes Selskab, Copenhagen.

[22] Hodges, J. L. (1955). *Discriminatory Analysis: Survey of Discriminatory Analysis*. Air University, School of aviation medicine, USAF.

[23] Hotelling, H. (1931). The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3):360 – 378.

[24] Huberty, C. J. (1975). Discriminant analysis. *Review of Educational Research*, 45(4):543–598.

[25] Huberty, C. J. and Olejnik, S. (2006). *Applied MANOVA and discriminant analysis*. John Wiley & Sons, New York.

[26] Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning*, volume 978. Springer Science & Business Media, New York.

[27] James, G., Witten, D., Hastie, T., and Robert, T. (2021). *An Introduction to Statistical Learning with Applications in R*. Springer, New York.

[28] Kendall, M. (1957). A course in multivariate analysis, griffin's statistical monographs & courses. *Hafner Publishing Co., New York*, 2(9):5.

[29] Khatri, C. G. (1966). A note on a MANOVA model applied to problems in growth curve. *Annals of the Institute of Statistical Mathematics*, 18(1):75–86.

[30] Kollo, T., Roos, A., and von Rosen, D. (2007). Approximation of the distribution of the location parameter in the growth curve model. *Scandinavian Journal of Statistics*, 34(3):499–510.

[31] Kollo, T. and von Rosen, D. (1998). A unified approach to the approximation of multivariate densities. *Scandinavian journal of statistics*, 25(1):93–109.

[32] Kollo, T. and von Rosen, D. (2005). *Advanced Multivariate Statistics with Matrices*, volume 579. Springer Science & Business Media, Dordrecht, The Netherlands.

[33] Kudo, A. (1959). The classificatory problem viewed as a two-decision problem. *Memoirs of the Faculty of Science, Kyushu University. Series A, Mathematics*, 13(2):96–125.

[34] Kudo, A. (1960). The classifactory problem viewed as a two-decision problem-II. *Memoirs of the Faculty of Science, Kyushu University. Series A, Mathematics*, 14(1):63–83.

[35] Laplace, P.-S. (1811). Mémoire sur les intégrales définies et leur application aux probabilités, et spécialement à la recherche du milieu qu'il faut choisir entre les résultats des observations. *Mémoires de l'Acadêmie Royale des Sciences de Paris*, pages 279–347.

[36] Lauritzen, S. L., Thiele, T. N., Thiele, J., and Hald, A. (2002). *Thiele: Pioneer in Statistics*. Oxford University Press Inc., New York.

[37] Lee, J. C. (1977). Bayesian classification of data from growth curves. *South African Statistical Journal*, 11(2):155–166.

[38] Lee, J. C. (1982). Classification of growth curves. In *Classification Pattern Recognition and Reduction of Dimensionality*, volume 2 of *Handbook of Statistics*, pages 121–137. Elsevier, North-Holland, Amsterdam.

[39] Mahalanobis, P. C. (1925). Analysis of race-mixture in Bengal. *Journal and Proceedings of Asiatic Society of Bengal New series*, 23:301–333.

[40] Mahalanobis, P. C. (1930). On test and measures of group divergence: theoretical formulae. *Journal and Proceedings of Asiatic Society of Bengal New series*, 26:541–588.

[41] Mahalanobis, P. C. (1936). On the generalized distance in statistics. National Institute of Science of India.

[42] Mathai, A. and Provost, S. B. (1992). *Quadratic Forms in Random Variables: Theory and Applications*, volume 87. Marcel Dekker Inc., New York.

[43] Matloff, N. (2017). *Statistical Regression and Classification: From Linear Models to Machine Learning*. CRC Press, New York.

[44] McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, New York.

[45] McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, Inc., Hoboken, New Jersey.

[46] Memon, A. Z. and Okamoto, M. (1971). Asymptotic expansion of the distribution of the Z statistic in discriminant analysis. *Journal of Multivariate Analysis*, 1(3):294–307.

[47] Mentz, G. B. and Kshirsagar, A. M. (2005). Classification using growth curves. *Communications in Statistics - Theory and Methods*, 33(10):2487–2502.

[48] Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. John Wiley & Sons, Inc., Hoboken, New Jersey.

[49] Ngailo, E. K., von Rosen, D., and Singull, M. (2021). Asymptotic approximation of misclassification probabilities in linear discriminant analysis with repeated measurements. *Acta et Commentationes Universitatis Tartuensis de Mathematica*, 25(1):67–85.

[50] Okamoto, M. (1963). An asymptotic expansion for the distribution of the linear discriminant function. *The Annals of Mathematical Statistics*, 34(4):1286–1301.

[51] Pearson, K. (1911). On the probability that two independent distributions of frequency are really samples from the same population. *Biometrika*, 8(1):250–254.

[52] Pearson, K. (1915). On the problem of sexing osteometric material. *Biometrika*, 10(4):479–487.

[53] Pearson, K. (1926). On the coefficient of racial likeness. *Biometrika*, 18:105–117.

[54] Potthoff, R. F. and Roy, S. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51(3-4):313–326.

[55] Ranga Rao, R. (1960). *Some Problems in Probability Theory*. PhD thesis, D. Phil., Thesis, Calcutta University.

[56] Rao, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society: Series B (Methodological)*, 10(2):159–193.

[57] Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics*, 14(1):1–17.

[58] Rao, C. R. (1959). Some problems involving linear hypotheses in multivariate analysis. *Biometrika*, 46:49–58.

[59] Rao, C. R. (1965). The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. *Biometrika*, 52(3-4):447–458.

[60] Rao, C. R. (1966). Discriminant function between composite hypotheses and related problems. *Biometrika*, 53:339–345.

[61] Schaafsma, W. and van Vark, G. N. (1979). Classification and discrimination problems with applications, part II. *Statistica Neerlandica*, 33(2):91–126.

[62] Siotani, M. (1977). *Asymptotic Expansions for Error Rates and Comparison of the W-procedure and the Z-procedure in Discriminant Analysis.* North-Holland, Amsterdam.

[63] Siotani, M. (1982). Large sample approximations and asymptotic expansions of classification statistics. In *Classification Pattern Recognition and Reduction of Dimensionality*, volume 2 of *Handbook of Statistics*, pages 61–100. Elsevier, North-Holland, Amsterdam.

[64] Siotani, M. and Wang, R.-H. (1975). *Further Expansion Formulae for Error Rates and Comparison of the W-and Z-procedures in Discriminant Analysis.* Department of Statistics, Kansas State University, Manhattan, Kansas.

[65] Sitgreaves, R. (1952). On the distribution of two random matrices used in classification procedures. *The Annals of Mathematical Statistics*, 23(2):263–270.

[66] Sitgreaves, R. (1961). Some results on the distribution of the W-classification statistic. *Chapter 15 in Solomon, H. (ed.). Studies in Item Analysis and Prediction*, pages 241–261.

[67] Skovgaard, I. M. (1986). On the multivariate Edgeworth expansions. *International Statistical Review*, 54(2):169–186.

[68] Smith, H. F. (1936). A discriminant function for plant selection. *Annals of Eugenics*, 7(3):240–250.

[69] Srivastava, M. S. and Khatri, C. (1979). *An Introduction to Multivariate Statistics.* Elsevier North-Holland, New York.

[70] Srivastava, M. S. and von Rosen, D. (1999). Growth curve models. In *Multivariate analysis, design of experiments, and survey sampling. Statistics: A series of textbooks and monographs*, volume 159, pages 547–578. Dekker, New York.

[71] Thiele, T. N. (1873). Om en tilnærmelsesformel. *Tidsskrift for mathematik*, 3:22–31.

[72] Thiele, T. N. (1889). *Forlaesinger over Almindelig Iagttagelseslaere: Sandsynlighedsregning og Mindste Kvadraters Methode.* København, Reitzel.

[73] Umunoza Gasana, E., von Rosen, D., and Singull, M. (2022a). Approximated misclassification errors for the likelihood based discriminant function via edgetworth-type expansion. *Linköping University Electronic Press, LiTH-MAT-R–2021/08–SE.*

[74] Umunoza Gasana, E., von Rosen, D., and Singull, M. (2022b). Moments of the likelihood-based discriminant function. *Communications in Statistics - Theory and Methods*, pages 1–13.

[75] Umunoza Gasana, E., von Rosen, D., and Singull, M. (2023a). Edgeworth-type expansion of the density of the classifier when growth curves are classified via the likelihood. *Linköping University Electronic Press, LiTH-MAT-R-2023/02.*

[76] Umunoza Gasana, E., von Rosen, D., and Singull, M. (2023b). Moments of the likelihood-based classification function using growth curves. *Linköping University Electronic Press, LiTH-MAT-R–2023/01–SE*.

[77] von Rosen, D. (1988). Moments for the inverted Wishart distribution. *Scandinavian Journal of Statistics*, 15(2):97–109.

[78] von Rosen, D. (1991). The growth curve model: A review. *Communications in Statistics-Theory and Methods*, 20(9):2791–2822.

[79] von Rosen, D. (2018). *Bilinear Regression Analysis: An Introduction*, volume 220. Springer, New York.

[80] von Rosen, D. and Singull, M. (2022). *Classification of repeated measurements using growth curves*. Linköping University Electronic Press, LiTH-MAT-R–2021/01–SE.

[81] Wald, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *The Annals of Mathematical Statistics*, 15(2):145–162.

[82] Wallace, N. and Travers, R. M. (1938). A psychometric sociological study of a group of speciality salesmen. *Annals of Eugenics*, 8(3):266–302.

[83] Wishart, J. (1928). The generalised product moment distribution in samples from a normal multivariate population. *Biometrika*, 20:32–52.

[84] Wishart, J. (1938). Growth-rate determinations in nutrition studies with the bacon pig, and their analysis. *Biometrika*, 30:16–28.

[85] Woolson, R. F. and Leeper, J. D. (1980). Growth curve analysis of complete and incomplete longitudinal data. *Communications in Statistics - Theory and Methods*, 9(14):1491–1513.

Part II

Papers

Papers

The papers associated with this thesis have been removed for copyright reasons. For more details about these see:

https://doi.org/10.3384/9789180751537

FACULTY OF SCIENCE AND ENGINEERING

Linköping Studies in Science and Technology, Dissertation No. 2311, 2023
Department of Mathematics

Linköping University
SE-581 83 Linköping, Sweden

www.liu.se

LINKÖPING UNIVERSITY