Misclassification Probabilities through Edgeworth-type Expansion for the Distribution of the Maximum Likelihood based Discriminant Function

Emelyne Umunoza Gasana

Linköping Studies in Science and Technology.
Licentiate Thesis No. 1911

Misclassification Probabilities through Edgeworth-type Expansion for the Distribution of the Maximum Likelihood based Discriminant Function

Emelyne Umunoza Gasana

Department of Mathematics
Division of Mathematical Statistics
Linköping University
SE-581 83 Linköping, Sweden
Linköping 2021

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License.

<https://creativecommons.org/licenses/by-nc/4.0/>

Linköping Studies in Science and Technology.

Licentiate Thesis No. 1911

Misclassification Probabilities through Edgeworth-type Expansion for the Distribution of the Maximum Likelihood based Discriminant Function

Emelyne Umunoza Gasana

emelyne.umunoza.gasana@liu.se

www.mai.liu.se

Mathematical Statistics

Department of Mathematics

Linköping University

SE-581 83 Linköping, Sweden

ISBN 978-91-7929-619-3

ISSN 0280-7971

Copyright © 2021 Emelyne Umunoza Gasana

Printed by LiU-Tryck, Linköping, Sweden 2021

*To the memory of my father, Gasana, and to fulfilling my promise to him and to
Mukabirasa, Juru and Iriza*

Abstract

Emelyne Umunoza Gasana (2021). Misclassification Probabilities through Edgeworth-type Expansion for the Distribution of the Maximum Likelihood based Discriminant Function.

Doctoral licentiate. ISBN 978-91-7929-619-3. ISSN 0280-7971.

This thesis covers misclassification probabilities via an Edgeworth-type expansion of the maximum likelihood based discriminant function. When deriving misclassification errors, first the expectation and variance in the population are assumed to be known where the variance is the same across populations and thereafter we consider the case where those parameters are unknown. Cumulants of the discriminant function for discriminating between two multivariate normal populations are derived. Approximate probabilities of the misclassification errors are established via an Edgeworth-type expansion using a standard normal distribution.

Keywords

Cumulants of discriminant function, Edgeworth-type expansion, linear discriminant analysis, misclassification errors.

Populärvetenskaplig sammanfattning

Statistisk klassificering är en multivariat teknik som används för att bestämma tillhörigheten av nya observationer till redan definierade grupper (så kallade klasser). Vid klassificering är det viktigt att studera sannolikheterna för felklassificering. En berömd metod för klassificering är att använda Fishers linjära diskriminantfunktion och när alla parametrar är kända så kan man enkelt beräkna sannolikheterna för felklassificering. Däremot, när parametrar är okända så måste man approximera dessa sannolikheter och detta kan göras på olika sätt.

I den här avhandlingen studeras en likelihoodbaserad diskriminantfunktion. Genom att härleda de första två momenten för diskriminantfunktionen kan sedan dess fördelningen approximeras genom en Edgeworth-utvidgning och sannolikheterna för felklassificering kan därigenom uppskattas. För att förstå skillnaden mellan metoder jämförs sannolikheterna för felklassificering för olika diskriminantfunktioner.

Acknowledgments

My acknowledgements to the Almighty GOD, from whom all knowledge, wisdom and understanding rightly come, He who has guided me throughout the whole period of my studies, for all that He has so graciously allowed me to achieve.

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

My sincere acknowledgements are addressed to the Linköping University (LiU) for its appreciable policy of promoting education at all levels, to Sida bilateral program for their financial support that allows me to pursue my PhD degree studies fruitfully. To Theresa Lagali, Karin Johansson and the entire administrative staff of Mathematics department at LiU for their support, and the whole academic staff, to all who generously gave their time and helped me to sharpen my knowledge during this period I have been training in LiU and contributed significantly to the realization of this step in my studies.

To my friend and colleague Denise for her moral support and the fruitful exchanges we have, to Cigdem, Roghi, Vincent, Mainza, Edward, Stefane, Elias and everyone, who contribute to these studies in terms of prayers, ideas, technical knowledge and support, I would like to extend my deepest and most heartfelt gratitude.

Last, but not least, to all my family and friends, especially to my beloved late father to whom I promised that I will do a PhD, my wonderful mother, there are no profound words to express my gratitude for the love and support that you have given me since what seems like forever, 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 and continued support is the light at the end of the tunnel. It is therefore for lack of a better word that I only say thank you.

Linköping, May 18, 2021
Emelyne Umunoza Gasana

Contents

1	Introduction	1
1.1	Background	1
1.2	Motivation	3
1.3	Aims and outline	3
1.3.1	Objectives of the thesis	3
1.3.2	Thesis outline	4
2	First and Second Cumulants of Discriminant Function	5
2.1	First and Second Cumulants of Discriminant Function when the Parameters are Known	5
2.2	First and Second Cumulants of Discriminant Function when the Parameters are Unknown	6
3	Edgeworth-type Expansion of the Discriminant Function	17
3.1	Edgeworth-type Expansion of the Discriminant Function $\tilde{D}$	18
3.2	Edgeworth-type Expansion of the discriminant function $\hat{D}$	21
4	Misclassification Errors	27
5	Conclusion	33
5.1	Discussion	33
5.2	Future research	37
6	Appendices	39
	Bibliography	43

Introduction

THIS thesis is concerned with the problem of classification of an observation into one of two populations. Distributions connected to the populations can be known as well as unknown.

Several researchers in multivariate statistics have been interested in discriminant analysis and classification techniques. These multivariate techniques are concerned with discriminating between distinct sets of observations and classifying new observations into predefined populations. Though discriminant analysis and classification methodology go hand in hand in several studies, the two approaches are not the same. Classification techniques are however usually derived via discriminant analysis. Discriminant analysis is about discriminating between populations, which can be done using a significance test or defining a distance measure that reflects the separation between the populations. On the other hand, classification is about allocating, for instance in the case of two populations, new units into one of the two populations.

1.1 Background

Fisher (1936, 1938), is considered to be one of the pioneers of the concept of classification. He showed how a set of multiple measurements can separate two populations. Fisher developed the so called linear discriminant analysis (LDA). To discriminate between taxonomic species, Fisher (1936) derived an equation that maximizes the distance between the two species using a set of multiple measurements from the groups. Moreover, Wallace and Travers (1938) who communicated with Fisher applied his technique in metrical psychology and Smith (1936) in plant selection. Furthermore, Rao (1948) used the same technique in biology. Wald (1944) studied a problem arising in the classification of a new individual into one of two groups assuming that from past experience, the test scores of a certain number of individuals who are known to belong to the first population as well as the test scores of some individuals known to belong to the second population

are available. Furthermore, Anderson (1951) analysed classification techniques choosing the rule of classification based on costs of misclassification. In addition, assuming normality and a common variance, Schaafsma and van Vark (1977) derived results for univariate two-group classification problems and later on Schaafsma and van Vark (1979) derived asymptotic results for the multivariate case under similar assumptions. Recently, in modern machine learning literature, classification and discrimination are described as a subcategory of supervised learning. Izenman (2008) refers to these two techniques together as class prediction.

In classification methodology we have $\pi_i, i \in \{1, \dots, q\}$, populations and we wish to allocate an observation $\mathbf{x}$ as a p -dimensional random vector $\mathbf{x} = (x_1, \dots, x_p)^\top$ to one of these populations. In LDA, it is often assumed that the observations in the i^{th} group are drawn from a multivariate normal distribution. For simplicity, assume one is interested in separating two populations and allocate new observations to one of the two normal populations with equal co-variance matrices, namely, $\mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_i, i \in \{1, 2\}$, is the p -dimensional vector of means of the i^{th} population and $\boldsymbol{\Sigma}$ is the common covariance matrix of each population. Fisher's discriminant function is given by:

$$L(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2), \quad (1.1)$$

and the classification procedure for a new observation $\mathbf{x}$ is that $\mathbf{x}$ is classified into π_1 if $L(\mathbf{x}; \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}) > 0$ and otherwise $\mathbf{x}$ is classified into π_2 .

In classification problems, as in some cases above, we consider that the distributions are completely known beforehand whereas in other cases, they can depend on unknown parameters which must be estimated from a sample drawn from respective population Anderson (1951). In this thesis, we are going to consider a case of two multivariate normal populations, $\mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}), i \in \{1, 2\}$, with the same covariance, $\boldsymbol{\Sigma}$. Assuming a common dispersion is not as limiting as it sounds because it means that classification of the new observation is reliant to the mean values of the distribution. If not, one can for instance use the medians. In LDA, the aim is to allocate the p -dimensional vector $\mathbf{x} = (x_1, \dots, x_p)^\top$ and to interpret the differences between the populations in terms of linear functions of the original variables. Such functions constitute the discriminant function or classification function.

Whenever classifying an object, there is a chance that it might be wrongly classified. Therefore, in statistics, we are interested in the probability of such an error. Fujikoshi et al. (2011) defines the probability of misclassification as a measure of the goodness of a classification rule. Generally, misclassification errors involve unknown parameters, thus they need to be estimated. To develop a classification rule we need to examine for the characteristics of individuals that best distinguish between the populations. Let R_i denote the region defined by the set of characteristics of individuals in population π_i . Therefore, when a new individual falls into R_i , it is classified into π_i . The rule is to introduce a region $R_i \in \mathcal{R}^p$ such that the risk for misclassification becomes small. Misclassifications happen when some units fall into wrong regions. Let Ω denote the sample space. Let R_1 be the set of $\mathbf{x}$ -values where $\mathbf{x}$ is classified into π_1 , then $R_2 = \Omega - R_1$ is the set of values where $\mathbf{x} \in \pi_2$. Furthermore, since every observation belongs into one of the two populations, we have that $\Omega = R_1 \cup R_2$. There exists other features of classification than

the characteristics of objects; prior probabilities and cost for misclassification. Observations are more likely to be allocated in a more prevalent population and stronger evidence would be needed to classify an object to the population with smaller prior probability. In addition, it may cost more to misclassify a severely sick subject as healthy than a healthy subject as being ill. However, these features will not be taken into account in this thesis. There are different approaches of deriving a discriminant function. The likelihood approach will be applied based on the fact that the random observation x either belongs to π_1 or π_2 .

In his book, Sitgreaves (1961) pointed out that the distribution of the classification function is often too complicated to allow for easy numerical calculations of misclassification probabilities. Several authors, such as Okamoto (1963), Anderson (1951, 2003) have investigated the distribution of the linear discriminant function. One way of getting around the problem of such complexity is to approximate the discriminant distribution function using asymptotic expansions. Edgeworth (1907) introduced the so-called Edgeworth series that approximates a probability distribution in terms of its cumulants. Gupta and Panchapakesan (1982), reviewed Edgeworth expansions in statistics. This concept attracted several applications in many areas of statistics such as Ranga Rao (1960) who obtained Edgeworth expansions and Davis (1976) who introduced a multivariate Edgeworth expansion. In addition, Fujikoshi et al. (2011) investigated the Edgeworth expansion and its validity. Furthermore, Kollo and von Rosen (2005) described Edgeworth-type expansions where a complicated density function is described through a simpler one. In this thesis, we will use an Edgeworth-type expansion to approximate the distribution of the discriminant function through a standard normal distribution and derive approximate probabilities of misclassification.

1.2 Motivation

This thesis mainly consists of two topics; derivation of the first two cumulants of the maximum likelihood based discriminant function and derivation of misclassification errors via an Edgeworth-type expansion. The distribution of the discriminant function is complicated, which makes it difficult to obtain exact misclassification errors. This is what motivates our research. Several researchers investigated a plug-in approach for deriving a linear discriminant function. In this thesis, a maximum likelihood approach is used to derive discriminant functions of quadratic form. Thereafter, probabilities of misclassifications are approximated via Edgeworth-type expansions which can be generalized to cases with classification of observations into more than two normal populations.

1.3 Aims and outline

1.3.1 Objectives of the thesis

The aim of this thesis is to develop a new approach for deriving approximations of the misclassification errors for the likelihood based discriminant function. The specific objectives are the following:

- (i) Calculate the first two cumulants of the likelihood based discriminant function;

- (ii) Approximate the misclassification errors through Edgeworth-type expansions.

1.3.2 Thesis outline

In Chapter 2, we discuss the first and second cumulants of the maximum likelihood based discriminant function. First we assume the means and the covariance to be completely known. Thereafter, all parameters are assumed to be unknown. In Chapter 3, Edgeworth-type expansions of the discriminant functions defined in Chapter 2 are developed in terms of their first two cumulants found in the same chapter. In Chapter 4, probabilities of misclassification are approximated based on Edgeworth-type expansions expressed in Chapter 3.

Notations used throughout this thesis are defined in Appendix A. The main theorems of this thesis are proved throughout the chapters but for standard theorems, some are proved in Appendices and others are cited and left with no proofs.

2

First and Second Cumulants of Discriminant Function

EXPRESSIONS of the first and second cumulants for the maximum likelihood based discriminant functions are obtained in this chapter. We will only consider the case with two populations. At first, it is assumed that the two populations have known means and a common covariance matrix. Thereafter, we will express the first two cumulants of the linear discriminant function with estimated parameters.

2.1 First and Second Cumulants of Discriminant Function when the Parameters are Known

In this section, we assume that a new observation is to be classified into one of two known multivariate normal populations with known means and a common known covariance matrix. We want to calculate the likelihood ratio based on whether $\mathbf{x}$ belongs to π_1 or π_2 . Let $\mathbf{y} \in \pi_1$, $\mathbf{z} \in \pi_2$ and $\mathbf{x}$ be the new observation to be classified. If $\mathbf{x} \in \pi_i$, $i \in \{1, 2\}$, the likelihood is given by

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

The ratio of the likelihood functions equals

$$\frac{L_1(\mathbf{x})}{L_2(\mathbf{x})} = e^{-\frac{1}{2} [(\mathbf{x} - \boldsymbol{\mu}_1)^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) - (\mathbf{x} - \boldsymbol{\mu}_2)^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_2)]}. \quad (2.1)$$

Thereupon, the observation $\mathbf{x}$ is classified into population π_1 if the ratio (2.1) is greater than or equal to 1 and otherwise into π_2 . Taking the logarithm of the ratio yields

$$\begin{aligned} D &= \ln \left(\frac{L_1(\mathbf{x})}{L_2(\mathbf{x})} \right) = -\frac{1}{2} [(\mathbf{x} - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_1) - (\mathbf{x} - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_2)] \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)), \end{aligned} \quad (2.2)$$

which is a linear function in $\mathbf{x}$. The decision of classifying $\mathbf{x}$ depends on whether $D > 0$ or $D < 0$. Since D in (2.2) is a linear function in $\mathbf{x}$, D is normally distributed.

Theorem 2.1

Consider the discriminant function in (2.2). Then

$$\begin{cases} D \sim \mathcal{N}(\frac{1}{2}\Delta^2, \Delta^2), & \text{if } \mathbf{x} \in \pi_1, \\ D \sim \mathcal{N}(-\frac{1}{2}\Delta^2, \Delta^2), & \text{if } \mathbf{x} \in \pi_2, \end{cases}$$

where

$$\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad (2.3)$$

is the Mahalanobis squared distance.

Proof: Assume $\mathbf{x} \in \pi_1$. Then,

$$\begin{aligned} E[D] &= E \left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)) \right] \\ &= \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \frac{1}{2} \Delta^2, \\ \text{Var}[D] &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = \Delta^2. \end{aligned}$$

Similarly, for $\mathbf{x} \in \pi_2$,

$$E[D] = -\frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

with the same variance as when $\mathbf{x} \in \pi_1$. □

2.2 First and Second Cumulants of Discriminant Function when the Parameters are Unknown

In Section 2.1 we have assumed that the two populations are known exactly. Suppose the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are unknown, which is the most common case in applications. One approach is to replace in the discriminant function the unknown parameters by their estimators, which is referred to as the plug-in strategy. Let the samples $\mathbf{y}_i$, $i \in \{1, \dots, n_1\}$ and $\mathbf{z}_j$, $j \in \{1, \dots, n_2\}$ from each population be available. Thus we can estimate the unknown parameters in Δ^2 by

$$\hat{\boldsymbol{\mu}}_1 = \frac{\mathbf{Y} \mathbf{1}_{n_1}}{n_1} = \bar{\mathbf{y}}, \quad \hat{\boldsymbol{\mu}}_2 = \frac{\mathbf{Z} \mathbf{1}_{n_2}}{n_2} = \bar{\mathbf{z}}, \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n_1 + n_2} \mathbf{X}(\mathbf{I}_{n_1 + n_2} - \mathbf{C}^\top (\mathbf{C} \mathbf{C}^\top)^{-1} \mathbf{C}) \mathbf{X}^\top, \quad (2.4)$$

where

$$\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_{n_1}), \mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_{n_2}), \mathbf{X} = (\mathbf{Y} : \mathbf{Z}), \mathbf{C} = (\mathbf{1}_{n_1}^\top \otimes \begin{pmatrix} 1 \\ 0 \end{pmatrix} : \mathbf{1}_{n_2}^\top \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix}),$$

and $\mathbf{1}_n$ is a column vector of n ones. In fact

$$(\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2) = \mathbf{X} \mathbf{C}^\top (\mathbf{C} \mathbf{C}^\top)^{-1}.$$

Another approach which can be used in classification to handle unknown parameters is the maximum likelihood approach with $\mathbf{x}$ involved in the estimation. Assume observations to be normally distributed. Let $\mathbf{y}_i \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $i \in \{1, \dots, n_1\}$, be a sample from π_1 collected in $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_{n_1})$ and $\mathbf{z}_j \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, $j \in \{1, \dots, n_2\}$ be a sample from π_2 collected in $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_{n_2})$. Moreover $\mathbf{Y}$, $\mathbf{Z}$, and $\mathbf{x}$ are supposed to be jointly independent. If $\mathbf{x} \in \pi_1$ the following two models emerge in terms of $\mathbf{Y}$, $\mathbf{Z}$ and $\mathbf{x}$,

$$\begin{aligned} (\mathbf{Y} : \mathbf{x}) &= \boldsymbol{\mu}_1 \mathbf{1}_{n_1+1}^\top + \boldsymbol{\varepsilon}, & \boldsymbol{\varepsilon} &\sim \mathcal{N}_{p, n_1+1}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_{n_1+1}), \\ \mathbf{Z} &= \boldsymbol{\mu}_2 \mathbf{1}_{n_2}^\top + \boldsymbol{\varepsilon}, & \boldsymbol{\varepsilon} &\sim \mathcal{N}_{p, n_2}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_{n_2}). \end{aligned} \quad (2.5)$$

If $\mathbf{x} \in \pi_2$ the model is specified as follows:

$$\begin{aligned} \mathbf{Y} &= \boldsymbol{\mu}_1 \mathbf{1}_{n_1}^\top + \boldsymbol{\varepsilon}, & \boldsymbol{\varepsilon} &\sim \mathcal{N}_{p, n_1+1}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_{n_1}), \\ (\mathbf{Z} : \mathbf{x}) &= \boldsymbol{\mu}_2 \mathbf{1}_{n_2+1}^\top + \boldsymbol{\varepsilon}, & \boldsymbol{\varepsilon} &\sim \mathcal{N}_{p, n_2+1}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_{n_2+1}). \end{aligned} \quad (2.6)$$

The equations (2.5) and (2.6) above will be used to create the likelihood function when $\mathbf{x} \in \pi_1$ and $\mathbf{x} \in \pi_2$, respectively. Put

$$\begin{aligned} \mathbf{X} &= (\mathbf{Y} : \mathbf{x} : \mathbf{Z}), \quad \mathbf{C} = \begin{pmatrix} \mathbf{1}_{n_1+1}^\top & \mathbf{0}_{n_2}^\top \\ \mathbf{0}_{n_1+1}^\top & \mathbf{1}_{n_2}^\top \end{pmatrix}, \quad \tilde{\mathbf{C}} = \begin{pmatrix} \mathbf{1}_{n_1}^\top & \mathbf{0}_{n_2+1}^\top \\ \mathbf{0}_{n_1}^\top & \mathbf{1}_{n_2+1}^\top \end{pmatrix}, \\ \mathbf{S} &= \mathbf{X}(\mathbf{I} - \mathbf{C}^\top (\mathbf{C} \mathbf{C}^\top)^{-1} \mathbf{C}) \mathbf{X}^\top, \quad \tilde{\mathbf{S}} = \mathbf{X}(\mathbf{I} - \tilde{\mathbf{C}}^\top (\tilde{\mathbf{C}} \tilde{\mathbf{C}}^\top)^{-1} \tilde{\mathbf{C}}) \mathbf{X}^\top, \end{aligned} \quad (2.7)$$

where $\mathbf{0}_n$ is a column vector of n zeros.

The following theorem by Kollo and von Rosen (2005) presents an inequality which will be used later.

Theorem 2.2

Let $\boldsymbol{\Sigma}$ and $\mathbf{S}$ be positive definite matrices of size $p \times p$. Then

$$|\boldsymbol{\Sigma}|^{-\frac{1}{2}n} e^{-\frac{1}{2} \text{tr}(\boldsymbol{\Sigma}^{-1} \mathbf{S})} \leq \left| \frac{1}{n} \mathbf{S} \right|^{-\frac{1}{2}n} e^{-\frac{1}{2}np},$$

and equality holds if and only if $\boldsymbol{\Sigma} = \frac{1}{n} \mathbf{S}$.

Lemma 2.1

Let $\mathbf{C}$, $\tilde{\mathbf{C}}$, $\mathbf{S}$ and $\tilde{\mathbf{S}}$ be defined as in (2.7), and $n = n_1 + n_2$. Then, If $\mathbf{x} \in \pi_1$, the maximum likelihood estimators of the unknown parameters in (2.5) equal

$$\begin{aligned} (\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2) &= \mathbf{X} \mathbf{C}^\top (\mathbf{C} \mathbf{C}^\top)^{-1}, \\ (n+1) \hat{\boldsymbol{\Sigma}} &= \mathbf{S}. \end{aligned}$$

If $\mathbf{x} \in \pi_2$, the maximum estimators of the unknown parameters in (2.6) equal

$$\begin{aligned} (\hat{\boldsymbol{\mu}}_1, \hat{\boldsymbol{\mu}}_2) &= \mathbf{X} \tilde{\mathbf{C}}^\top (\tilde{\mathbf{C}} \tilde{\mathbf{C}}^\top)^{-1}, \\ (n+1) \hat{\boldsymbol{\Sigma}} &= \tilde{\mathbf{S}}. \end{aligned}$$

Proof: If $\mathbf{x} \in \pi_1$, the likelihood, $L_{(\mathbf{Y}:\mathbf{x})}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, is given by

$$L_{(\mathbf{Y}:\mathbf{x})}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) = (2\pi)^{-(n_1+1)p/2} |\boldsymbol{\Sigma}|^{-(n_1+1)/2} e^{-\frac{1}{2} \sum_{i=1}^{n_1+1} (\mathbf{y}_i - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_1)}, \quad (2.8)$$

whereas the likelihood, $L_{\mathbf{Z}}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$, is given by

$$L_{\mathbf{Z}}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}) = (2\pi)^{-n_2 p/2} |\boldsymbol{\Sigma}|^{-n_2/2} e^{-\frac{1}{2} \sum_{j=1}^{n_2} (\mathbf{z}_j - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{z}_j - \boldsymbol{\mu}_2)}. \quad (2.9)$$

Hence, for $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 : \boldsymbol{\mu}_2)$, and $n = n_1 + n_2$ the likelihood, $L_{\mathbf{X}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, is given by:

$$L_{\mathbf{X}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-(n+1)p/2} |\boldsymbol{\Sigma}|^{-(n+1)/2} e^{-\frac{1}{2} \text{tr}\{\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}\mathbf{C})(\mathbf{X} - \boldsymbol{\mu}\mathbf{C})^\top\}}.$$

Taking the logarithm of the likelihood function above yields

$$\ln L_{\mathbf{X}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = c - \frac{n+1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \text{tr}\{\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}\mathbf{C})(\mathbf{X} - \boldsymbol{\mu}\mathbf{C})^\top\}, \quad (2.10)$$

where $c = |n+1| \frac{p}{2} \ln(2\pi)$. Write

$$\mathbf{X} - \boldsymbol{\mu}\mathbf{C} = [\mathbf{X} - \mathbf{X}\mathbf{C}^\top(\mathbf{C}\mathbf{C}^\top)^{-1}\mathbf{C}] + [\mathbf{X}\mathbf{C}^\top(\mathbf{C}\mathbf{C}^\top)^{-1}\mathbf{C} - \boldsymbol{\mu}\mathbf{C}] \equiv \mathbf{A}_1 + \mathbf{A}_2,$$

and note that $\mathbf{A}_1 \mathbf{A}_2^\top = \mathbf{0}$. Consequently,

$$\text{tr}\{\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}\mathbf{C})(\mathbf{X} - \boldsymbol{\mu}\mathbf{C})^\top\} = \text{tr}\{\boldsymbol{\Sigma}^{-1} \mathbf{A}_1 \mathbf{A}_1^\top\} + \text{tr}\{\boldsymbol{\Sigma}^{-1} \mathbf{A}_2 \mathbf{A}_2^\top\}. \quad (2.11)$$

The first term in (2.11) equals

$$\text{tr}\{\boldsymbol{\Sigma}^{-1} \mathbf{A}_1 \mathbf{A}_1^\top\} = \text{tr}\{\boldsymbol{\Sigma}^{-1} \mathbf{X}(\mathbf{I} - \mathbf{C}^\top(\mathbf{C}\mathbf{C}^\top)^{-1}\mathbf{C})\mathbf{X}^\top\} = \text{tr}\{\boldsymbol{\Sigma}^{-1} \mathbf{S}\},$$

which is independent of $\boldsymbol{\mu}$. The second term in (2.11) becomes

$$\begin{aligned} \text{tr}\{\boldsymbol{\Sigma}^{-1} \mathbf{A}_2 \mathbf{A}_2^\top\} &= \text{tr}\{\boldsymbol{\Sigma}^{-1}(\mathbf{X}\mathbf{C}^\top(\mathbf{C}\mathbf{C}^\top)^{-1} - \boldsymbol{\mu})\mathbf{C}\mathbf{C}^\top(\mathbf{X}\mathbf{C}^\top(\mathbf{C}\mathbf{C}^\top)^{-1} - \boldsymbol{\mu})^\top\} \\ &= \text{tr}\{\boldsymbol{\Sigma}^{-1/2}(\mathbf{X}\mathbf{C}^\top(\mathbf{C}\mathbf{C}^\top)^{-1} - \boldsymbol{\mu})\mathbf{C}\mathbf{C}^\top(\mathbf{X}\mathbf{C}^\top(\mathbf{C}\mathbf{C}^\top)^{-1} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1/2}\} \geq 0, \end{aligned}$$

which is equal to zero if and only if $\mathbf{X}\mathbf{C}^\top(\mathbf{C}\mathbf{C}^\top)^{-1} - \boldsymbol{\mu} = \mathbf{0}$, irrespective to the actual value of $\boldsymbol{\Sigma}$. Therefore, the MLE of $\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}} = \mathbf{X}\mathbf{C}^\top(\mathbf{C}\mathbf{C}^\top)^{-1}$, and substituting it for $\hat{\boldsymbol{\mu}}$, the last term of (2.11) vanishes. Then we maximize the remaining part with respect to $\boldsymbol{\Sigma}$ or equivalently with respect to $\boldsymbol{\Sigma}^{-1}$. That is, we maximize the function

$$g(\boldsymbol{\Sigma}^{-1}) = c + \frac{n+1}{2} \ln |\boldsymbol{\Sigma}^{-1}| - \frac{1}{2} \text{tr}\{\boldsymbol{\Sigma}^{-1} \mathbf{S}\}. \quad (2.12)$$

Therefore, Theorem 2.2 can be used to maximize the function in (2.12),

$$g(\boldsymbol{\Sigma}^{-1}) \geq c + \frac{n+1}{2} \ln \left| \frac{1}{n+1} \mathbf{S} \right|^{-1} - \frac{1}{2} p(n+1), \quad (2.13)$$

and equality holds if and only if $\boldsymbol{\Sigma} = \frac{1}{n+1} \mathbf{S}$. Then the proof of Lemma 2.1 follows. Similarly, if $\mathbf{x} \in \pi_2$, the likelihood under the second model in Lemma 2.1 is maximized using similar calculations as when $\mathbf{x} \in \pi_1$. $\square$

Replacing the unknown parameters with their estimators in the likelihood functions, the likelihood ratio for classification of $\mathbf{x}$ into π_1 or π_2 is given by

$$\frac{(2\pi)^{-\frac{p}{2}(n+1)} |\hat{\Sigma}|^{-\frac{1}{2}(n+1)} e^{-\frac{1}{2} \text{tr}\{\hat{\Sigma}^{-1}(\mathbf{X} - (\hat{\mu}_1, \hat{\mu}_2)\mathbf{C})(\mathbf{X} - (\hat{\mu}_1, \hat{\mu}_2)\mathbf{C})^\top\}}}{(2\pi)^{-\frac{p}{2}(n+1)} |\hat{\Sigma}|^{-\frac{1}{2}(n+1)} e^{-\frac{1}{2} \text{tr}\{\hat{\Sigma}^{-1}(\mathbf{X} - (\hat{\mu}_1, \hat{\mu}_2)\mathbf{C})(\mathbf{X} - (\hat{\mu}_1, \hat{\mu}_2)\mathbf{C})^\top\}}} = \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}|} \right)^{\frac{n+1}{2}}, \quad (2.14)$$

where $\hat{\Sigma}$ and $\hat{\Sigma}$ are generally different. The new observation $\mathbf{x}$ is classified into π_1 if the ratio (2.14) is greater than 1 and into π_2 if the ratio (2.14) is smaller than 1. Furthermore, consider the averages defined by (2.4). Let

$$\mathbf{S}_y = \mathbf{Y}\mathbf{Y}^\top - n_1 \bar{\mathbf{y}}\bar{\mathbf{y}}^\top, \quad \mathbf{S}_z = \mathbf{Z}\mathbf{Z}^\top - n_2 \bar{\mathbf{z}}\bar{\mathbf{z}}^\top$$

and put $\mathbf{S}_{yz} = \mathbf{S}_y + \mathbf{S}_z$; the sum of squares matrix based on the joint sample from populations π_1 and π_2 , without taking into consideration the new observation. Moreover $\mathbf{S}_{yz} \sim W_p(\Sigma, n_1 + n_2 - 2)$. Consequently, estimates $\mathbf{S}$ and $\tilde{\mathbf{S}}$ can be expressed as

$$\begin{aligned} \mathbf{S} &= \mathbf{X}(\mathbf{I} - \mathbf{C}^\top(\mathbf{C}\mathbf{C}^\top)^{-1})\mathbf{X}^\top \\ &= \mathbf{Y}\mathbf{Y}^\top + \mathbf{x}\mathbf{x}^\top + \mathbf{Z}\mathbf{Z}^\top - (\mathbf{Y}\mathbf{1}_{n_1}^\top + \mathbf{x}, \mathbf{Z}\mathbf{1}_{n_2}^\top) \begin{pmatrix} \frac{1}{n_1+1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix} (\mathbf{Y}\mathbf{1}_{n_1}^\top + \mathbf{x}, \mathbf{Z}\mathbf{1}_{n_2}^\top)^\top \\ &= \mathbf{S}_y + \mathbf{S}_z + \mathbf{x}\mathbf{x}^\top + n_1 \bar{\mathbf{y}}\bar{\mathbf{y}}^\top - \frac{n_1^2}{n_1 + 1} \bar{\mathbf{y}}\bar{\mathbf{y}}^\top - \frac{n_1}{n_1 + 1} (\bar{\mathbf{y}}\mathbf{x}^\top + \mathbf{x}\bar{\mathbf{y}}^\top) - \frac{1}{n_1 + 1} \mathbf{x}\mathbf{x}^\top \\ &= \mathbf{S}_{yz} + \frac{n_1}{n_1 + 1} (\mathbf{x}\mathbf{x}^\top + \bar{\mathbf{y}}\bar{\mathbf{y}}^\top - \bar{\mathbf{y}}\mathbf{x}^\top - \mathbf{x}\bar{\mathbf{y}}^\top) \\ &= \mathbf{S}_{yz} + \frac{n_1}{n_1 + 1} (\bar{\mathbf{y}} - \mathbf{x})(\bar{\mathbf{y}} - \mathbf{x})^\top. \end{aligned}$$

Similarly, it can be shown that

$$\tilde{\mathbf{S}} = \mathbf{S}_{yz} + \frac{n_2}{n_2 + 1} (\bar{\mathbf{z}} - \mathbf{x})(\bar{\mathbf{z}} - \mathbf{x})^\top.$$

As a result, the likelihood ratio in (2.14) is equivalent to

$$\frac{|\tilde{\mathbf{S}}|}{|\mathbf{S}|} = \frac{|\mathbf{S}_{yz} + \frac{n_2}{n_2+1} (\bar{\mathbf{z}} - \mathbf{x})(\bar{\mathbf{z}} - \mathbf{x})^\top|}{|\mathbf{S}_{yz} + \frac{n_1}{n_1+1} (\bar{\mathbf{y}} - \mathbf{x})(\bar{\mathbf{y}} - \mathbf{x})^\top|} = \frac{1 + \frac{n_2}{n_2+1} (\bar{\mathbf{z}} - \mathbf{x})^\top \mathbf{S}_{yz}^{-1} (\bar{\mathbf{z}} - \mathbf{x})}{1 + \frac{n_1}{n_1+1} (\bar{\mathbf{y}} - \mathbf{x})^\top \mathbf{S}_{yz}^{-1} (\bar{\mathbf{y}} - \mathbf{x})}, \quad n_1 + n_2 - 2 \geq p. \quad (2.15)$$

Note that $\mathbf{S}_{yz}^{-1}$ exists if $n_1 + n_2 - 2 \geq p$. A large likelihood suggests that $\mathbf{x}$ is to be classified into π_1 . Therefore, $\mathbf{x}$ is classified into π_1 if

$$\frac{n_2}{n_2 + 1} (\bar{\mathbf{z}} - \mathbf{x})^\top \mathbf{S}_{yz}^{-1} (\bar{\mathbf{z}} - \mathbf{x}) \geq \frac{n_1}{n_1 + 1} (\bar{\mathbf{y}} - \mathbf{x})^\top \mathbf{S}_{yz}^{-1} (\bar{\mathbf{y}} - \mathbf{x}). \quad (2.16)$$

Since, $\mathbf{S}_{yz} \sim W_p(\Sigma, n)$, for $n = n_1 + n_2 - 2$, then using Theorem 6.4 (i) in Appendix B, $E[\mathbf{S}_{yz}^{-1}] = \frac{1}{n_1 + n_2 - p - 3} \Sigma^{-1}$, we can transform (2.16) by multiplying both sides by $\frac{1}{2}(n_1 + n_2 - p - 3)$ and to make it comparable with (2.2) we apply $\frac{1}{2}$.

Proposition 2.1

Consider the models given in (2.5), suppose $m = n_1 + n_2 - p - 3 > 0$, and put

$$\hat{D} = \frac{n_2 m}{n_2 + 1} \frac{1}{2} (\bar{\mathbf{z}} - \mathbf{x})^\top \mathbf{S}_{yz}^{-1} (\bar{\mathbf{z}} - \mathbf{x}) - \frac{n_1 m}{n_1 + 1} \frac{1}{2} (\bar{\mathbf{y}} - \mathbf{x})^\top \mathbf{S}_{yz}^{-1} (\bar{\mathbf{y}} - \mathbf{x}). \quad (2.17)$$

The observation $\mathbf{x}$ is classified in π_1 if $\hat{D} \geq 0$ and if $\hat{D} < 0$ then $\mathbf{x}$ is classified in π_2 .

Theorem 2.3

Consider the discriminant function in Proposition 2.1. The expected value of the discriminant function is given by

$$E[\hat{D}] = \begin{cases} \frac{1}{2} \frac{n_2}{n_2 + 1} \Delta^2, & \text{if } \mathbf{x} \in \pi_1, \\ -\frac{1}{2} \frac{n_1}{n_1 + 1} \Delta^2, & \text{if } \mathbf{x} \in \pi_2, \end{cases} \quad (2.18)$$

where Δ^2 is the Mahalanobis squared distance given in (2.3).

Proof: Note that $\bar{\mathbf{y}}$, $\bar{\mathbf{z}}$, $\mathbf{S}_{yz}$ and $\mathbf{x}$ are independently distributed. Let $\boldsymbol{\mu}_x = E[\mathbf{x}]$. Put $\mathbf{u}^\top = (\mathbf{u}_1^\top, \mathbf{u}_2^\top)$, where

$$\begin{aligned} \mathbf{u}_1 &= \sqrt{\frac{n_2}{n_2 + 1}} \frac{1}{\sqrt{2}} (\bar{\mathbf{z}} - \mathbf{x}) \sim \mathcal{N}_p \left(\sqrt{\frac{n_2}{n_2 + 1}} \frac{1}{\sqrt{2}} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_x), \boldsymbol{\Sigma} \right) \\ \mathbf{u}_2 &= \sqrt{\frac{n_1}{n_1 + 1}} \frac{1}{\sqrt{2}} (\bar{\mathbf{y}} - \mathbf{x}) \sim \mathcal{N}_p \left(\sqrt{\frac{n_1}{n_1 + 1}} \frac{1}{\sqrt{2}} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_x), \boldsymbol{\Sigma} \right). \end{aligned} \quad (2.19)$$

Thus, $\mathbf{u} \sim \mathcal{N}_{2p}(\boldsymbol{\mu}, \boldsymbol{\psi})$, where

$$\begin{aligned} \boldsymbol{\mu}^\top &= \left(\sqrt{\frac{n_2}{n_2 + 1}} \frac{1}{\sqrt{2}} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_x)^\top, \sqrt{\frac{n_1}{n_1 + 1}} \frac{1}{\sqrt{2}} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_x)^\top \right), \\ \boldsymbol{\psi} &= \begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix} \otimes \frac{1}{2} \boldsymbol{\Sigma}, \end{aligned} \quad (2.20)$$

where

$$q = \sqrt{\frac{n_1}{n_1 + 1}} \sqrt{\frac{n_2}{n_2 + 1}}. \quad (2.21)$$

Now, let $\mathbf{Q} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \otimes m \mathbf{S}_{yz}^{-1}$. Hence,

$$\hat{D} = \mathbf{u}^\top \mathbf{Q} \mathbf{u} = \frac{m}{2} \left(\frac{n_2}{n_2 + 1} (\bar{\mathbf{z}} - \mathbf{x})^\top \mathbf{S}_{yz}^{-1} (\bar{\mathbf{z}} - \mathbf{x}) - \frac{n_1}{n_1 + 1} (\bar{\mathbf{y}} - \mathbf{x})^\top \mathbf{S}_{yz}^{-1} (\bar{\mathbf{y}} - \mathbf{x}) \right).$$

The vectors $\mathbf{u}$ and $\mathbf{Q}$ are independently distributed. Then, since $\mathbf{u} \sim \mathcal{N}_{2p}(\boldsymbol{\mu}, \boldsymbol{\psi})$, using Theorem 6.3 (i), Appendix B, we get

$$E[\hat{D} | \mathbf{S}_{yz}] = E[\mathbf{u}^\top \mathbf{Q} \mathbf{u} | \mathbf{S}_{yz}] = \underbrace{\text{tr}(\boldsymbol{\psi} \mathbf{Q})}_{=0} + \boldsymbol{\mu}^\top \mathbf{Q} \boldsymbol{\mu} = \boldsymbol{\mu}^\top \mathbf{Q} \boldsymbol{\mu}.$$

Since $\text{tr}(\boldsymbol{\psi} \mathbf{Q}) = \text{tr} \left\{ \begin{pmatrix} 1 & -q \\ q & -1 \end{pmatrix} \otimes \frac{1}{2} m \boldsymbol{\Sigma} \mathbf{S}_{yz}^{-1} \right\} = 0$. As a result,

$$\begin{aligned} E[\hat{D}] &= E[E[\hat{D} | \mathbf{S}_{yz}]] = E[\boldsymbol{\mu}^\top \mathbf{Q} \boldsymbol{\mu}] \\ &= E \left[\frac{1}{2} \frac{n_2}{n_2 + 1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_x)^\top m \mathbf{S}_{yz}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_x) - \frac{1}{2} \frac{n_1}{n_1 + 1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_x)^\top m \mathbf{S}_{yz}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_x) \right]. \end{aligned}$$

Thus, if $\mathbf{x} \in \pi_1$, then $\boldsymbol{\mu}_{\mathbf{x}} = \boldsymbol{\mu}_1$. Therefore, since from Theorem 6.4 (i), Appendix B, $E[m\mathbf{S}_{yz}^{-1}] = \boldsymbol{\Sigma}^{-1}$, we get

$$E[\widehat{D}] = \frac{1}{2} \frac{n_2}{n_2 + 1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top E[m\mathbf{S}_{yz}^{-1}] (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) = \frac{1}{2} \frac{n_2}{n_2 + 1} \Delta^2.$$

Similarly, if $\mathbf{x} \in \pi_2$, then $\boldsymbol{\mu}_{\mathbf{x}} = \boldsymbol{\mu}_2$ and

$$E[\widehat{D}] = -\frac{1}{2} \frac{n_1}{n_1 + 1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top E[m\mathbf{S}_{yz}^{-1}] (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) = -\frac{1}{2} \frac{n_1}{n_1 + 1} \Delta^2.$$

□

Definition 2.1. Let $m = n_1 + n_2 - p - 3$. Then

(i) $c_1 = \frac{1}{m}$; (ii) $c_2 = \frac{m-1}{m(m+1)(m-2)}$; (iii) $c_3 = \frac{1}{m-1} c_2$.

Theorem 2.4

Consider the discriminant function in Proposition 2.1. For $m = n_1 + n_2 - p - 3 > 2$, the variance of the discriminant function is given by

$$\text{Var}[\widehat{D}] = \begin{cases} (1-q^2)m^2c_0p + m^2 \frac{n_2}{n_2+1} c_0 \Delta^2 + \left(\frac{n_2}{n_2+1}\right)^2 \frac{1}{2(m-2)} (\Delta^2)^2, & \text{if } \mathbf{x} \in \pi_1, \\ (1-q^2)m^2c_0p + m^2 \frac{n_1}{n_1+1} c_0 \Delta^2 + \left(\frac{n_1}{n_1+1}\right)^2 \frac{1}{2(m-2)} (\Delta^2)^2, & \text{if } \mathbf{x} \in \pi_2, \end{cases} \quad (2.22)$$

where Δ^2 is the Mahalanobis squared distance given by (2.3) and

$$c_0 = \frac{m+p}{m(m-2)(m+1)}. \quad (2.23)$$

Proof: It will be utilized that

$$\text{Var}[\widehat{D}] = \text{Var}[E[\widehat{D}|\mathbf{S}_{yz}]] + E[\text{Var}[\widehat{D}|\mathbf{S}_{yz}]]. \quad (2.24)$$

Moreover, $\widehat{D} = \mathbf{u}^\top \mathbf{Q} \mathbf{u}$, $\mathbf{u} \sim \mathcal{N}_{2p}(\boldsymbol{\mu}, \boldsymbol{\psi})$, where $\mathbf{u}$, $\mathbf{Q}$ and $\boldsymbol{\psi}$ are defined in the proof of Theorem 2.3. Using Theorem 6.2 (ii) in Appendix B, we get,

$$\text{Var}[\widehat{D}|\mathbf{S}_{yz}] = 2\text{tr}\{\boldsymbol{\psi} \mathbf{Q} \boldsymbol{\psi} \mathbf{Q}\} + 4\boldsymbol{\mu}^\top \mathbf{Q} \boldsymbol{\psi} \mathbf{Q} \boldsymbol{\mu}. \quad (2.25)$$

Evaluating (2.25) step by step,

$$\begin{aligned} \mathbf{Q} \boldsymbol{\psi} \mathbf{Q} &= \left[\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \otimes m\mathbf{S}_{yz}^{-1} \right] \left[\begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix} \otimes \frac{1}{2} \boldsymbol{\Sigma} \right] \left[\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \otimes m\mathbf{S}_{yz}^{-1} \right] \\ &= \begin{pmatrix} 1 & -q \\ -q & 1 \end{pmatrix} \otimes \frac{1}{2} m^2 \mathbf{S}_{yz}^{-1} \boldsymbol{\Sigma} \mathbf{S}_{yz}^{-1}, \end{aligned} \quad (2.26)$$

$$\begin{aligned} \boldsymbol{\psi} \mathbf{Q} \boldsymbol{\psi} \mathbf{Q} &= \left[\begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix} \otimes \frac{1}{2} \boldsymbol{\Sigma} \right] \begin{pmatrix} 1 & -q \\ -q & 1 \end{pmatrix} \otimes \frac{1}{2} m^2 \mathbf{S}_{yz}^{-1} \boldsymbol{\Sigma} \mathbf{S}_{yz}^{-1} \\ &= \begin{pmatrix} 1-q^2 & 0 \\ 0 & 1-q^2 \end{pmatrix} \otimes \frac{1}{4} m^2 \boldsymbol{\Sigma} \mathbf{S}_{yz}^{-1} \boldsymbol{\Sigma} \mathbf{S}_{yz}^{-1}. \end{aligned} \quad (2.27)$$

Then the first term in (2.25) becomes

$$\text{tr}\{\psi \bar{Q} \psi Q\} = \frac{1}{2}(1 - q^2)m^2 \text{tr}\{\Sigma S_{yz}^{-1} \Sigma S_{yz}^{-1}\}.$$

Furthermore, for $q = \sqrt{\frac{n_1}{n_1+1}} \sqrt{\frac{n_2}{n_2+1}}$, the second term of (2.25) equals

$$\begin{aligned} \mu^\top Q \psi Q \mu &= \left[\left(\sqrt{\frac{n_2}{n_2+1}} \frac{1}{\sqrt{2}} (\mu_2 - \mu_x)^\top - q \sqrt{\frac{n_1}{n_1+1}} \frac{1}{\sqrt{2}} (\mu_1 - \mu_x)^\top, \right. \right. \\ &\quad \left. \left. - q \sqrt{\frac{n_2}{n_2+1}} \frac{1}{\sqrt{2}} (\mu_2 - \mu_x)^\top + \sqrt{\frac{n_1}{n_1+1}} \frac{1}{\sqrt{2}} (\mu_1 - \mu_x)^\top \right) \otimes \frac{1}{2} m^2 S_{yz}^{-1} \Sigma S_{yz}^{-1} \right] \\ &\quad \times \left(\sqrt{\frac{n_2}{n_2+1}} \frac{1}{\sqrt{2}} (\mu_2 - \mu_x), \sqrt{\frac{n_1}{n_1+1}} \frac{1}{\sqrt{2}} (\mu_1 - \mu_x) \right)^\top \\ &= \frac{1}{4} m^2 \frac{n_2}{n_2+1} (\mu_2 - \mu_x)^\top S_{yz}^{-1} \Sigma S_{yz}^{-1} (\mu_2 - \mu_x) - \frac{1}{4} q^2 m^2 (\mu_1 - \mu_x)^\top S_{yz}^{-1} \Sigma S_{yz}^{-1} (\mu_2 - \mu_x) \\ &\quad - \frac{1}{4} q^2 m^2 (\mu_2 - \mu_x)^\top S_{yz}^{-1} \Sigma S_{yz}^{-1} (\mu_1 - \mu_x) + \frac{1}{4} m^2 \frac{n_1}{n_1+1} (\mu_1 - \mu_x)^\top S_{yz}^{-1} \Sigma S_{yz}^{-1} (\mu_1 - \mu_x) \\ &= \frac{1}{4} m^2 \frac{n_2}{n_2+1} (\mu_2 - \mu_x)^\top S_{yz}^{-1} \Sigma S_{yz}^{-1} (\mu_2 - \mu_x) - \underbrace{\frac{1}{2} q^2 m^2 (\mu_2 - \mu_x)^\top S_{yz}^{-1} \Sigma S_{yz}^{-1} (\mu_1 - \mu_x)}_{=0} \\ &\quad + \frac{1}{4} m^2 \frac{n_1}{n_1+1} (\mu_1 - \mu_x)^\top S_{yz}^{-1} \Sigma S_{yz}^{-1} (\mu_1 - \mu_x). \end{aligned} \quad (2.28)$$

The term $\frac{1}{2} q^2 m^2 (\mu_2 - \mu_x)^\top S_{yz}^{-1} \Sigma S_{yz}^{-1} (\mu_1 - \mu_x)$ is always zero because either $x \in \pi_1$ or $x \in \pi_2$.

Moreover, for $m > 2$, Theorem 6.4 (ii) in Appendix B yields $E[S_{yz}^{-1} \Sigma S_{yz}^{-1}] = c_0 \Sigma^{-1}$, where $c_0 = \frac{m+p}{m(m-2)(m+1)}$. Then, we can evaluate

$$\begin{aligned} E[\text{Var}[\hat{D} | S_{yz}]] &= (1 - q^2) m^2 \text{tr}\{E[\Sigma S_{yz}^{-1} \Sigma S_{yz}^{-1}]\} + m^2 \frac{n_2}{n_2+1} (\mu_2 - \mu_x)^\top E[S_{yz}^{-1} \Sigma S_{yz}^{-1}] (\mu_2 - \mu_x) \\ &\quad + m^2 \frac{n_1}{n_1+1} (\mu_1 - \mu_x)^\top E[S_{yz}^{-1} \Sigma S_{yz}^{-1}] (\mu_1 - \mu_x) \\ &= (1 - q^2) m^2 c_0 p + m^2 \frac{n_2}{n_2+1} c_0 (\mu_2 - \mu_x)^\top \Sigma^{-1} (\mu_2 - \mu_x) \\ &\quad + m^2 \frac{n_1}{n_1+1} c_0 (\mu_1 - \mu_x)^\top \Sigma^{-1} (\mu_1 - \mu_x). \end{aligned}$$

If $x \in \pi_1$ then

$$\begin{aligned} E[\text{Var}[\hat{D} | S_{yz}]] &= (1 - q^2) m^2 c_0 p + m^2 \frac{n_2}{n_2+1} c_0 (\mu_1 - \mu_2)^\top \Sigma^{-1} (\mu_1 - \mu_2) \\ &= (1 - q^2) m^2 c_0 p + m^2 \frac{n_2}{n_2+1} c_0 \Delta^2. \end{aligned}$$

Now, let's calculate $\text{Var}[E[\hat{D} | S_{yz}]]$. First we have

$$E[\hat{D} | S_{yz}] = \frac{1}{2} \frac{n_2}{n_2+1} (\mu_2 - \mu_x)^\top m S_{yz}^{-1} (\mu_2 - \mu_x) - \frac{1}{2} \frac{n_1}{n_1+1} (\mu_1 - \mu_x)^\top m S_{yz}^{-1} (\mu_1 - \mu_x).$$

Take $M = \frac{1}{2} \frac{n_2}{n_2+1} (\mu_2 - \mu_x)(\mu_2 - \mu_x)^\top - \frac{1}{2} \frac{n_1}{n_1+1} (\mu_1 - \mu_x)(\mu_1 - \mu_x)^\top$. Then, $E[\hat{D} | S_{yz}] = \text{tr}\{m S_{yz}^{-1} M\}$.

Hence, using Theorem 6.4 (vii) in Appendix B, with c_1 , c_2 and c_3 defined as in Definition 2.1 we get

$$\begin{aligned} Var[E[\hat{D}|\mathbf{S}_{yz}]] &= Var[\text{tr}\{m\mathbf{S}_{yz}^{-1}\mathbf{M}\}] = m^2 \text{vec}^\top \mathbf{M} Var[\mathbf{S}_{yz}^{-1}] \text{vec} \mathbf{M} \\ &= 2c_3 m^2 \text{vec}^\top \mathbf{M} (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) \text{vec} \mathbf{M} + (c_2 - c_1^2) (\text{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{M}\})^2. \end{aligned}$$

If $\mathbf{x} \in \pi_1$, $\mathbf{M} = \frac{1}{2} \frac{n_2}{n_2+1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top$ and if $\mathbf{x} \in \pi_2$, $\mathbf{M} = \frac{1}{2} \frac{n_1}{n_1+1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top$. Therefore, if $\mathbf{x} \in \pi_1$,

$$Var[E[\hat{D}|\mathbf{S}_{yz}]] = \frac{1}{2} m^2 \left(\frac{n_2}{n_2+1} \right)^2 (c_3 + \frac{1}{2}(c_2 - c_1^2))(\Delta^2)^2,$$

and if $\mathbf{x} \in \pi_2$,

$$Var[E[\hat{D}|\mathbf{S}_{yz}]] = \frac{1}{2} m^2 \left(\frac{n_1}{n_1+1} \right)^2 (c_3 + \frac{1}{2}(c_2 - c_1^2))(\Delta^2)^2,$$

where $c_3 + \frac{1}{2}c_2 - \frac{1}{2}c_1^2 = \frac{1}{m^2(m-2)}$. Thus the proof follows. $\square$

The classification function $\hat{D}$ given in Proposition 2.1 follows a complicated distribution since it expresses a difference of two non-central F-distributions. Its basic properties are given in (2.18) and (2.22) and the variance (2.22) can be large depending on the Mahalanobis squared distance and the sample sizes. An alternative idea is to consider asymptotics when both n_1 and n_2 become large but p is fixed.

Proposition 2.2

Consider the models given in (2.5) and put

$$\tilde{D} = \frac{1}{2} \frac{n_2}{n_2+1} (\bar{\mathbf{z}} - \mathbf{x})^\top \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{z}} - \mathbf{x}) - \frac{1}{2} \frac{n_1}{n_1+1} (\bar{\mathbf{y}} - \mathbf{x})^\top \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}} - \mathbf{x}). \quad (2.29)$$

If $\tilde{D} \geq 0$, $\mathbf{x}$ is classified in π_1 and if $\tilde{D} < 0$ then $\mathbf{x}$ is classified in π_2 .

Using Theorem 6.4 (i), in Appendix B, we get $E[m\mathbf{S}_{yz}^{-1}] = \boldsymbol{\Sigma}^{-1}$, $m = n_1 + n_2 - p - 3$. Hence, replacing $m\mathbf{S}_{yz}^{-1}$ in Proposition 2.1 by $\boldsymbol{\Sigma}^{-1}$ yields Proposition 2.2. Thereafter, we want to show that $\hat{D}$ and $\tilde{D}$ are asymptotically equivalently distributed, that is, for any ε ,

$$P((\hat{D} - \tilde{D})^2 > \varepsilon) \longrightarrow 0, \quad n_i \rightarrow \infty, \quad i \in \{1, 2\}. \quad (2.30)$$

Lemma 2.2

(Markov's Inequality). Let X be a random variable that takes only nonnegative values. Then for any positive real number a ,

$$P(X \geq a) \leq \frac{E[X]}{a},$$

provided that $E[X]$ exists.

Lemma 2.2 implies that (2.30) is true if it can be shown that $E[(\hat{D} - \tilde{D})^2]$ converges to 0. For large n_1 , $\frac{n_1}{n_1+1} \sim 1$ and the same with n_2 . Furthermore,

$$\begin{aligned}\hat{D} - \tilde{D} &= \frac{1}{2} \frac{n_2}{n_2 + 1} (\bar{\mathbf{z}} - \mathbf{x})^\top m \mathbf{S}_{yz}^{-1} (\bar{\mathbf{z}} - \mathbf{x}) - \frac{1}{2} \frac{n_1}{n_1 + 1} (\bar{\mathbf{y}} - \mathbf{x})^\top m \mathbf{S}_{yz}^{-1} (\bar{\mathbf{y}} - \mathbf{x}) \\ &\quad - \frac{1}{2} \frac{n_2}{n_2 + 1} (\bar{\mathbf{z}} - \mathbf{x})^\top \Sigma^{-1} (\bar{\mathbf{z}} - \mathbf{x}) + \frac{1}{2} \frac{n_1}{n_1 + 1} (\bar{\mathbf{y}} - \mathbf{x})^\top \Sigma^{-1} (\bar{\mathbf{y}} - \mathbf{x}) \\ &= \frac{1}{2} \frac{n_2}{n_2 + 1} (\bar{\mathbf{z}} - \mathbf{x})^\top (m \mathbf{S}_{yz}^{-1} - \Sigma^{-1}) (\bar{\mathbf{z}} - \mathbf{x}) \\ &\quad - \frac{1}{2} \frac{n_1}{n_1 + 1} (\bar{\mathbf{y}} - \mathbf{x})^\top (m \mathbf{S}_{yz}^{-1} - \Sigma^{-1}) (\bar{\mathbf{y}} - \mathbf{x}).\end{aligned}$$

Again let

$$\mathbf{u} = \left(\frac{1}{\sqrt{2}} \sqrt{\frac{n_2}{n_2 + 1}} (\bar{\mathbf{z}} - \mathbf{x})^\top, \frac{1}{\sqrt{2}} \sqrt{\frac{n_1}{n_1 + 1}} (\bar{\mathbf{y}} - \mathbf{x})^\top \right)^\top$$

and

$$\Psi = \begin{pmatrix} m \mathbf{S}_{yz}^{-1} - \Sigma^{-1} & 0 \\ 0 & \Sigma^{-1} - m \mathbf{S}_{yz}^{-1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \otimes (m \mathbf{S}_{yz}^{-1} - \Sigma^{-1}). \quad (2.31)$$

Moreover, $\mathbf{u} \sim \mathcal{N}_{2p}(\boldsymbol{\mu}, \boldsymbol{\psi})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\psi}$ are defined in (2.20). Hence, $\hat{D} - \tilde{D} = \mathbf{u}^\top \Psi \mathbf{u}$, and $(\hat{D} - \tilde{D})^2 = (\mathbf{u}^\top \otimes \mathbf{u}^\top) \text{vec} \Psi \text{vec}^\top \Psi (\mathbf{u} \otimes \mathbf{u})$. Applying Lemma 6.1 in Appendix B to (2.31),

$$\text{vec} \Psi = (\mathbf{I}_2 \otimes \mathbf{K}_{p,2} \otimes \mathbf{I}_p) \left(\text{vec} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \otimes \text{vec}(m \mathbf{S}_{yz}^{-1} - \Sigma^{-1}) \right),$$

where $\mathbf{K}_{p,2}$ is the commutation matrix defined in Definition 6.1, in Appendix A. Then,

$$\begin{aligned}E[(\hat{D} - \tilde{D})^2] &= E_{\mathbf{u}} [(\mathbf{u}^\top \otimes \mathbf{u}^\top) (\mathbf{I}_2 \otimes \mathbf{K}_{p,2} \otimes \mathbf{I}_p) \left(\text{vec} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \text{vec}^\top \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \otimes \text{Var}(m \mathbf{S}_{yz}^{-1}) \right) \\ &\quad \times (\mathbf{I}_2 \otimes \mathbf{K}_{p,2} \otimes \mathbf{I}_p) (\mathbf{u} \otimes \mathbf{u})],\end{aligned} \quad (2.32)$$

where, $E_{\mathbf{u}}[\cdot]$ is the expectation taken with respect to $\mathbf{u}$ and the expression $\text{Var}[\mathbf{S}_{yz}^{-1}]$ is given by Theorem 6.4 (iii), in Appendix B. The vector $\mathbf{u}$ is normally distributed with constant variance, thus the expectation $E_{\mathbf{u}}[\cdot]$ in (2.32) will not increase when n_1 and n_2 increase and hence the expectation depends only on $\text{Var}[m \mathbf{S}_{yz}^{-1}]$. It can be verified that using $\text{Var}[m \mathbf{S}_{yz}^{-1}] \rightarrow 0$ when $n_i \rightarrow \infty$, with $i \in \{1, 2\}$. Therefore, we can conclude that $\hat{D}$ and $\tilde{D}$ are asymptotically equivalent.

The distribution of the classification function $\tilde{D}$ is a difference between two non-central χ^2 distributions, which is less complicated to be evaluated than the distribution of $\hat{D}$. In addition, the distributions for $\Sigma^{-\frac{1}{2}}(\bar{\mathbf{z}} - \mathbf{x})$ and $\Sigma^{-\frac{1}{2}}(\bar{\mathbf{y}} - \mathbf{x})$ do not depend on Σ . The distribution for $\tilde{D}$ does not depend on the covariance, which makes sense since we have assumed a common variance for both populations and therefore it is reasonable that Σ is not involved in classification of $\mathbf{x}$ to either π_1 or π_2 .

In the following theorem, the expectations and variances for the classification function $\tilde{D}$ are presented, for comparison with the basic properties of $\hat{D}$, given in Theorem 2.3 and Theorem 2.4.

Theorem 2.5

Let $\tilde{D}$ be defined in Proposition 2.2. Then

$$\begin{aligned} (i) \quad & \text{if } \mathbf{x} \in \pi_1, \begin{cases} E[\tilde{D}] &= \frac{1}{2} \frac{n_2}{n_2+1} \Delta^2, \\ \text{Var}[\tilde{D}] &= p(1-q^2) + \frac{n_2}{n_2+1} \Delta^2, \end{cases} \\ (ii) \quad & \text{if } \mathbf{x} \in \pi_2, \begin{cases} E[\tilde{D}] &= -\frac{1}{2} \frac{n_1}{n_1+1} \Delta^2, \\ \text{Var}[\tilde{D}] &= p(1-q^2) + \frac{n_1}{n_1+1} \Delta^2, \end{cases} \end{aligned} \quad (2.33)$$

where $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $q = \sqrt{\frac{n_1}{n_1+1}} \sqrt{\frac{n_2}{n_2+1}}$.

Proof: Let $\mathbf{u}_1$ and $\mathbf{u}_2$ be defined in (2.19) and hence, $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2) \sim \mathcal{N}_{2p}(\boldsymbol{\mu}, \boldsymbol{\psi})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\psi}$ are defined in (2.20) and q in (2.21). Let $\mathbf{P} = \left[\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \otimes \boldsymbol{\Sigma}^{-1} \right]$. Then, $\tilde{D}$ in Proposition 2.2 equals $\mathbf{u}^\top \mathbf{P} \mathbf{u}$. Therefore, using Theorem 6.3 (i), in Appendix B, it follows that

$$E[\tilde{D}] = \frac{1}{2} \frac{n_2}{n_2+1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_x)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_x) - \frac{1}{2} \frac{n_1}{n_1+1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_x)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_x). \quad (2.34)$$

If $\mathbf{x} \in \pi_1$, $\boldsymbol{\mu}_x = \boldsymbol{\mu}_1$. Thus, the proof follows. Furthermore, using Theorem 6.2 (ii), in Appendix B we get

$$\text{Var}[\tilde{D}] = 2\text{tr}\{\boldsymbol{\psi} \mathbf{P} \boldsymbol{\psi} \mathbf{P}\} + 4\boldsymbol{\mu}^\top \mathbf{P} \boldsymbol{\psi} \mathbf{P} \boldsymbol{\mu}, \quad (2.35)$$

where we have

$$\mathbf{P} \boldsymbol{\psi} \mathbf{P} = \begin{pmatrix} 1 & -q \\ -q & -1 \end{pmatrix} \otimes \frac{1}{2} \boldsymbol{\Sigma}^{-1}. \quad (2.36)$$

Therefore,

$$\boldsymbol{\psi} \mathbf{P} \boldsymbol{\psi} \mathbf{P} = \begin{pmatrix} 1-q^2 & 0 \\ 0 & 1-q^2 \end{pmatrix} \otimes \frac{1}{4} \mathbf{I}_p. \quad (2.37)$$

Hence, the terms in (2.35) are given by

$$\text{tr}\{\boldsymbol{\psi} \mathbf{P} \boldsymbol{\psi} \mathbf{P}\} = \frac{1}{2} (1-q^2) \text{tr} \mathbf{I}_p = \frac{1}{2} p(1-q^2) \quad (2.38)$$

and

$$\begin{aligned} \boldsymbol{\mu}^\top \mathbf{P} \boldsymbol{\psi} \mathbf{P} \boldsymbol{\mu} = & \frac{1}{4} \left[\frac{n_2}{n_2+1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_x)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_x) - q^2 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_x)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_x) \right. \\ & \left. - q^2 (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_x)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_x) + \frac{n_1}{n_1+1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_x)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_x) \right]. \end{aligned} \quad (2.39)$$

Since $\boldsymbol{\mu}_x$ equals either $\boldsymbol{\mu}_1$ or $\boldsymbol{\mu}_2$, the two middle terms will always be zero. If $\mathbf{x} \in \pi_1$, replacing (2.38) and (2.39) in (2.35) and considering Δ^2 defined by (2.3) yields

$$\text{Var}[\tilde{D}] = p(1-q^2) + \frac{n_2}{n_2+1} \Delta^2. \quad (2.40)$$

The proof of (ii) can be obtained using similar calculations. $\square$

The distribution of $\tilde{D}$ defined in Proposition 2.2 is comparably simpler than the distribution of $\hat{D}$ in Proposition 2.1. In addition, the expectations of $\hat{D}$ and $\tilde{D}$ are the same, thus the proposed approximation $\tilde{D}$ is reasonable. However, it is still difficult to derive useful results from this distributions. Nevertheless, it can be used when n_1 and n_2 tend to infinity. The following theorem summarizes the distribution of $\tilde{D}$ in that case.

Theorem 2.6

Assume $\tilde{D}$ is defined as in Proposition 2.2, and Δ^2 is as in (2.3). For large n_1 and n_2 , $\tilde{D}$ is normally distributed,

$$\begin{cases} \tilde{D} \sim N(\frac{1}{2}\Delta^2, \Delta^2), & \text{if } \mathbf{x} \in \pi_1, \\ \tilde{D} \sim N(-\frac{1}{2}\Delta^2, \Delta^2), & \text{if } \mathbf{x} \in \pi_2, \end{cases} \quad (2.41)$$

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

Proof: As $n_i \rightarrow \infty, i \in \{1, 2\}$, we have

$$\frac{n_1}{n_1 + 1} \rightarrow 1, \quad \frac{n_2}{n_2 + 1} \rightarrow 1, \quad \bar{\mathbf{y}} \xrightarrow{p} \boldsymbol{\mu}_1, \quad \bar{\mathbf{z}} \xrightarrow{p} \boldsymbol{\mu}_2, \quad p(1 - q^2) \rightarrow 0. \quad (2.42)$$

Therefore, as $n_i \rightarrow \infty, i \in \{1, 2\}$,

$$\begin{aligned} \tilde{D} &\longrightarrow \frac{1}{2}(\boldsymbol{\mu}_2 - \mathbf{x})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \mathbf{x}) - \frac{1}{2}(\boldsymbol{\mu}_1 - \mathbf{x})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \mathbf{x}) \\ &= \mathbf{x}^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), \end{aligned} \quad (2.43)$$

which is the same as D defined in (2.2) and the proof follows from Theorem 2.1. $\square$

Note that Theorem 2.6 is identical to Theorem 2.1, which is understandable because by the law of large number and central limit theorem for large samples n_1 and n_2 information about the two populations is precise.

Theorem 2.7

Let $\hat{D}$ be given by Proposition 2.1 and $\tilde{D}$ by Proposition 2.2, with $\bar{\mathbf{y}}$ the mean of a sample of size n_1 from $\mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $\bar{\mathbf{z}}$ the mean of a sample of size n_2 from $\mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. The limiting distribution of both $\hat{D}$ and $\tilde{D}$ as $n_1 \rightarrow \infty$ and $n_2 \rightarrow \infty$ is $\mathcal{N}(\frac{1}{2}\Delta^2, \Delta^2)$ if $\mathbf{x}$ is distributed as $\mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $\mathcal{N}(-\frac{1}{2}\Delta^2, \Delta^2)$ if $\mathbf{x}$ is distributed as $\mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$.

Proof: The limiting distribution of $\tilde{D}$ is already proven in Theorem 2.6. Moreover, as $n_i \rightarrow \infty, i \in \{1, 2\}$, $m\mathbf{S}^{-1} \rightarrow \boldsymbol{\Sigma}^{-1}$. Thus the proof for $\hat{D}$ follows Theorem 2.6. $\square$

Hence it follows that the limiting distribution of both classification rules $\tilde{D}$ and $\hat{D}$ follows the distribution of D given by (2.2). This result will be used for the standardization of the two classification rules in Chapter 4.

Edgeworth-type Expansion of the Discriminant Function

SEVERAL researchers such as Anderson (1951, 2003), Okamoto (1963), to mention two, have investigated the distribution of the linear discriminant function. The exact distributions of the discriminant functions $\hat{D}$ and $\tilde{D}$, given in Proposition 2.1 and Proposition 2.2, respectively, are very difficult to use for numerical calculations, as mentioned earlier. Therefore, it is practical to provide reasonable approximations of the distribution of these discriminant functions to get around the problem. In this chapter, following Kollo and von Rosen (2005), we present Edgeworth-type expansions of the two discriminant functions $\hat{D}$ and $\tilde{D}$. With the help of the expansions misclassification probabilities can be obtained. However, it is important to note that an Edgeworth expansion needs not to be a density.

Definition 3.1. The j^{th} -degree Hermite polynomial, $H_j(x, m, \sigma^2)$, for the mean m and the variance $\sigma^2 > 0$ is given by

$$H_j(x, m, \sigma^2) = \frac{(-1)^j}{f_x(x)} \frac{d^j f_x(x)}{dx^j}, \quad j \in \mathbb{N}_0, \quad (3.1)$$

where $f_x(x)$ is the density function of the normal distribution $\mathcal{N}(m, \sigma^2)$, i.e.,

$$f_x(x) = (2\pi)^{-\frac{1}{2}} |\sigma|^{-1} e^{-\frac{1}{2}(\frac{x-m}{\sigma})^2}. \quad (3.2)$$

Theorem 3.1

Consider the case of a standard normal distribution, $\mathcal{N}(0, 1)$. The j^{th} univariate Hermite polynomials $H_j(x, 1)$, $j \in \{0, 1, 2, 3\}$, equal

$$H_0(x, 1) = 1; \quad H_1(x, 1) = x; \quad H_2(x, 1) = x^2 - 1; \quad H_3(x, 1) = x^3 - 3x. \quad (3.3)$$

The proof is given in Appendix B.

The Edgeworth-type expansion of a density presented by Kollo and von Rosen (2005) is shown in the following theorem.

Theorem 3.2

Let d be a random variable with finite first three cumulants, then its density $f_d(x)$ can be presented through the density $f_N(x)$ of the standard normal distribution, $\mathcal{N}(0, 1)$, by the Edgeworth-type expansion

$$f_d(x) = f_N(x) \left\{ 1 + E[d]H_1(x, 1) + \frac{1}{2} (Var[d] - 1 + (E[d])^2) H_2(x, 1) + \frac{1}{6} (c_3[d] + 3(Var[d] - 1)E[d] + (E[d])^3) H_3(x, 1) \right\}, \quad (3.4)$$

where $c_3[\cdot]$ is the third cumulant, $H_i(x, 1)$, $i \in \{1, 2, 3\}$, are obtained from Theorem 3.1 and $f_N(x) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2}$.

Suppose the third cumulant is zero, then inserting the Hermite polynomials $H_i(x, 1)$, $i \in \{1, 2, 3\}$, and the density of $\mathcal{N}(0, 1)$ by their values, (3.4) can be rewritten as

$$\begin{aligned} f_d(x) &= f_N(x) \left\{ 1 - \frac{1}{2} (Var[d] - 1 + (E[d])^2) + \left(E[d] - \frac{1}{2} (3Var[d] - 1)E[d] + (E[d])^3 \right) x \right. \\ &\quad \left. + \frac{1}{2} (Var[d] - 1 + (E[d])^2) x^2 + \frac{1}{6} (3Var[d] - 1)E[d] + (E[d])^3 x^3 \right\} \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{1}{2} \left\{ 3 - Var[d] - (E[d])^2 - (3Var[d]E[d] - 5E[d] + (E[d])^3) x \right. \\ &\quad \left. + (Var[d] - 1 + (E[d])^2) x^2 + \left((Var[d] - 1)E[d] + \frac{1}{3}(E[d])^3 \right) x^3 \right\}. \end{aligned} \quad (3.5)$$

Moreover $f_d(x)$ in (3.5) can be presented as

$$f_d(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{1}{2} \{ a_0 - a_1x + a_2x^2 + a_3x^3 \}, \quad (3.6)$$

where

$$\begin{aligned} a_0 &= 3 - Var[d] - (E[d])^2, \quad a_1 = 3Var[d]E[d] - 5E[d] + (E[d])^3, \\ a_2 &= Var[d] - 1 + (E[d])^2, \quad a_3 = (Var[d] - 1)E[d] + \frac{1}{3}(E[d])^3. \end{aligned} \quad (3.7)$$

Thus $f_d(x)$ is a polynomial function of x and the coefficients are in terms of the first and second cumulants of d . In Chapter 2 we have found expressions of the first two cumulants of the discriminant function. Hence, an Edgeworth-type expansion, from Theorem 3.2, of the discriminant function can be expressed up to three terms, where the term including $c_3[\cdot]$ will be assumed to be zero. Thereupon, since the discriminant function is one-dimensional, the Edgeworth-type expansions of the densities of $\widehat{D}$ and $\widetilde{D}$ through the density of $\mathcal{N}(0, 1)$ in the next sections will take the form of (3.6), where a_i , $i \in \{0, 1, 2, 3\}$ will be in terms of the expectation and variance of the corresponding discriminant function.

3.1 Edgeworth-type Expansion of the Discriminant Function $\widetilde{D}$

The basic properties of the discriminant function $\widetilde{D}$, defined in Proposition 2.2, such as its expectation and variance are given in Theorem 2.5. In the theorem below, an Edgeworth-

type expansion of $\tilde{D}$ is presented in terms of its mean and variance via the standard normal distribution density.

Theorem 3.3

Let $\tilde{D}$ be the discriminant function defined in Proposition 2.2 and Δ^2 is as in (2.3). Let q be given by (2.21), then the Edgeworth-type expansion, $f_{\tilde{D}}(x)$, if x is classified into π_1 , equals

$$\begin{aligned} f_{\tilde{D}}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{1}{2} \left\{ 3 - p(1 - q^2) - \frac{n_2}{n_2 + 1} \Delta^2 - \frac{1}{4} \left(\frac{n_2}{n_2 + 1} \right)^2 (\Delta^2)^2 \right. \\ - \frac{1}{2} \left((3p(1 - q^2) - 5) \frac{n_2}{n_2 + 1} \Delta^2 + 3 \left(\frac{n_2}{n_2 + 1} \right)^2 (\Delta^2)^2 + \frac{1}{4} \left(\frac{n_2}{n_2 + 1} \right)^3 (\Delta^2)^3 \right) x \\ + \left(-1 + p(1 - q^2) + \frac{n_2}{n_2 + 1} \Delta^2 + \frac{1}{4} \left(\frac{n_2}{n_2 + 1} \right)^2 (\Delta^2)^2 \right) x^2 \\ \left. + \frac{1}{2} \left((p(1 - q^2) - 1) \frac{n_2}{n_2 + 1} \Delta^2 + \left(\frac{n_2}{n_2 + 1} \right)^2 (\Delta^2)^2 + \frac{1}{12} \left(\frac{n_2}{n_2 + 1} \right)^3 (\Delta^2)^3 \right) x^3 \right\}. \end{aligned} \quad (3.8)$$

If $x \in \pi_2$ we have

$$\begin{aligned} f_{\tilde{D}}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{1}{2} \left\{ 3 - p(1 - q^2) - \frac{n_1}{n_1 + 1} \Delta^2 - \frac{1}{4} \left(\frac{n_1}{n_1 + 1} \right)^2 (\Delta^2)^2 \right. \\ + \frac{1}{2} \left((3p(1 - q^2) - 5) \frac{n_1}{n_1 + 1} \Delta^2 + 3 \left(\frac{n_1}{n_1 + 1} \right)^2 (\Delta^2)^2 + \frac{1}{4} \left(\frac{n_1}{n_1 + 1} \right)^3 (\Delta^2)^3 \right) x \\ + \left(-1 + p(1 - q^2) + \frac{n_1}{n_1 + 1} \Delta^2 + \frac{1}{4} \left(\frac{n_1}{n_1 + 1} \right)^2 (\Delta^2)^2 \right) x^2 \\ \left. - \frac{1}{2} \left((p(1 - q^2) - 1) \frac{n_1}{n_1 + 1} \Delta^2 + \left(\frac{n_1}{n_1 + 1} \right)^2 (\Delta^2)^2 + \frac{1}{12} \left(\frac{n_1}{n_1 + 1} \right)^3 (\Delta^2)^3 \right) x^3 \right\}. \end{aligned} \quad (3.9)$$

Proof: Consider the Edgeworth-type expansion of the form (3.6). For $\tilde{D}$ in (2.2) with the first moments defined from Theorem 2.5, suppose a_i , $i \in \{0, 1, 2, 3\}$, are given by equations (3.7). Then, if $x \in \pi_1$ replacing the moments into a_i , $i \in \{0, 1, 2, 3\}$, with their values we get,

$$\begin{aligned} a_0 &= 3 - p(1 - q^2) - \frac{n_2}{n_2 + 1} \Delta^2 - \frac{1}{4} \left(\frac{n_2}{n_2 + 1} \right)^2 (\Delta^2)^2, \\ a_1 &= \frac{1}{2} \left((3p(1 - q^2) - 5) \frac{n_2}{n_2 + 1} \Delta^2 + 3 \left(\frac{n_2}{n_2 + 1} \right)^2 (\Delta^2)^2 + \frac{1}{4} \left(\frac{n_2}{n_2 + 1} \right)^3 (\Delta^2)^3 \right), \\ a_2 &= -1 + p(1 - q^2) + \frac{n_2}{n_2 + 1} \Delta^2 + \frac{1}{4} \left(\frac{n_2}{n_2 + 1} \right)^2 (\Delta^2)^2, \\ a_3 &= \frac{1}{2} \left((p(1 - q^2) - 1) \frac{n_2}{n_2 + 1} \Delta^2 + \left(\frac{n_2}{n_2 + 1} \right)^2 (\Delta^2)^2 + \frac{1}{12} \left(\frac{n_2}{n_2 + 1} \right)^3 (\Delta^2)^3 \right). \end{aligned} \quad (3.10)$$

When $x \in \pi_2$, we have the same values as when $x \in \pi_1$, except that n_1 and n_2 are interchanged and we have opposite signs on coefficients of x and x^3 which establishes the proof. $\square$

According to the results from Theorem 2.6, the background is that the limiting distribution of $\tilde{D}$ is $\mathcal{N}(\frac{1}{2}\Delta^2, \Delta^2)$ if x is from π_1 and is $\mathcal{N}(-\frac{1}{2}\Delta^2, \Delta^2)$ if x is from π_2 .

In the following theorem we evaluate the Edgeworth-type expansion of the standardized discriminant function with respect to its limiting distribution.

Theorem 3.4

Let $\tilde{D}$ be the discriminant function defined in Proposition 2.2 and Δ^2 is as in (2.3). Let $y = \frac{\tilde{D} - \frac{1}{2}\Delta^2}{\Delta}$ and q be given by (2.21), then the Edgeworth-type expansion, $f_y(x)$, if x is classified into π_1 , equals

$$\begin{aligned}
 f_y(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{1}{2} \left\{ \frac{1}{\Delta^2} p(1-q^2) + 3 - \frac{n_2}{n_2+1} - \frac{1}{2}\Delta^2 \left(\left(\frac{1}{2} \frac{n_2}{n_2+1} \right)^2 - \frac{n_2}{n_2+1} + \frac{1}{2} \right) \right. \\
 - \left[\frac{3}{2} \frac{1}{\Delta} p(1-q^2) \left(\frac{n_2}{n_2+1} - 1 \right) + \Delta \left(\frac{3}{2} \left(\frac{n_2}{n_2+1} \right)^2 - 4 \frac{n_2}{n_2+1} + \frac{5}{2} \right) \right. \\
 \left. \left. + \frac{1}{8} \Delta^3 \left(\left(\frac{n_2}{n_2+1} \right)^3 + 3 \left(\frac{n_2}{n_2+1} \right)^2 + 3 \frac{n_2}{n_2+1} - 1 \right) \right] x \right. \\
 - \left[\frac{1}{\Delta^2} p(1-q^2) + 1 - \frac{n_2}{n_2+1} - \frac{1}{2}\Delta^2 \left(\frac{1}{2} \left(\frac{n_2}{n_2+1} \right)^2 - \frac{n_2}{n_2+1} + \frac{1}{2} \right) \right] x^2 \\
 \left. + \left[\frac{1}{2} \frac{1}{\Delta} p(1-q^2) \left(\frac{n_2}{n_2+1} - 1 \right) + \Delta \left(\frac{1}{2} \left(\frac{n_2}{n_2+1} \right)^3 - \frac{n_2}{n_2+1} + \frac{1}{2} \right) \right. \right. \\
 \left. \left. + \frac{1}{24} \Delta^3 \left(\left(\frac{n_2}{n_2+1} \right)^3 - 3 \left(\frac{n_2}{n_2+1} \right)^2 + 3 \frac{n_2}{n_2+1} - 1 \right) \right] x^3 \right\}. \tag{3.11}
 \end{aligned}$$

If $x \in \pi_2$ we have

$$\begin{aligned}
 f_y(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{1}{2} \left\{ \frac{1}{\Delta^2} p(1-q^2) + 3 - \frac{n_2}{n_2+1} - \frac{1}{2}\Delta^2 \left(\left(\frac{1}{2} \frac{n_2}{n_2+1} \right)^2 - \frac{n_2}{n_2+1} + \frac{1}{2} \right) \right. \\
 + \left[\frac{3}{2} \frac{1}{\Delta} p(1-q^2) \left(\frac{n_2}{n_2+1} - 1 \right) + \Delta \left(\frac{3}{2} \left(\frac{n_2}{n_2+1} \right)^2 - 4 \frac{n_2}{n_2+1} + \frac{5}{2} \right) \right. \\
 \left. \left. + \frac{1}{8} \Delta^3 \left(\left(\frac{n_2}{n_2+1} \right)^3 + 3 \left(\frac{n_2}{n_2+1} \right)^2 + 3 \frac{n_2}{n_2+1} - 1 \right) \right] x \right. \\
 - \left[\frac{1}{\Delta^2} p(1-q^2) + 1 - \frac{n_2}{n_2+1} - \frac{1}{2}\Delta^2 \left(\frac{1}{2} \left(\frac{n_2}{n_2+1} \right)^2 - \frac{n_2}{n_2+1} + \frac{1}{2} \right) \right] x^2 \\
 - \left[\frac{1}{2} \frac{1}{\Delta} p(1-q^2) \left(\frac{n_2}{n_2+1} - 1 \right) + \Delta \left(\frac{1}{2} \left(\frac{n_2}{n_2+1} \right)^3 - \frac{n_2}{n_2+1} + \frac{1}{2} \right) \right. \\
 \left. \left. + \frac{1}{24} \Delta^3 \left(\left(\frac{n_2}{n_2+1} \right)^3 - 3 \left(\frac{n_2}{n_2+1} \right)^2 + 3 \frac{n_2}{n_2+1} - 1 \right) \right] x^3 \right\}. \tag{3.12}
 \end{aligned}$$

Proof: Given the standardized $y = \frac{\tilde{D} - \frac{1}{2}\Delta^2}{\Delta}$ then its expectation and variance are $E(y) = \frac{1}{\Delta} E[\tilde{D}] - \frac{1}{2}\Delta = \frac{1}{2}\Delta \left(\frac{n_2}{n_2+1} - 1 \right)$ and $Var[y] = \frac{1}{\Delta^2} Var[\tilde{D}] = \frac{1}{\Delta^2} p(1-q^2) + \frac{n_2}{n_2+1}$, where $E[\tilde{D}]$ and $Var[\tilde{D}]$ are given by Theorem 2.5. Now consider the Edgeworth-type expansion of the form (3.6). Therefore, if $x \in \pi_1$, substituting the mean and variance of

y into a_i , $i \in \{0, 1, 2, 3\}$ given by (3.7) yields

$$\begin{aligned}
 a_0 &= 3 - \frac{1}{\Delta^2} \text{Var}[\tilde{D}] - \left(\frac{1}{\Delta^2} (E[\tilde{D}])^2 - E[\tilde{D}] - \frac{1}{4} \Delta^2 \right), \\
 &= \frac{1}{\Delta^2} p(1 - q^2) + 3 - \frac{n_2}{n_2 + 1} - \frac{1}{2} \Delta^2 \left(\left(\frac{1}{2} \frac{n_2}{n_2 + 1} \right)^2 - \frac{n_2}{n_2 + 1} + \frac{1}{2} \right) \\
 a_1 &= 3 \frac{1}{\Delta^2} \text{Var}[\tilde{D}] \left(\frac{1}{\Delta} E[\tilde{D}] - \frac{1}{2} \Delta \right) - 5 \left(\frac{1}{\Delta} - \frac{1}{2} \Delta \right) \\
 &\quad + \left(\frac{1}{\Delta^3} ([\tilde{D}])^3 - \frac{3}{2} \frac{1}{\Delta} (E[\tilde{D}])^2 + \frac{3}{4} E[\tilde{D}] \Delta - \frac{1}{8} \Delta^3 \right) \\
 &= \frac{3}{2} \frac{1}{\Delta} p(1 - q^2) \left(\frac{n_2}{n_2 + 1} - 1 \right) + \Delta \left(\frac{3}{2} \left(\frac{n_2}{n_2 + 1} \right)^2 - 4 \frac{n_2}{n_2 + 1} + \frac{5}{2} \right) \\
 &\quad + \frac{1}{8} \Delta^3 \left(\left(\frac{n_2}{n_2 + 1} \right)^3 + 3 \left(\frac{n_2}{n_2 + 1} \right)^2 + 3 \frac{n_2}{n_2 + 1} - 1 \right), \\
 a_2 &= 2 - a_0 \\
 &= \frac{1}{\Delta^2} p(1 - q^2) + 1 - \frac{n_2}{n_2 + 1} - \frac{1}{2} \Delta^2 \left(\frac{1}{2} \left(\frac{n_2}{n_2 + 1} \right)^2 - \frac{n_2}{n_2 + 1} + \frac{1}{2} \right), \\
 a_3 &= \frac{2}{3} E[y] + \frac{1}{3} a_1 = \frac{2}{3} \left(\frac{1}{\Delta} E[\tilde{D}] - \frac{1}{2} \right) \\
 &= \frac{1}{2} \frac{1}{\Delta} p(1 - q^2) \left(\frac{n_2}{n_2 + 1} - 1 \right) + \Delta \left(\frac{1}{2} \left(\frac{n_2}{n_2 + 1} \right)^3 - \frac{n_2}{n_2 + 1} + \frac{1}{2} \right) \\
 &\quad + \frac{1}{24} \Delta^3 \left(\left(\frac{n_2}{n_2 + 1} \right)^3 - 3 \left(\frac{n_2}{n_2 + 1} \right)^2 + 3 \frac{n_2}{n_2 + 1} - 1 \right). \tag{3.13}
 \end{aligned}$$

Which establishes the expression in (3.11). Similar calculations if $\mathbf{x} \in \pi_2$ yield the expression in (3.12), which concludes the proof. $\square$

3.2 Edgeworth-type Expansion of the discriminant function $\hat{D}$

The classification function $\hat{D}$ given in Proposition 2.1 follows a complicated distribution since it is a difference between two non-independent non-central F-distributions. In this section, an Edgeworth-type expansion of the density of $\hat{D}$ is provided in order to get around this problem in terms of its expected value and variance via a standard normal distribution density. Let $f_{\hat{D}}(x)$ be such an expansion. Then since $\hat{D}$ is one-dimensional, $f_{\hat{D}}$ will be a polynomial function of the form (3.6). In the following theorem the expansion is presented.

Theorem 3.5

Let $\hat{D}$ be the discriminant function defined in Proposition 2.1 and Δ^2 is as in (2.3). Then, for $m = n_1 + n_2 - p - 3$, q in (2.21) and c_0 in (2.23), the Edgeworth-type expansion,

$f_{\hat{D}}(x)$ when the observation $\mathbf{x}$ is classified in population π_1 equals

$$\begin{aligned}
 f_{\hat{D}}(x) = & \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{1}{2} \left\{ 3 - (1 - q^2)m^2 c_0 p - m^2 \frac{n_2}{n_2 + 1} c_0 \Delta^2 + \left(\frac{n_2}{n_2 + 1} \right)^2 \frac{m}{4(m-2)} (\Delta^2)^2 \right. \\
 & - \frac{1}{2} \left[(3(1 - q^2)m^2 c_0 p - 5) \frac{n_2}{n_2 + 1} \Delta^2 + 3m^2 \left(\frac{n_2}{n_2 + 1} \right)^2 c_0 (\Delta^2)^2 \right. \\
 & \left. \left. + \left(\frac{n_2}{n_2 + 1} \right)^3 \frac{m+4}{4(m-2)} (\Delta^2)^3 \right] x + \left[(1 - q^2)m^2 c_0 p - 1 \right. \right. \\
 & \left. + m^2 \frac{n_2}{n_2 + 1} c_0 \Delta^2 - \left(\frac{n_2}{n_2 + 1} \right)^2 \frac{m}{4(m-2)} (\Delta^2)^2 \right] x^2 \\
 & + \frac{1}{2} \left[((1 - q^2)m^2 c_0 p - 1) \frac{n_2}{n_2 + 1} \Delta^2 + m^2 \left(\frac{n_2}{n_2 + 1} \right)^2 c_0 (\Delta^2)^2 \right. \\
 & \left. \left. + \left(\frac{n_2}{n_2 + 1} \right)^3 \frac{m+4}{12(m-2)} (\Delta^2)^3 \right] x^3 \right\}, \tag{3.14}
 \end{aligned}$$

Moreover, when $\mathbf{x} \in \pi_2$ we have

$$\begin{aligned}
 f_{\hat{D}}(x) = & \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{1}{2} \left\{ 3 - (1 - q^2)m^2 c_0 p - m^2 \frac{n_1}{n_1 + 1} c_0 \Delta^2 + \left(\frac{n_1}{n_1 + 1} \right)^2 \frac{m}{4(m-2)} (\Delta^2)^2 \right. \\
 & + \frac{1}{2} \left[(3(1 - q^2)m^2 c_0 p - 5) \frac{n_1}{n_1 + 1} \Delta^2 + 3m^2 \left(\frac{n_1}{n_1 + 1} \right)^2 c_0 (\Delta^2)^2 \right. \\
 & \left. + \left(\frac{n_1}{n_1 + 1} \right)^3 \frac{m+4}{4(m-2)} (\Delta^2)^3 \right] x + \left[(1 - q^2)m^2 c_0 p - 1 \right. \\
 & \left. + m^2 \frac{n_1}{n_1 + 1} c_0 \Delta^2 - \left(\frac{n_1}{n_1 + 1} \right)^2 \frac{m}{4(m-2)} (\Delta^2)^2 \right] x^2 \\
 & - \frac{1}{2} \left[((1 - q^2)m^2 c_0 p - 1) \frac{n_1}{n_1 + 1} \Delta^2 + m^2 \left(\frac{n_1}{n_1 + 1} \right)^2 c_0 (\Delta^2)^2 \right. \\
 & \left. + \left(\frac{n_1}{n_1 + 1} \right)^3 \frac{m+4}{12(m-2)} (\Delta^2)^3 \right] x^3 \right\}. \tag{3.15}
 \end{aligned}$$

Proof: If $\mathbf{x} \in \pi_1$, for $m = n_1 + n_2 - p - 3$, consider the form (3.6) of the Edgeworth-type expansion, where $\hat{D}$ is defined as (2.1) with expectation (2.18) and variance (2.22). Substituting the cumulants into a_i , $i \in \{0, 1, 2, 3\}$ given by (3.7), we get

$$\begin{aligned}
 a_0 = & 3 - (1 - q^2)m^2 c_0 p - m^2 \frac{n_2}{n_2 + 1} c_0 \Delta^2 + \left(\frac{n_2}{n_2 + 1} \right)^2 \frac{m}{4(m-2)} (\Delta^2)^2, \\
 a_1 = & \frac{1}{2} \left\{ (3(1 - q^2)m^2 c_0 p - 5) \frac{n_2}{n_2 + 1} \Delta^2 + 3m^2 \left(\frac{n_2}{n_2 + 1} \right)^2 c_0 (\Delta^2)^2 \right. \\
 & \left. + \left(\frac{n_2}{n_2 + 1} \right)^3 \frac{m+4}{4(m-2)} (\Delta^2)^3 \right\}, \\
 a_2 = & (1 - q^2)m^2 c_0 p - 1 + m^2 \frac{n_2}{n_2 + 1} c_0 \Delta^2 - \left(\frac{n_2}{n_2 + 1} \right)^2 \frac{m}{4(m-2)} (\Delta^2)^2, \\
 a_3 = & \frac{1}{2} \left[((1 - q^2)m^2 c_0 p - 1) \frac{n_2}{n_2 + 1} \Delta^2 + m^2 \left(\frac{n_2}{n_2 + 1} \right)^2 c_0 (\Delta^2)^2 \right. \\
 & \left. + \left(\frac{n_2}{n_2 + 1} \right)^3 \frac{m+4}{12(m-2)} (\Delta^2)^3 \right]. \tag{3.16}
 \end{aligned}$$

Similarly, substituting the results from Theorem 2.3 and Theorem 2.4 into a_i , $i \in \{0, 1, 2, 3\}$, given by (3.7) when $\mathbf{x} \in \pi_2$ concludes the proof. $\square$

Moreover, from Theorem 2.7, the limiting distribution of $\hat{D}$ is $\mathcal{N}(\frac{1}{2}\Delta^2, \Delta^2)$ if $\mathbf{x}$ is said to come from π_1 and is $\mathcal{N}(-\frac{1}{2}\Delta^2, \Delta^2)$ if $\mathbf{x}$ is said to come from π_2 . The following theorem depicts the standardized Edgeworth-type expansion of $\hat{D}$.

Theorem 3.6

Let $\hat{D}$ be the discriminant function defined in Proposition 2.1 and Δ^2 is as in (2.3) and let $z = \frac{\hat{D} - \frac{1}{2}\Delta^2}{\Delta}$. Then, for $m = n_1 + n_2 - p - 3$, q in (2.21) and c_0 in (2.23), the Edgeworth-type expansion, $f_z(x)$ when the observation $\mathbf{x}$ is classified in population π_1 equals

$$\begin{aligned}
 f_z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{1}{2} \Big\{ & -\frac{1}{\Delta^2} (1-q^2) m^2 c_0 p + 3 - m^2 c_0 \frac{n_2}{n_2+1} \\
 & - \frac{1}{2} \Delta^2 \left(\left(\frac{n_2}{n_2+1} \right)^2 \frac{m}{2(m-2)} - \frac{n_2}{n_2+1} + \frac{1}{2} \right) \\
 & - \frac{1}{2} \left[\frac{1}{\Delta} 3(1-q^2) m^2 c_0 p \left(\frac{n_2}{n_2+1} - 1 \right) + \Delta \left(3m^2 c_0 \frac{n_2}{n_2+1} - 5 \right) \right. \\
 & \times \left(\frac{n_2}{n_2+1} - 1 \right) + \frac{1}{4} \Delta^3 \left(3 \left(\frac{n_2}{n_2+1} \right)^3 \frac{1}{m-2} - 3 \left(\frac{n_2}{n_2+1} \right)^2 \frac{1}{m-2} \right. \\
 & \left. \left. + \frac{n_2}{n_2+1} - 1 \right) \right] x + \left[\frac{1}{\Delta^2} (1-q^2) m^2 c_0 p - 1 + m^2 c_0 \frac{n_2}{n_2+1} \right. \\
 & \left. + \frac{1}{2} \Delta^2 \left(\left(\frac{n_2}{n_2+1} \right)^2 \frac{m}{2(m-2)} - \frac{n_2}{n_2+1} + \frac{1}{2} \right) \right] x^2 \\
 & + \frac{1}{2} \left[\frac{1}{\Delta} (1-q^2) m^2 c_0 p \left(\frac{n_2}{n_2+1} - 1 \right) + \Delta \left(m^2 c_0 \frac{n_2}{n_2+1} - \frac{4}{3} \right) \right. \\
 & \times \left(\frac{n_2}{n_2+1} - 1 \right) + \frac{1}{4} \Delta^3 \left(\left(\frac{n_2}{n_2+1} \right)^2 \frac{1}{m-2} \left(\frac{n_2}{n_2+1} - 1 \right) \right. \\
 & \left. \left. + \frac{1}{3} \frac{n_2}{n_2+1} - \frac{1}{3} \right) \right] x^3 \Big\}, \tag{3.17}
 \end{aligned}$$

Moreover, when $\mathbf{x} \in \pi_2$ we have

$$\begin{aligned}
 f_z(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \frac{1}{2} \Bigg\{ & -\frac{1}{\Delta^2} (1-q^2) m^2 c_0 p + 3 - m^2 c_0 \frac{n_1}{n_1+1} \\
 & - \frac{1}{2} \Delta^2 \left(\left(\frac{n_1}{n_1+1} \right)^2 \frac{m}{2(m-2)} - \frac{n_1}{n_1+1} + \frac{1}{2} \right) \\
 & - \frac{1}{2} \left[\frac{1}{\Delta} 3(1-q^2) m^2 c_0 p \left(\frac{n_1}{n_1+1} - 1 \right) + \Delta \left(3m^2 c_0 \frac{n_1}{n_1+1} - 5 \right) \right. \\
 & \times \left(\frac{n_1}{n_1+1} - 1 \right) + \frac{1}{4} \Delta^3 \left(3 \left(\frac{n_1}{n_1+1} \right)^3 \frac{1}{m-2} - 3 \left(\frac{n_1}{n_1+1} \right)^2 \frac{1}{m-2} \right. \\
 & \left. \left. + \frac{n_1}{n_1+1} - 1 \right) \right] x + \left[\frac{1}{\Delta^2} (1-q^2) m^2 c_0 p - 1 + m^2 c_0 \frac{n_1}{n_1+1} \right. \\
 & \left. + \frac{1}{2} \Delta^2 \left(\left(\frac{n_1}{n_1+1} \right)^2 \frac{m}{2(m-2)} - \frac{n_1}{n_1+1} + \frac{1}{2} \right) \right] x^2 \\
 & + \frac{1}{2} \left[\frac{1}{\Delta} (1-q^2) m^2 c_0 p \left(\frac{n_1}{n_1+1} - 1 \right) + \Delta \left(m^2 c_0 \frac{n_1}{n_1+1} - \frac{4}{3} \right) \right. \\
 & \times \left(\frac{n_1}{n_1+1} - 1 \right) + \frac{1}{4} \Delta^3 \left(\left(\frac{n_1}{n_1+1} \right)^2 \frac{1}{m-2} \left(\frac{n_1}{n_1+1} - 1 \right) \right. \\
 & \left. \left. + \frac{1}{3} \frac{n_1}{n_1+1} - \frac{1}{3} \right) \right] x^3 \Bigg\}, \tag{3.18}
 \end{aligned}$$

Proof: Consider $E[\hat{D}]$ given by Theorem 2.3 and $Var[\hat{D}]$ by Theorem 2.4. If $\mathbf{x} \in \pi_1$ then,

$$E[z] = \frac{1}{\Delta} E[\hat{D}] - \frac{1}{2} \Delta = \frac{1}{2} \Delta \left(\frac{n_2}{n_2+1} - 1 \right)$$

and

$$\begin{aligned}
 Var[z] &= \frac{1}{\Delta^2} Var[\hat{D}] \\
 &= \frac{1}{\Delta^2} (1-q^2) m^2 c_0 p + m^2 c_0 \frac{n_2}{n_2+1} + \left(\frac{n_2}{n_2+1} \right)^2 \frac{1}{2(m-2)} \Delta^2.
 \end{aligned}$$

Now consider the Edgeworth-type expansion of the form (3.6). Hence, if $\mathbf{x} \in \pi_1$, replacing the expected value and variance of z in a_i , $i \in \{0, 1, 2, 3\}$ given by (3.7) by their values yields

$$\begin{aligned}
 a_0 &= 3 - Var[z] - \left(E[z] \right)^2 \\
 &= -\frac{1}{\Delta^2} (1-q^2) m^2 c_0 p + 3 - m^2 c_0 \frac{n_2}{n_2+1} - \frac{1}{2} \Delta^2 \left(\left(\frac{n_2}{n_2+1} \right)^2 \frac{m}{2(m-2)} - \frac{n_2}{n_2+1} + \frac{1}{2} \right),
 \end{aligned}$$

$$\begin{aligned}
a_1 &= 3\text{Var}[z]E[z] - 5E[z] + \left(E[z]\right)^2 \\
&= \frac{1}{\Delta} \frac{3}{2} (1 - q^2) m^2 c_0 p \left(\frac{n_2}{n_2 + 1} - 1 \right) + \frac{1}{2} \Delta \left(3m^2 c_0 \frac{n_2}{n_2 + 1} - 5 \right) \left(\frac{n_2}{n_2 + 1} - 1 \right) \\
&\quad + \frac{1}{4} \Delta^3 \left(3 \left(\frac{n_2}{n_2 + 1} \right)^3 \frac{1}{m-2} - 3 \left(\frac{n_2}{n_2 + 1} \right)^2 \frac{1}{m-2} + \frac{n_2}{n_2 + 1} - 1 \right), \\
a_2 &= 2 - a_0 \\
&= \frac{1}{\Delta^2} (1 - q^2) m^2 c_0 p - 1 + m^2 c_0 \frac{n_2}{n_2 + 1} + \frac{1}{2} \Delta^2 \left(\left(\frac{n_2}{n_2 + 1} \right)^2 \frac{m}{2(m-2)} - \frac{n_2}{n_2 + 1} + \frac{1}{2} \right), \\
a_3 &= \frac{2}{3} E[z] + \frac{1}{3} a_1 \\
&= \frac{1}{\Delta} \frac{1}{2} (1 - q^2) m^2 c_0 p \left(\frac{n_2}{n_2 + 1} - 1 \right) + \Delta \frac{1}{2} \left(m^2 c_0 \frac{n_2}{n_2 + 1} - \frac{4}{3} \right) \left(\frac{n_2}{n_2 + 1} - 1 \right) \\
&\quad + \frac{1}{4} \Delta^3 \left(\left(\frac{n_2}{n_2 + 1} \right)^2 \frac{1}{m-2} \left(\frac{n_2}{n_2 + 1} - 1 \right) + \frac{1}{3} \frac{n_2}{n_2 + 1} - \frac{1}{3} \right). \tag{3.19}
\end{aligned}$$

Which establishes the expression in (3.17). Similar calculations if $\mathbf{x} \in \pi_2$ yield the expression in (3.18), which concludes the proof. $\square$

Misclassification Errors

IT is not enough to make a decision based on a classification rule. In addition to a classification rule, it is important to study probabilities of misclassifications in order to know if a classification approach is reliable. The probability of misclassification is a measure of the goodness of the proposed classification rule as stated by Fujikoshi et al. (2011). The classification of the new observation $\mathbf{x}$ depends on whether the discriminant function is greater than zero or not. Denote by $P(\mathbf{x} \rightarrow \pi_i)$ the probability that an observation $\mathbf{x}$ is classified into population π_i . Let $i, j \in \{1, 2\}$, then $P(\mathbf{x} \rightarrow \pi_i | \mathbf{x} \in \pi_i)$ is the probability of correctly classifying the observation $\mathbf{x}$ into population π_i and $P(\mathbf{x} \rightarrow \pi_j | \mathbf{x} \in \pi_i)$ is the probability that the observation $\mathbf{x}$ comes from population π_i but is classified as coming from π_j . In this section, we investigate misclassification errors of the linear classification D in (2.2), the quadratic classifications $\hat{D}$ given in Proposition 2.1 and $\tilde{D}$ in Proposition 2.2. The probabilities of misclassifications for the classification functions $\hat{D}$ and $\tilde{D}$ can be estimated by integrating the Edgeworth-type expansions of the discriminant functions through the density of the standard normal distribution. The next theorem is well known.

Theorem 4.1

Consider the linear classification function given by (2.2). Let $\Phi(\cdot)$ be the standard normal cumulative distribution function. Then,

$$P(\mathbf{x} \rightarrow \pi_2 | \mathbf{x} \in \pi_1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\frac{1}{2}\Delta} e^{-\frac{1}{2}x^2} dx = \Phi\left(-\frac{1}{2}\Delta\right), \quad (4.1)$$

$$P(\mathbf{x} \rightarrow \pi_1 | \mathbf{x} \in \pi_2) = \frac{1}{\sqrt{2\pi}} \int_{\frac{1}{2}\Delta}^{\infty} e^{-\frac{1}{2}x^2} dx = 1 - \Phi\left(\frac{1}{2}\Delta\right). \quad (4.2)$$

Proof: We have that when $\mathbf{x} \in \pi_1$, D is distributed according to $\mathcal{N}(\frac{1}{2}\Delta^2, \Delta^2)$ and ac-

according to $\mathcal{N}(-\frac{1}{2}\Delta^2, \Delta^2)$ when $\mathbf{x} \in \pi_2$. Then,

$$\begin{aligned}
 P(\mathbf{x} \rightarrow \pi_2 | \mathbf{x} \in \pi_1) &= P(D < 0 | \mathbf{x} \in \pi_1) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^0 \frac{1}{\Delta} e^{-\frac{1}{2}(D - \frac{1}{2}\Delta^2)^2 / \Delta^2} dD \\
 &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\frac{1}{2}\Delta} e^{-\frac{1}{2}x^2} dx = \Phi(-\frac{1}{2}\Delta), \\
 P(\mathbf{x} \rightarrow \pi_1 | \mathbf{x} \in \pi_2) &= P(D \geq 0 | \mathbf{x} \in \pi_2) = \frac{1}{\sqrt{2\pi}} \int_0^{\infty} \frac{1}{\Delta} e^{-\frac{1}{2}(D + \frac{1}{2}\Delta^2)^2 / \Delta^2} dD \\
 &= \frac{1}{\sqrt{2\pi}} \int_{\frac{1}{2}\Delta}^{\infty} e^{-\frac{1}{2}x^2} dx = 1 - \Phi(\frac{1}{2}\Delta).
 \end{aligned}$$

□

We can see from the expressions in Theorem 4.1 that the misclassification errors are monotone functions of the Mahalanobis distance Δ . Now, let us consider the discriminant function with unknown means and co-variance. As we observed before, the exact distributions of $\hat{D}$ and $\tilde{D}$ are complicated to calculate. Therefore, one would like to approximate the misclassification errors to know how good these classification rules are.

Okamoto (1963) and Memon and Okamoto (1971) investigated asymptotic expansions for the discriminant function distribution up to terms of order n^{-2} , $n = n_1 + n_2 - 2$, Siotani and Wang (1975); Siotani (1977) and Anderson (2003) extended the expansions to terms of $O(n^{-3})$ the asymptotic expansions of misclassification errors for the distribution of the discriminant functions when n_1 and n_2 grow to infinity and the ratio n_1/n_2 approaches a positive limit. From Theorem 2.7 we assume the asymptotic distribution of $\hat{D}$ and $\tilde{D}$ to be $\mathcal{N}(\frac{1}{2}\Delta^2, \Delta^2)$ if $\mathbf{x}$ comes from π_2 and to be $\mathcal{N}(-\frac{1}{2}\Delta^2, \Delta^2)$ if $\mathbf{x}$ comes from π_1 .

Lemma 4.1

Let a_i , $i \in \{0, 1, 2, 3\}$ be given by (3.7) and Δ in (2.3). Let $\Phi(\cdot)$ and $\phi(\cdot)$ be the cumulative distribution function and density of $\mathcal{N}(0, 1)$, respectively. For the discriminant function D , where D is either $\hat{D}$ defined in Proposition 2.1 or $\tilde{D}$ defined in Proposition 2.2

$$P\left\{\frac{D - \frac{1}{2}\Delta^2}{\Delta} \leq -\frac{1}{2}\Delta | \mathbf{x} \in \pi_1\right\} \approx \Phi(-\frac{1}{2}\Delta) + \phi(\frac{1}{2}\Delta) \left(\frac{1}{2}a_1 + \frac{1}{4}\Delta a_2 - (1 + \frac{1}{8}\Delta^2)a_3\right), \quad (4.3)$$

and $P\left\{-\frac{D + \frac{1}{2}\Delta^2}{\Delta} \leq -\frac{1}{2}\Delta | \mathbf{x} \in \pi_2\right\}$ is the same but with opposite signs on coefficients of a_1 and a_3 .

Remark 4.1. Note that a_1 , a_2 and a_3 depend on if $\hat{D}$ or $\tilde{D}$ is approximated.

Proof: Consider the Edgeworth-type expansion of the form (3.6) and put $t = \frac{D - \frac{1}{2}\Delta^2}{\Delta}$. Then, performing partial integration

$$\begin{aligned}
 P(\mathbf{x} \rightarrow \pi_2 | \mathbf{x} \in \pi_1) &= P\left\{\frac{D - \frac{1}{2}\Delta^2}{\Delta} \leq -\frac{1}{2}\Delta | \mathbf{x} \in \pi_1\right\} = \int_{-\infty}^{-\frac{1}{2}\Delta} f_t(x) dx \\
 &\approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\frac{1}{2}\Delta} e^{-\frac{1}{2}x^2} \frac{1}{2} \{a_0 - a_1x + a_2x^2 + a_3x^3\} dx \\
 &= \frac{1}{2}a_0\Phi\left(-\frac{1}{2}\Delta\right) + \frac{1}{2\sqrt{2\pi}}a_1e^{-\frac{1}{8}\Delta^2} + \frac{1}{2\sqrt{2\pi}}a_2\left[\frac{1}{2}\Delta e^{-\frac{1}{8}\Delta^2} \right. \\
 &\quad \left. + \frac{\sqrt{2\pi}}{2}(2\Phi\left(-\frac{1}{2}\right) - 1)\right] + \frac{1}{4}a_2 - \frac{1}{2\sqrt{2\pi}}a_3e^{-\frac{1}{8}\Delta^2}\left(-2 - \frac{\Delta^2}{4}\right) \\
 &= \frac{1}{2}a_0\Phi\left(-\frac{1}{2}\Delta\right) + \frac{1}{2}a_1\phi\left(-\frac{1}{2}\Delta\right) + \frac{1}{2}a_2\left(\frac{1}{2}\Delta\phi\left(-\frac{1}{2}\Delta\right) + \Phi\left(-\frac{1}{2}\Delta\right)\right) \\
 &\quad - a_3\left(1 + \frac{1}{8}\Delta^2\right)\phi\left(-\frac{1}{2}\Delta\right) \\
 &= \Phi\left(-\frac{1}{2}\Delta\right)\frac{1}{2}(a_0 + a_2) + \phi\left(-\frac{1}{2}\Delta\right)\left(\frac{1}{2}a_1 + \frac{1}{4}\Delta a_2 - \left(1 + \frac{1}{8}\Delta^2\right)a_3\right) \\
 &= \Phi\left(-\frac{1}{2}\Delta\right) + \phi\left(\frac{1}{2}\Delta\right)\left(\frac{1}{2}a_1 + \frac{1}{4}\Delta a_2 - \left(1 + \frac{1}{8}\Delta^2\right)a_3\right),
 \end{aligned}$$

since

$$\frac{1}{2}(a_0 + a_2) = \frac{1}{2}\left(3 - \text{Var}(t) - (E(t))^2 + \text{Var}(t) - 1 + (E(t))^2\right) = 1,$$

□

The classification rule using $\tilde{D}$ is to assign $\mathbf{x}$ to π_1 if $\tilde{D} \geq 0$ and to π_2 if $\tilde{D} < 0$. The probabilities of misclassification are given by Theorem 4.2 below.

Theorem 4.2

Consider the classification function, $\tilde{D}$, given by (2.2). Let Δ^2 be given by (2.3) and let $\Phi(\cdot)$ and $\phi(\cdot)$ be the cumulative distribution function and density of $\mathcal{N}(0, 1)$, respectively. Then, with q given by (2.21),

$$\begin{aligned}
P(\mathbf{x} \rightarrow \pi_2 | \mathbf{x} \in \pi_1) &\approx \Phi\left(-\frac{1}{2}\Delta\right) + \frac{1}{2}\phi\left(\frac{1}{2}\Delta\right) \left\{ \frac{1}{2} \frac{1}{\Delta} \frac{n_2}{n_2+1} p(1-q^2) + \frac{3}{2} \left(\frac{n_2}{n_2+1}\right)^2 - 4 \frac{n_2}{n_2+1} + \frac{5}{2} \right. \\
&\quad - \Delta \left(\left(\frac{n_2}{n_2+1}\right)^3 - \frac{3}{2} \frac{n_2}{n_2+1} - \frac{1}{16} p(1-q^2) \left(\frac{n_2}{n_2+1} - 1\right) + \frac{1}{2} \right) \\
&\quad + \frac{1}{2} \Delta^3 \left(\frac{1}{3} \left(\frac{n_2}{n_2+1}\right)^3 + \left(\frac{n_2}{n_2+1}\right)^2 - \frac{1}{2} \frac{n_2}{n_2+1} - \frac{3}{4} \right) \\
&\quad \left. + \frac{1}{64} \Delta^5 \left(\left(\frac{n_2}{n_2+1}\right)^3 - 3 \left(\frac{n_2}{n_2+1}\right)^2 + 3 \frac{n_2}{n_2+1} - 1 \right) \right\}, \tag{4.4}
\end{aligned}$$

$$\begin{aligned}
P(\mathbf{x} \rightarrow \pi_1 | \mathbf{x} \in \pi_2) &\approx \Phi\left(-\frac{1}{2}\Delta\right) + \frac{1}{4}\phi\left(\frac{1}{2}\Delta\right) \left\{ \frac{1}{2} \frac{1}{\Delta} \frac{n_1}{n_1+1} p(1-q^2) + \frac{3}{2} \left(\frac{n_1}{n_1+1}\right)^2 - 4 \frac{n_1}{n_1+1} + \frac{5}{2} \right. \\
&\quad - \Delta \left(\left(\frac{n_1}{n_1+1}\right)^3 - \frac{3}{2} \frac{n_1}{n_1+1} - \frac{1}{16} p(1-q^2) \left(\frac{n_1}{n_1+1} - 1\right) + \frac{1}{2} \right) \\
&\quad + \frac{1}{2} \Delta^3 \left(\frac{1}{3} \left(\frac{n_1}{n_1+1}\right)^3 + \left(\frac{n_1}{n_1+1}\right)^2 - \frac{1}{2} \frac{n_1}{n_1+1} - \frac{3}{4} \right) \\
&\quad \left. + \frac{1}{64} \Delta^5 \left(\left(\frac{n_1}{n_1+1}\right)^3 - 3 \left(\frac{n_1}{n_1+1}\right)^2 + 3 \frac{n_1}{n_1+1} - 1 \right) \right\}, \tag{4.5}
\end{aligned}$$

The difference between the two expressions in (4.4) and (4.5) is that n_1 and n_2 are interchanged.

Proof: Consider the misclassification probability in Lemma 4.1 with the constants a_i , $i \in \{1, 2, 3\}$, in (3.13) and the cumulants of $\bar{D}$ from Theorem 2.5. If $\mathbf{x} \in \pi_1$

$$\begin{aligned}
\frac{1}{2}a_1 + \frac{1}{4}\Delta a_2 - (1 + \frac{1}{8}\Delta^2)a_3 &= \frac{3}{4} \frac{1}{\Delta} p(1-q^2) \left(\frac{n_2}{n_2+1} - 1\right) + \frac{1}{2} \left(\frac{3}{2} \left(\frac{n_2}{n_2+1}\right)^2 - 4 \frac{n_2}{n_2+1} + \frac{5}{2} \right) \\
&\quad + \frac{1}{16} \Delta^3 \left(\left(\frac{n_2}{n_2+1}\right)^3 + 3 \left(\frac{n_2}{n_2+1}\right)^2 - 1 \right) + \frac{1}{4\Delta} p(1-q^2) \\
&\quad - \left(\frac{n_2}{n_2+1} - 1 \right) \frac{1}{4} \Delta - \frac{1}{8} \Delta^3 \left(\frac{1}{2} \left(\frac{n_2}{n_2+1}\right)^2 - \frac{n_2}{n_2+1} + \frac{1}{2} \right) \\
&\quad - \frac{1}{2} \frac{1}{\Delta} p(1-q^2) \left(\frac{n_2}{n_2+1} - 1\right) - \Delta \left(\frac{1}{2} \left(\frac{n_2}{n_2+1}\right)^3 - \frac{n_2}{n_2+1} + \frac{1}{2} \right) \\
&\quad - \frac{1}{24} \Delta^3 \left(\left(\frac{n_2}{n_2+1}\right)^3 - 3 \left(\frac{n_2}{n_2+1}\right)^2 + 3 \frac{n_2}{n_2+1} - 1 \right) \\
&\quad + \frac{1}{16} \Delta p(1-q^2) \left(\frac{n_2}{n_2+1} - 1\right) + \frac{1}{8} \Delta^3 \left(\left(\frac{n_2}{n_2+1}\right)^3 + 3 \frac{n_2}{n_2+1} - 1 \right) \\
&\quad + \frac{1}{128} \Delta^5 \left(\left(\frac{n_2}{n_2+1}\right)^3 - 3 \left(\frac{n_2}{n_2+1}\right)^2 + 3 \frac{n_2}{n_2+1} - 1 \right) \\
&= (p(1-q^2) - 1)\Delta + (p(1-q^2) - 3) \frac{n_2}{n_2+1} \Delta^2 + \frac{n_2}{n_2+1} \Delta^3 \\
&\quad + \left(\frac{n_2}{n_2+1} - \frac{1}{2} p(1-q^2) \right) \frac{n_2}{n_2+1} \Delta^4 + \frac{1}{4} \left(\frac{n_2}{n_2+1} \right)^2 \Delta^5 \\
&\quad - \frac{1}{2} \left(\frac{1}{6} \frac{n_2}{n_2+1} + 1 \right) \left(\frac{n_2}{n_2+1} \right)^2 \Delta^6 - \frac{1}{24} \left(\frac{n_2}{n_2+1} \right)^3 \Delta^8.
\end{aligned}$$

Hence the expression in (4.4) follows. Furthermore, similar calculations when $\mathbf{x} \in \pi_2$ conclude the proof. $\square$

The classification rule using $\widehat{D}$ is to classify $\mathbf{x}$ to π_1 if $\widehat{D} \geq 0$ and to π_2 if $\widehat{D} < 0$. The theorem below expresses the misclassification errors for this rule.

Theorem 4.3

Consider the classification function, $\widehat{D}$, given by (2.1) and Δ^2 given by (2.3). Then, for $m = n_1 + n_2 - p - 3$, q in (2.21) and c_0 in (2.23), and let $\Phi(\cdot)$ and $\phi(\cdot)$ be the cumulative distribution function and density of $\mathcal{N}(0, 1)$, respectively. Then,

$$\begin{aligned}
 P(\mathbf{x} \rightarrow \pi_2 | \mathbf{x} \in \pi_1) \approx & \Phi\left(-\frac{1}{2}\Delta\right) + \frac{1}{2}\phi\left(\frac{1}{2}\Delta\right) \left\{ \frac{1}{\Delta} \frac{1}{4}(1-q^2)m^2c_0p \frac{n_2}{n_2+1} \right. \\
 & + \frac{1}{2}\Delta \left[\frac{1}{2}m^2c_0 \left(\frac{n_2}{n_2+1} \right)^2 - \frac{1}{8}(1-q^2)m^2c_0p \left(\frac{n_2}{n_2+1} - 1 \right) \right. \\
 & - \frac{3}{4} \frac{n_2}{n_2+1} + \frac{1}{2} \left] + \frac{1}{2}\Delta^3 \left[\frac{1}{24} \left(\frac{n_2}{n_2+1} \right)^3 \frac{12+m^2(m-2)}{m^2(m-2)} \right. \\
 & - \frac{1}{16} \left(\frac{n_2}{n_2+1} \right)^2 \frac{4+m^2(m-2)}{m^2(m-2)} - \frac{1}{8}m^2c_0 \frac{n_2}{n_2+1} \left(\frac{n_2}{n_2+1} - 1 \right) - \frac{1}{24} \left] \right. \\
 & - \frac{1}{192}\Delta^5 \left[\left(\frac{n_2}{n_2+1} \right)^3 \frac{12+m^2(m-2)}{m^2(m-2)} \right. \\
 & \left. \left. - 3 \left(\frac{n_2}{n_2+1} \right)^2 \frac{4+m^2(m-2)}{m^2(m-2)} + 3 \frac{n_2}{n_2+1} - 1 \right] \right\}, \tag{4.6}
 \end{aligned}$$

$$\begin{aligned}
 P(\mathbf{x} \rightarrow \pi_1 | \mathbf{x} \in \pi_2) \approx & \Phi\left(-\frac{1}{2}\Delta\right) + \frac{1}{2}\phi\left(\frac{1}{2}\Delta\right) \left\{ \frac{1}{\Delta} \frac{1}{4}(1-q^2)m^2c_0p \frac{n_1}{n_1+1} \right. \\
 & + \frac{1}{2}\Delta \left[\frac{1}{2}m^2c_0 \left(\frac{n_1}{n_1+1} \right)^2 - \frac{1}{8}(1-q^2)m^2c_0p \left(\frac{n_1}{n_1+1} - 1 \right) \right. \\
 & - \frac{3}{4} \frac{n_1}{n_1+1} + \frac{1}{2} \left] + \frac{1}{2}\Delta^3 \left[\frac{1}{24} \left(\frac{n_1}{n_1+1} \right)^3 \frac{12+m^2(m-2)}{m^2(m-2)} \right. \\
 & - \frac{1}{16} \left(\frac{n_1}{n_1+1} \right)^2 \frac{4+m^2(m-2)}{m^2(m-2)} - \frac{1}{8}m^2c_0 \frac{n_1}{n_1+1} \left(\frac{n_1}{n_1+1} - 1 \right) - \frac{1}{24} \left] \right. \\
 & - \frac{1}{192}\Delta^5 \left[\left(\frac{n_1}{n_1+1} \right)^3 \frac{12+m^2(m-2)}{m^2(m-2)} \right. \\
 & \left. \left. - 3 \left(\frac{n_1}{n_1+1} \right)^2 \frac{4+m^2(m-2)}{m^2(m-2)} + 3 \frac{n_1}{n_1+1} - 1 \right] \right\}. \tag{4.7}
 \end{aligned}$$

Proof: Consider a_i , $i \in \{1, 2, 3\}$, given by (3.19). If $\mathbf{x} \in \pi_2$, then from Lemma 4.1 we

get

$$\begin{aligned}
\frac{1}{2}a_1 + \frac{1}{4}\Delta a_2 - (1 + \frac{1}{8}\Delta^2)a_3 &= \frac{3}{4}\frac{1}{\Delta}(1-q^2)m^2c_0p\left(\frac{n_2}{n_2+1}-1\right) + \frac{1}{4}\Delta\left(\frac{n_2}{n_2+1}-1\right)\left(3m^2c_0\frac{n_2}{n_2+1}-5\right) \\
&+ \frac{1}{16}\Delta^3\left(\left(\frac{n_2}{n_2+1}\right)^3\frac{12+m^2(m-2)}{m^2(m-2)}-3\Delta^3\left(\frac{n_2}{n_2+1}\right)^3\frac{4+m^2(m-2)}{m^2(m-2)}\right. \\
&+ 3\frac{n_2}{n_2+1}-1\Big) + \frac{1}{4}\frac{1}{\Delta}(1-q^2)m^2c_0p - \frac{1}{4}\Delta\left(1-m^2c_0\frac{n_2}{n_2+1}\right) \\
&+ \frac{1}{8}\Delta^3\left(\left(\frac{n_2}{n_2+1}\right)^2\frac{4+m^2(m-2)}{m^2(m-2)} - \frac{n_2}{n_2+1} + 1\right) \\
&- \frac{1}{2}\frac{1}{\Delta}(1-q^2)m^2c_0p\left(\frac{n_2}{n_2+1}-1\right) - \frac{1}{2}\Delta\left(\frac{n_2}{n_2+1}-1\right)\left(m^2c_0\frac{n_2}{n_2+1}-1\right) \\
&- \frac{1}{8}\Delta^3\left(\frac{1}{3}\left(\frac{n_2}{n_2+1}\right)^3\frac{12+m^2(m-2)}{m^2(m-2)} + \left(\frac{n_2}{n_2+1}\right)^2\frac{4+m^2(m-2)}{m^2(m-2)}\right. \\
&- \frac{n_2}{n_2+1} + \frac{1}{3}\Big) - \frac{1}{16}\Delta(1-q^2)m^2c_0p\left(\frac{n_2}{n_2+1}-1\right) \\
&- \frac{1}{16}\Delta^3\left(\frac{n_2}{n_2+1}-1\right)\left(m^2c_0\frac{n_2}{n_2+1}-1\right) + \frac{1}{192}\Delta^5\left(\frac{1}{3}\left(\frac{n_2}{n_2+1}\right)^3\right. \\
&\times \frac{12+m^2(m-2)}{m^2(m-2)} - 3\left(\frac{n_2}{n_2+1}\right)^2\frac{4+m^2(m-2)}{m^2(m-2)} + 3\frac{n_2}{n_2+1}-1\Big) \\
&= \frac{1}{4}\frac{1}{\Delta}(1-q^2)m^2c_0p\frac{n_2}{n_2+1} + \frac{1}{2}\Delta\left[\frac{1}{2}m^2c_0\left(\frac{n_2}{n_2+1}\right)^2\right. \\
&- \frac{1}{8}(1-q^2)m^2c_0p\left(\frac{n_2}{n_2+1}-1\right) - \frac{3}{4}\frac{n_2}{n_2+1} + \frac{1}{2}\Big] \\
&+ \frac{1}{2}\Delta^3\left[\frac{1}{24}\left(\frac{n_2}{n_2+1}\right)^3\frac{12+m^2(m-2)}{m^2(m-2)} - \frac{1}{16}\left(\frac{n_2}{n_2+1}\right)^2\right. \\
&\times \frac{4+m^2(m-2)}{m^2(m-2)} - \frac{1}{8}m^2c_0\frac{n_2}{n_2+1}\left(\frac{n_2}{n_2+1}-1\right) - \frac{1}{24}\Big] \\
&- \frac{1}{192}\Delta^5\left[\left(\frac{n_2}{n_2+1}\right)^3\frac{12+m^2(m-2)}{m^2(m-2)}\right. \\
&- 3\left(\frac{n_2}{n_2+1}\right)^2\frac{4+m^2(m-2)}{m^2(m-2)} + 3\frac{n_2}{n_2+1}-1\Big].
\end{aligned}$$

Therefore, substituting the above expression into (4.3) the proof follows. To prove for when $\mathbf{x} \in \pi_2$ we can use the same approach as when $\mathbf{x} \in \pi_1$ and the result is established. $\square$

Conclusion

OUR aim in this thesis was to develop a different approach for deriving approximations of the misclassification errors for the likelihood based classification rule. These approximations can be generalized to the case of classification into one of several populations. In this chapter, we will discuss the results of the thesis in comparison to already existing approaches and their contribution in regards to classification techniques. Furthermore, we will introduce our future research interests.

5.1 Discussion

In this thesis we start by proposing classification rules using a maximum likelihood approach. We consider the case with classification into two populations, π_1 and π_2 . It is assumed that the two populations have a common covariance. There exists several approaches of deriving those rules when the parameters of the population in which we wish to classify the new observation are estimated. Anderson (1951) proposed a classification rule that he labeled as the W -rule using a plug-in approach and Kudo (1959, 1960) introduced the Z -rule; a maximum likelihood criterion as an alternative to Wald-Anderson's W -rule:

$$W = \mathbf{x}' \mathbf{S}^{-1} (\bar{\mathbf{y}} - \bar{\mathbf{z}}) - \frac{1}{2} (\bar{\mathbf{y}} + \bar{\mathbf{z}})' \mathbf{S}^{-1} (\bar{\mathbf{y}} - \bar{\mathbf{z}}),$$

which is a linear function in $\mathbf{x}$. If $W \geq 0$ then $\mathbf{x}$ is classified to π_1 and if $W < 0$ to π_2 . Moreover,

$$Z = \frac{1}{2} \left[\frac{n_2}{n_2 + 1} (\mathbf{x} - \bar{\mathbf{z}})' \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{z}}) - \frac{n_1}{n_1 + 1} (\mathbf{x} - \bar{\mathbf{y}})' \mathbf{S}^{-1} (\mathbf{x} - \bar{\mathbf{y}}) \right],$$

which is a quadratic function in $\mathbf{x}$. The maximum likelihood rule is to classify $\mathbf{x}$ as coming from π_1 if Z is positive and π_2 if Z is negative. In this thesis, the discrimination is performed with respect to the proposed maximum likelihood based classification rules

for $\tilde{D}$ and $\hat{D}$, given in Proposition 2.2 and Proposition 2.1, respectively. We regard an observation x as coming from π_1 or π_2 according to whether the observed value of the discriminant functions $\tilde{D}$ or $\hat{D}$ is positive or negative. Note that when $n_1 = n_2 \equiv n$, $\hat{D} = -\frac{mn}{n+1}W = mZ$, where $m = 2n - p - 3$.

For known expectation and variance, the distribution of the classification rule D , follows from Theorem 2.1, that is, $D \sim \mathcal{N}(\frac{1}{2}\Delta^2, \Delta^2)$ when the observation x belongs to π_1 and $D \sim \mathcal{N}(-\frac{1}{2}\Delta^2, \Delta^2)$ when the observation x belongs to π_2 , which is also the limiting distribution of both $\tilde{D}$ and $\hat{D}$, see Theorem 2.7. It is important to know misclassification errors before and after drawing the two samples when forming the classification rule to know how reliable the classification can be.

We observed in Chapter 2 that it is difficult to evaluate the exact distributions of $\hat{D}$ and $\tilde{D}$. Therefore, we handled this by performing out Edgeworth-type expansions in order to further obtain misclassification probabilities. Note that an Edgeworth-type expansion does not have to be a density. In our case, the integration of both Edgeworth-type expansions of the discriminant functions developed in Chapter 3 over the entire space equals one, that is,

$$\int_{-\infty}^{+\infty} f_{\tilde{D}}(x)dx = \int_{-\infty}^{+\infty} f_{\hat{D}}(x)dx = 1.$$

However, we have not shown nonnegativity of the expansions. The Edgeworth-type expansion was performed in terms of cumulants of the discriminant functions $\tilde{D}$ and $\hat{D}$ through a standard normal distribution up to three terms. The expansion is a polynomial function in x .

In addition, in order to estimate misclassification probabilities for the classification rules whether $\tilde{D} > 0$ or $\tilde{D} < 0$ and whether $\hat{D} > 0$ or $\hat{D} < 0$, we used the Edgeworth-type expansions $f_{\tilde{D}}(x)$ and $f_{\hat{D}}(x)$ of standardized variables with their limiting distribution given in Theorem 2.7. Misclassification errors from Theorem 4.2 and Theorem 4.3 are functions of the Mahalanobis distance.

Figure 5.1 shows the change in misclassification probabilities using $\tilde{D}$ and $\hat{D}$ with respect to the Mahalanobis distance (Δ), where $P(x \rightarrow \pi_2 | x \in \pi_1)$ is the probability that an observation x is wrongly classified into population π_2 . The curves of rules with $\tilde{D}$ and $\hat{D}$ are more or less similar.

Anderson (2003) derived misclassification probabilities using W and Z , given in Theorem 5.1 and Theorem 5.2 below, respectively.

Theorem 5.1 (Anderson (2003))

For large n_1 and n_2 , and $n = n_1 + n_2 - 2$ and $u = -\frac{1}{2}\Delta$, where Δ is the Mahalanobis distance,

$$\begin{aligned} P\left\{\frac{W - \frac{1}{2}\Delta^2}{\Delta} \leq u | x \in \pi_1\right\} &= \Phi(u) - \phi(u) \left\{ \frac{1}{2n_1\Delta^2} [u^3 + (p-3)u - p\Delta] \right. \\ &\quad + \frac{1}{2n_2\Delta^2} [u^3 + 2\Delta u^2 + (p-3+\Delta^2)u + (p-2)\Delta] \\ &\quad \left. + \frac{1}{4n} [4u^3 + 4\Delta u^2 + (6p-6+\Delta^2)u + 2(p-1)\Delta] \right\} + O(n^{-2}) \end{aligned} \quad (5.1)$$

Figure 5.1: Misclassification errors for $\tilde{D}$ and $\hat{D}$ rules when $n_1 = n_2 = 100$ and $p = 10$.

and $P\left\{-\left(\frac{W + \frac{1}{2}\Delta^2}{\Delta}\right) \leq u | \mathbf{x} \in \pi_2\right\}$ is similar to (5.1) with n_1 and n_2 interchanged.

Theorem 5.2 (Anderson (2003))

For large n_1 and n_2 , n_1/n_2 and $n = n_1 + n_2 - 2$ and $u = -\frac{1}{2}\Delta$,

$$\begin{aligned}
 P\left\{\frac{Z - \frac{1}{2}\Delta^2}{\Delta} \leq u | \mathbf{x} \in \pi_1\right\} = & \Phi(u) - \phi(u) \left\{ \frac{1}{2n_1\Delta^2} [u^3 + \Delta u^2 + (p-3)u - \Delta] \right. \\
 & + \frac{1}{2n_2\Delta^2} [u^3 + \Delta u^2 + (p-3+\Delta^2)u - \Delta^3 - \Delta] \\
 & \left. + \frac{1}{4n} [4u^3 + 4\Delta u^2 + (6p-6+\Delta^2)u + 2(p-1)\Delta] \right\} + O(n^{-2}),
 \end{aligned} \tag{5.2}$$

and $P\left\{-\left(\frac{Z + \frac{1}{2}\Delta^2}{\Delta}\right) \leq u | \mathbf{x} \in \pi_2\right\}$ equals (5.2) with n_1 and n_2 interchanged.

Figure 5.2 and Figure 5.3 display the change in misclassification probabilities using the rules $\tilde{D}$, $\hat{D}$, W and Z with respect to Δ . We can see from both figures that the rules $\tilde{D}$ and $\hat{D}$ are comparably similar to Anderson's W and Z rules. However, Figure 5.1 also shows that the rule with $\hat{D}$ may have negative misclassification error. Moreover, for small n , ($n < 10$), see Figure 5.3, misclassification errors using $\hat{D}$ can be negative. This is due to the fact that the Egdeworth-type expansion is not always a density. Hence, misclassification errors are not really probabilities and might yield irrelevant values. However, a negative value can be interpreted that the misclassification error is small.

Figure 5.2: Misclassification errors for $\tilde{D}$, $\hat{D}$, W and Z rules when $n_1 = n_2 = 100$ and $p = 10$.

We consider classification only if the misclassification error is less than 50%. Therefore, it is not meaningful to classify with $\tilde{D}$ when $\Delta < 0.4$. On the other hand, the misclassification error using $\hat{D}$ is always less than 50%.

Figure 5.3: Misclassification errors for $\tilde{D}$, $\hat{D}$, W and Z rules when $n_1 = 5$, $n_2 = 4$ and $p = 3$.

5.2 Future research

In this thesis, we studied classification of an observation into one of two populations. In future, we want to extend the methods used to the classification of an observation into more than two population.

Furthermore, in order to overcome the complexity of the distribution of classification rules, we expanded their distributions using Edgeworth-type expansion. In addition, the expansions of standardized variables by the limiting distribution of the classification rules $\tilde{D}$ and $\hat{D}$ are used in order to derive misclassification probabilities. What are the other ways to standardize the rules $\tilde{D}$ and $\hat{D}$ other than the limiting distribution? In the next research we will study different ways of standardizing the classification rules to get simpler densities for the derivation of misclassification errors. The variance of both $\tilde{D}$ and $\hat{D}$ contain a constant term which does not contain the squared Mahalanobis distance (Δ^2). As a result, the misclassification probabilities contain the term $\frac{1}{\Delta}$. Though this term is also found in misclassification errors around the W - and Z -rules, it results in irrelevant

values for misclassification probabilities using $\tilde{D}$ and $\hat{D}$ since the Edgeworth-type expansion used to derive the probabilities is not really a density. The idea is to find an appropriate variance function of the squared Mahalanobis distance for the standardization which cancels out the constant term within the variance of the classification rules.

Appendices

In this part we have lemmas, theorems, notations, symbols and definitions that are referenced and used throughout this thesis.

Appendix A: Notations

Note that all vectors are column vectors. In general, bold lower case letters are used to denote vector valued and bold upper case letters are used for matrix valued variables. However, there might be exceptions from these general rules. The same symbol can be used for different purposes. The principal notation is listed below, any deviations are explained in the text.

Symbols and Operators

$\mathbf{0}_n$	Null vector of dimension n
$\mathbf{1}_n$	Vector of n ones
Σ	Covariance matrix
$\mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$	p -dimensional normal (Gaussian) distribution with mean value $\boldsymbol{\mu}$ and covariance matrix Σ
$W_p(\Sigma, n)$	p -dimensional Wishart distribution with n degrees of freedom and scale matrix Σ
$\Phi(\mathbf{x})$	Normal (Gaussian) distribution function
$f_{\mathbf{x}}(\mathbf{x})$	Probability density function of $\mathbf{x}$
$L(\cdot)$	Likelihood function
$l(\cdot)$	Log-likelihood function
$\frac{d(\cdot)}{dx}$	Derivative of a function with respect to x
$\otimes$	Kronecker product

$\sim$	Denotes "is distributed according to"
$\in$	Belongs to
$\xrightarrow{p}$	Convergence in probability
$P(\mathbf{x} \leq k)$	Probability that the random variable $\mathbf{x}$ is less than or equal to k
$\mathbf{I}_n$	Identity matrix of size n
$\mathbf{K}_{p,q}$	Commutation matrix
$ \mathbf{A} $	Determinant of matrix $\mathbf{A}$
$\text{tr} \mathbf{A}$	Trace of matrix $\mathbf{A}$
$\mathbf{A}^\top$	Transpose of matrix $\mathbf{A}$
$\mathbf{A}^{-1}$	Inverse of matrix $\mathbf{A}$
$E[\mathbf{x}]$	Expectation of the random variable $\mathbf{x}$
$E[\mathbf{x} \mathbf{y}]$	Conditional expectation of the random variable $\mathbf{x}$, given that the random variable $\mathbf{y} = \mathbf{y}$
$\text{Var}[\mathbf{x}]$	Variance of the random variable $\mathbf{x}$
$\min$	Minimum
$\max$	Maximum
$\text{vec}(\cdot)$	the vectorization of a matrix

Abbreviations and Acronyms

i.i.d.	independent identically distributed
LDA	Linear Discriminant Analysis
MLE	Maximum Likelihood Estimators (Estimate)
pdf	probability density function
s.t.	subject to
w.r.t.	with respect to

Definitions

Definition 6.1. The partitioned matrix $\mathbf{K}_{p,q} : pq \times pq$ consisting of $q \times p$ -blocks is called commutation matrix, if

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

Definition 6.2. The vectorization of a $m \times n$ matrix $\mathbf{A}$, denoted is the $mn \times 1$ column vector

$$\text{vec} \mathbf{A} = [a_{11}, \dots, a_{m1}, a_{12}, \dots, a_{m2}, \dots, a_{1n}, \dots, a_{mn}]^\top.$$

Definition 6.3. The Kronecker product of $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{kl})$ is defined as $\mathbf{A} \otimes \mathbf{B} = (a_{ij} b_{kl})$.

Definition 6.4. The random matrix $\mathbf{W}$ is Wishart-distributed with n degrees of freedom if and only if $\mathbf{W} = \mathbf{X}\mathbf{X}^\top$, for some $\mathbf{X} \sim \mathcal{N}_{p,n}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I})$, which is denoted by $\mathbf{W} \sim W_p(\mathbf{\Sigma}, n)$.

Appendix B: Useful results

Theorem 6.1

Consider the case of a standard normal distribution, $\mathcal{N}(0, 1)$. The j^{th} univariate Hermite polynomials $H_j(x, 1)$, $j \in \{0, 1, 2, 3\}$ equal

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

Proof: The density function of the standard normal distribution $\mathcal{N}(0, 1)$, is given by:

$$f_x(\mathbf{x}) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2}.$$

For $k = 0$, the Definition 3.1 provides a trivial proof of $H_0(x, 1)$.

For $k = 1$ the first order Hermite polynomial equals

$$H_1(x, 1) = \frac{(-1)}{(2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2}} \frac{d}{dx} \left[(2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2} \right] = -\frac{(-1)}{(2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2}} \frac{e^{-\frac{x^2}{2}} x}{2^{\frac{1}{2}} \pi^{\frac{1}{2}}} = x.$$

For $k = 2$ the second order Hermite polynomial is equal to

$$\begin{aligned} H_2(x, 1) &= \frac{1}{(2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2}} \frac{d^2}{dx^2} \left[(2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2} \right] \\ &= -\frac{1}{(2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2}} \frac{1}{(2\pi)^{\frac{1}{2}}} \left(-e^{-\frac{x^2}{2}} x^2 + e^{-\frac{x^2}{2}} \right) = x^2 - 1. \end{aligned}$$

Lastly, for $k = 3$ the third order Hermite polynomial becomes

$$\begin{aligned} H_3(x, 1) &= -\frac{1}{(2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2}} \frac{d^3}{dx^3} \left[(2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2} \right] \\ &= \frac{1}{(2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}x^2}} \frac{e^{-\frac{x^2}{2}} x^3 - 3e^{-\frac{x^2}{2}} x}{2^{\frac{1}{2}} \pi^{\frac{1}{2}}} = x^3 - 3x. \end{aligned}$$

□

Lemma 6.1

Let $\mathbf{A} : p \times q$ and $\mathbf{B} : r \times s$. Then

$$\text{vec}(\mathbf{A} \otimes \mathbf{B}) = (\mathbf{I}_q \otimes \mathbf{K}_{s,p} \otimes \mathbf{I}_r)(\text{vec} \mathbf{A} \otimes \text{vec} \mathbf{B}),$$

where, $\mathbf{K}_{s,p}$ is the commutation matrix.

Proof is derived in von Rosen (2018).

Theorem 6.2

Let $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, and let $\mathbf{A}$ be a $p \times p$ constant matrix, the h^{th} cumulant of the quadratic form $\mathbf{X}^T \mathbf{A} \mathbf{X}$ is

$$\psi_h(\mathbf{X}^T \mathbf{A} \mathbf{X}) = 2^{h-1} h! \left\{ \frac{\text{tr}\{(\mathbf{A} \boldsymbol{\Sigma})^h\}}{h} + \boldsymbol{\mu}^T (\mathbf{A} \boldsymbol{\Sigma})^{h-1} \mathbf{A} \boldsymbol{\mu} \right\}$$

This expression is derived in Mathai and Provost (1992).

Theorem 6.3

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

$$(i) E[\mathbf{X}^\top \mathbf{A} \mathbf{X}] = \text{tr}\{\mathbf{A} \boldsymbol{\Sigma}\} + \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu}$$

$$(ii) \text{Var}[\mathbf{X}^\top \mathbf{A} \mathbf{X}] = 2[\text{tr}\{(\mathbf{A} \boldsymbol{\Sigma})^2\} + 2\boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\mu}]$$

Proof: From Theorem 6.2 above we can directly deduce the first and second cumulants of the quadratic form

$$(i) \psi_1(\mathbf{X}^\top \mathbf{A} \mathbf{X}) = E[\mathbf{X}^\top \mathbf{A} \mathbf{X}] = \text{tr}\{\mathbf{A} \boldsymbol{\Sigma}\} + \boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\mu}$$

$$(ii) \psi_2(\mathbf{X}^\top \mathbf{A} \mathbf{X}) = \text{Var}[\mathbf{X}^\top \mathbf{A} \mathbf{X}] = 2[\text{tr}\{(\mathbf{A} \boldsymbol{\Sigma})^2\} + 2\boldsymbol{\mu}^\top \mathbf{A} \boldsymbol{\Sigma} \mathbf{A} \boldsymbol{\mu}]$$

□

Definition 6.5. The following constants are mostly referenced in Chapter 2 for their technical derivation from Theorem 6.4.

$$(i) c_1 = \frac{1}{n-p-1}; (ii) c_2 = \frac{n-p-2}{(n-p)(n-p-1)(n-p-3)}; (iii) c_3 = \frac{1}{n-p-2} c_2.$$

Theorem 6.4

Let $\mathbf{W} \sim W_p(n, \boldsymbol{\Sigma})$ and let $\mathbf{A}$ and $\mathbf{B}$ be $p \times p$ constant matrices and c_1, c_2 and c_3 are defined in Definition 6.5

$$(i) E[\mathbf{W}^{-1}] = c_1 \boldsymbol{\Sigma}^{-1}, n-p-1 > 0;$$

$$(ii) E[\mathbf{W}^{-1} \mathbf{A} \mathbf{W}^{-1}] = c_2 \boldsymbol{\Sigma}^{-1} \mathbf{A} \boldsymbol{\Sigma}^{-1} + c_3 [\boldsymbol{\Sigma}^{-1} \mathbf{A}^\top \boldsymbol{\Sigma}^{-1} + \text{tr}\{\mathbf{A} \boldsymbol{\Sigma}^{-1}\} \boldsymbol{\Sigma}^{-1}], n-p-3 > 0;$$

$$(iii) \text{Var}[\mathbf{W}^{-1}] = c_3 (\mathbf{I} + \mathbf{K}_{p,p}) (\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1}) + (c_2 - c_1^2) \text{vec} \boldsymbol{\Sigma}^{-1} \text{vec}^\top \boldsymbol{\Sigma}^{-1}, n-p-3 > 0.$$

The proofs and technical expressions in Theorem 6.4 (i) and (ii) above can be found in von Rosen (1988) and (iii) in Fujikoshi et al. (2011).

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*, 3rd ed. John Wiley & Sons, Inc., New York; Chapman & Hall, Ltd., London.
- [3] Davis, A. (1976). Statistical distributions in univariate and multivariate Edgeworth populations. *Biometrika*, 63(3):661–670.
- [4] Edgeworth, F. (1907). On the representation of statistical frequency by a series. *Journal of the Royal Statistical Society*, 70(1):102–106.
- [5] Fisher, R. A. (1936). The use of multiple measurements in 10 taxonomic problems. *Annals of Eugenics*, 7(2):179–188.
- [6] Fisher, R. A. (1938). The statistical utilization of multiple measurements. *Annals of Eugenics*, 8(4):376–386.
- [7] Fujikoshi, Y., Ulyanov, V. V., and Shimizu, R. (2011). *Multivariate Statistics: High-dimensional and Large-Sample Approximations*, volume 760. John Wiley & Sons, New Jersey.
- [8] Gupta, S. S. and Panchapakesan, S. (1982). Edgeworth expansions in statistics: A brief review. *Technical report, Purdue university Lafayette in department of statistics*.
- [9] Izenman, A. J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning*, volume 978. Springer Science & Business Media, New York.
- [10] Kollo, T. and von Rosen, D. (2005). *Advanced Multivariate Statistics with Matrices*, volume 579. Springer Science & Business Media, Dordrecht.

- [11] 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.
- [12] Kudo, A. (1960). The classificatory problem viewed as a two-decision problem-ii. *Memoirs of the Faculty of Science, Kyushu University. Series A, Mathematics*, 14(1):63–83.
- [13] Mathai, A. and Provost, S. B. (1992). *Quadratic Forms in Random Variables: Theory and Applications*, volume 87. Marcel Dekker, Inc.
- [14] 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.
- [15] Okamoto, M. (1963). An asymptotic expansion for the distribution of the linear discriminant function. *The Annals of Mathematical Statistics*, 34(4):1286–1301.
- [16] Ranga Rao, R. (1960). *Some Problems in Probability Theory*. PhD thesis, D. Phil., Thesis, Calcutta University.
- [17] 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.
- [18] von Rosen, D. (1988). Moments for the inverted Wishart distribution. *Scandinavian Journal of Statistics*, pages 97–109.
- [19] von Rosen, D. (2018). *Bilinear regression analysis*. Springer, New York.
- [20] Schaafsma, W. and van Vark, G. N. (1977). Classification and discrimination problems with applications, part I. *Statistica Neerlandica*, 31(1):25–45.
- [21] Schaafsma, W. and van Vark, G. N. (1979). Classification and discrimination problems with applications, part II. *Statistica Neerlandica*, 33(2):91–126.
- [22] Siotani, M. (1977). *Asymptotic Expansions for Error Rates and Comparison of the W-procedure and the Z-procedure in Discriminant Analysis*. North-Holland, Amsterdam.
- [23] Siotani, M. and Wang, R.-H. (1975). *Further Expansion Formulae for Error Rates and Comparison of the W-and Z-Procedures in Discriminant Analysis*. Technical Report No. 33, Department of Statistics, Kansas State University, Manhattan, Kansas.
- [24] 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.
- [25] Smith, H. F. (1936). A discriminant function for plant selection. *Annals of Eugenics*, 7(3):240–250.
- [26] 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.

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

FACULTY OF SCIENCE AND ENGINEERING

Linköping Studies in Science and Technology, Licentiate thesis No. 1911, 2021
Department of Mathematics

Linköping University
SE-581 83 Linköping, Sweden

www.liu.se