ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 25, Number 1, June 2021
Available online at http://acutm.math.ut.ee

Asymptotic approximation of misclassification probabilities in linear discriminant analysis with repeated measurements

Edward K. Ngailo, Dietrich von Rosen, and Martin Singull

Abstract. We propose asymptotic approximations for the probabilities of misclassification in linear discriminant analysis when the group means follow a growth curve structure. The discriminant function can classify a new observation vector of p repeated measurements into one of several multivariate normal populations with equal covariance matrix. We derive certain relations of the statistics under consideration in order to obtain asymptotic approximation of misclassification errors for the two group case. Finally, we perform Monte Carlo simulations to evaluate the reliability of the proposed results.

1. Introduction

One can say that discriminant analysis and classification were first really developed in the 1930's when multivariate statistics was a blossoming area and attracted researchers. One of the first to deal with linear discriminant analysis and classification as we know it today was the well known Sir R. A. Fisher [3]. Fisher published several papers on discriminant analysis, including [4] in which he reviewed his 1936 work and related it to the contributions by Hotelling and his T^2 statistic [6], and by Mahalanobis to his Δ^2 statistic [16] and other distance measures. Nowadays, of course several textbooks, for example [8, 9, 15, 22], have treated discriminant analysis in detail.

Linear discriminant analysis is a technique which is commonly used for the supervised classification problems. It is used for modeling differences in

Received August 19, 2020.
2020 *Mathematics Subject Classification.* 62H30.
Key words and phrases. Asymptotic approximation, Growth Curve model, linear discriminant function, probability of misclassification.
https://doi.org/10.12697/ACUTM.2021.25.05
Corresponding author: Martin Singull

groups π_i with $i = 1, 2, \ldots, k$, that is, separating two ($k = 2$) or more ($k > 2$) classes by maximizing the distance between means $\boldsymbol{\mu}_i$ under the assumption of the same variance-covariance matrix $\boldsymbol{\Sigma}$. The statistical problem treated by Fisher in [4] was that of assigning an unknown observation into one of two known groups on the basis of p measured characteristics, i.e., a feature vector $\mathbf{x} = (x_1, \ldots, x_p)'$. Furthermore, the groups are assumed to follow distributions with the same variance-covariance matrix. The sample based Fisher's discriminant function [27] equals

$$L(\mathbf{x}; \overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \mathbf{S}) = (\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\mathbf{S}^{-1}\mathbf{x} - \frac{1}{2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\mathbf{S}^{-1}(\overline{\mathbf{x}}_1 + \overline{\mathbf{x}}_2), \qquad (1)$$

where $\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_2$ denote the sample mean vectors of the two groups, i.e., $\overline{\mathbf{x}}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{x}_{ij}$, where $\mathbf{x}_{ij}$ is observation j from group i with sample size n_i, and $\mathbf{S}$ is the pooled sample covariance matrix given by

$$\mathbf{S} = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{n_1 + n_2 - 2}, \text{ where } \mathbf{S}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i} \big(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i\big)\big(\mathbf{x}_{ij} - \bar{\mathbf{x}}_i\big)',$$

The classification rule for a new observation $\mathbf{x}$ is: classify $\mathbf{x}$ into π_1 if $L(\mathbf{x}; \overline{\mathbf{x}}_1, \overline{\mathbf{x}}_2, \mathbf{S}) > 0$ holds, otherwise classify into π_2. In order to understand the performance of the classification rule it is important to evaluate the probability of misclassification.

An asymptotic expression for the misclassification errors for the linear discriminant rule given above and for large n_1 and n_2 is given by Okamoto in [18]. That is, the probability of misclassifying an observation from π_i in π_j, with $i, j = 1, 2$ and $i \neq j$, is given by

$$e(j|i) \approx \Phi\left(-\frac{\Delta}{2}\right) + \phi\left(\frac{\Delta}{2}\right)\left(\frac{a_i}{n_1} + \frac{a_j}{n_2} + \frac{a_3}{n_1 + n_2 - 2}\right), \qquad (2)$$

where $a_1 = \frac{\Delta^2 + 12(p-1)}{16\Delta}$, $a_2 = \frac{\Delta^2 - 4(p-1)}{16\Delta}$, $a_3 = \frac{\Delta}{4}(p - 1)$, and $\Phi(\cdot)$ and $\phi(\cdot)$ are the standard normal cumulative distribution and probability density functions, respectively. Furthermore, $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is the squared Mahalanobis distance [16] which simply can be estimated by the consistent estimator

$$\widehat{\Delta}^2 = \frac{n_1 + n_2 - p - 3}{n_1 + n_2 - 2}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2)'\mathbf{S}^{-1}(\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2).$$

Also in recent years there are studies in linear discriminant analysis which evaluate the misclassification errors in an asymptotic approach. For example, [7, 28] expanded the expression for the asymptotic approximation for the misclassification errors using Taylor series expansion, and [5] derives many details on the Taylor series expansion of the asymptotic misclassification expression and their possible errors of approximations.

Discriminant analysis is usually applied to multivariate statistical problems in which several features are collected simultaneously. However, [1,

11, 12, 20] among others considered discriminant analysis procedures for repeated measurements. Repeated measures discriminant analysis procedures are applied to data collected at multiple occasions on the same individual. Later, [24, 25] developed procedures based on univariate and multivariate repeated measures data, focusing on different covariance structures. Recently, [13] reviewed the literature on discriminant analysis for univariate and multivariate repeated measures data, focusing on covariance patterns and linear mixed-effects models with applications to psychological research.

One of the first to discuss discrimination between growth curves was Burnaby in his early paper [1] followed by Rao [20]. Burnaby's focus was to generalize procedures for constructing discriminant functions as well as to propose a generalized distance between the populations of repeated measurements. In our work, the classification of growth curves relies solely on the Growth Curve model given by [19]. There have been earlier attempts to modify the linear classification function given by [4] including Growth Curve models for the group means. [11] considered classification of individuals into one of two growth curves using a Bayesian approach, which was later extended by [14]. Again, [12] developed both non-Bayesian and Bayesian classification of growth curves, where he considered two different covariance structures, the arbitrary positive definite covariance matrix and Rao's simple covariance structure. More recently, [17] considered the Rao's classification scores [21] with group means following the Growth Curve model structure given by [19]. They used only simulations to show the performance of the classification rule, based on both an arbitrary covariance matrix and structured covariance matrices (compound symmetry and independence).

Hence, the modifications required for the classification function (1) when the repeated measurements of the populations obey some structure, have been studied. However, very little has been said about the misclassification error rate. In this paper we will study the linear discriminant function for repeated measurements, i.e., growth curves, and derive some asymptotic approximations of the misclassification probabilities. The results are derived following the ideas given by [5] for the cases when the covariance matrix is known and when it is unknown and hence have to be estimated. The follow-up analysis involves investigating the performance of the proposed approximations through Monte Carlo simulations.

The organization of this paper is as follows. In Section 2, the main idea is given and the linear classification function for growth curves is presented. In Section 3, the approximations of the probabilities of misclassifications are derived for both known and unknown covariance matrices and the proposed results are supported by a simulation study in Section 4. In Section 5, we finalize the paper by giving a brief summary of the results.

Throughout this paper matrices will be denoted by bold upper case letters, vectors by bold lower case letters, and elements of matrices by ordinary letters.

2. Classification into one of two growth curves

In this section we start by introducing the Growth Curve model of [19] which will be considered when formulating the linear discriminant function. The Growth Curve model is given by

$$\mathbf{X} = \mathbf{ABC} + \mathbf{E}, \quad \mathbf{E} \sim N_{p,n}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I}_n), \tag{3}$$

where $\mathbf{X}$ is a $p \times n$ data matrix of n independent individuals with p repeated measurements, $\mathbf{E}$ is the $p \times n$ error matrix with columns assumed to be independently p variate normally distributed with mean vector $\mathbf{0}_p$ and a positive definite variance-covariance matrix $\mathbf{\Sigma}$. We will assume a polynomial growth of order $q-1$ and $k = 2$ independent groups of n_i individuals, $i = 1, 2$, with the total sample size as $n = n_1 + n_2$. Then, the design matrices $\mathbf{A} : p \times q, \mathbf{C} : 2 \times n$ and parameter $\mathbf{B} : q \times 2$ can be given by

$$\mathbf{A} = \begin{pmatrix} 1 & t_1 & \dots & t_1^{q-1} \\ 1 & t_2 & \dots & t_2^{q-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_p & \dots & t_p^{q-1} \end{pmatrix}, \ \mathbf{C} = \begin{pmatrix} \mathbf{1}'_{n_1} & \mathbf{0}'_{n_2} \\ \mathbf{0}'_{n_1} & \mathbf{1}'_{n_2} \end{pmatrix}, \ \mathbf{B} = \begin{pmatrix} \mathbf{b}_1, \mathbf{b}_2 \end{pmatrix}, \tag{4}$$

where $\mathbf{1}'_{n_i}$ and $\mathbf{0}'_{n_i}$ are vectors of ones and zeros, respectively. The optimal classification rule that minimizes the probability of misclassification is to assign a vector $\mathbf{x}$ following the Growth Curve model to π_1 if

$$q_1 f_1(\mathbf{x}) > q_2 f_2(\mathbf{x}), \tag{5}$$

and to π_2 otherwise. Here $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are the density functions for $\mathbf{x}$ belonging to π_1 or π_2, respectively, and q_1 and q_2 are the prior probabilities about the classification. Given the Growth Curve model (3), we have the probability density functions

$$f_i(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \mathrm{tr}\Big\{ \mathbf{\Sigma}^{-1} (\mathbf{x} - \mathbf{A}\mathbf{b}_i)(\mathbf{x} - \mathbf{A}\mathbf{b}_i)' \Big\} \right\}, \quad i = 1, 2.$$

Assuming $q_1 = q_2$, i.e., no prior information, then the classification rule (5) can be given as:

$$\text{classify } \mathbf{x} \text{ to } \pi_1 \text{ if } L(\mathbf{x}; \mathbf{b}_1, \mathbf{b}_2, \mathbf{\Sigma}) > 0, \text{ and to } \pi_2 \text{ otherwise,} \tag{6}$$

where the linear classification function L is

$$L(\mathbf{x}; \mathbf{b}_1, \mathbf{b}_2, \mathbf{\Sigma}) = (\mathbf{b}_1 - \mathbf{b}_2)' \mathbf{A}' \mathbf{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} (\mathbf{b}_1 - \mathbf{b}_2)' \mathbf{A}' \mathbf{\Sigma}^{-1} \mathbf{A} (\mathbf{b}_1 + \mathbf{b}_2). \tag{7}$$

In this paper we will study two cases: $\mathbf{\Sigma}$ is known and $\mathbf{\Sigma}$ is unknown.

2.1. Estimators of the parameters B and Σ. There exist different approaches to estimate the model parameters $\mathbf{B}$ and $\mathbf{\Sigma}$, in the Growth Curve model (3), see e.g., [23, 27]. The maximum likelihood estimator of $\mathbf{B}$ when $\mathbf{A}$ and $\mathbf{C}$ are assumed to have full rank, is given by

$$\widehat{\mathbf{B}} = (\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2) = (\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1},$$

where $\mathbf{S}$ is the within sum of squares matrix

$$\mathbf{S} = \mathbf{X}(\mathbf{I}_n - \mathbf{P}_{\mathbf{C}'})\mathbf{X}', \quad \text{with} \quad \mathbf{P}_{\mathbf{C}'} = \mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1}\mathbf{C}.$$

With $\mathbf{C}$ given in (4) and $k = 2$, we have the estimators for the mean parameters as

$$\widehat{\mathbf{b}}_i = (\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\overline{\mathbf{x}}_i,$$

where $\overline{\mathbf{x}}_i = \frac{1}{n_i}\mathbf{X}_i\mathbf{1}_{n_i}$, with $\mathbf{X}_i = \begin{pmatrix}\mathbf{x}_{i1} & \dots & \mathbf{x}_{in_i}\end{pmatrix} : p \times n_i$ being the n_i observations from group $\pi_i, i = 1, 2$, i.e., $\mathbf{X} = \begin{pmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{pmatrix}$. Furthermore, the maximum likelihood estimator for $\mathbf{\Sigma}$ is given by

$$n\widehat{\mathbf{\Sigma}} = (\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}\mathbf{C})(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}\mathbf{C})' = \mathbf{S} + (\mathbf{I}_p - \mathbf{P}_{\mathbf{A},\mathbf{S}})\mathbf{X}\mathbf{P}_{\mathbf{C}'}\mathbf{X}(\mathbf{I}_p - \mathbf{P}_{\mathbf{A},\mathbf{S}}), \quad (8)$$

which is a positive definite with probability one when $p < n - 2$ and $\mathbf{P}_{\mathbf{A},\mathbf{S}} = \mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}$. When $\mathbf{\Sigma}$ is known the estimator of the mean parameter $\mathbf{B}$ is given by

$$\widehat{\mathbf{B}} = (\mathbf{A}'\mathbf{\Sigma}\mathbf{A})^{-1}\mathbf{A}'\mathbf{\Sigma}\mathbf{X}\mathbf{C}'(\mathbf{C}\mathbf{C}')^{-1} = \begin{pmatrix}\widehat{\mathbf{b}}_1 & \widehat{\mathbf{b}}_2\end{pmatrix},$$

where

$$\widehat{\mathbf{b}}_i = (\mathbf{A}'\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{\Sigma}^{-1}\overline{\mathbf{x}}_i \sim N_q\left(\mathbf{b}_i, \frac{1}{n_i}(\mathbf{A}'\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}\right). \quad (9)$$

We will now give an useful lemma that we will use later in the paper.

Lemma 1. *Let* $\widehat{\mathbf{\Sigma}}$ *be given in* (8) *and suppose that all included inverses exist. Then*

$$\mathbf{A}'(n\widehat{\mathbf{\Sigma}})^{-1} = \mathbf{A}'\mathbf{S}^{-1}.$$

The proof of this lemma can be found, for example, in [23].

3. Asymptotic approximation of probabilities of misclassification

In this section we consider the approximations of the probabilities of misclassification using the linear classification function given in (7). While in general, it is hard to obtain the exact probability of misclassifications, there have been extensive studies for their asymptotic approximations including asymptotic expansions (see, for example, [5, 18, 26]). The main purpose of this section is to derive the approximations for the probabilities of misclassification through expressing the linear classification function in (7) as a location

and scale mixture of the standard normal distribution. The probabilities of misclassification by the linear classification function (7) are denoted

$$e(2|1) = \Pr(L \leq 0|\mathbf{x} \in \pi_1), \quad e(1|2) = \Pr(L > 0|\mathbf{x} \in \pi_2),$$

where $e(2|1)$ is the probability of allocating $\mathbf{x}$ of p repeated measurements into π_2, although it is known that they come from π_1 and similarly for $e(1|2)$. We are interested in deriving an asymptotic approximation of $e(2|1)$. Note that in this article, probabilities of misclassification are used interchangeably with misclassification errors.

3.1. Asymptotic approximation of the misclassification errors with known $\mathbf{\Sigma}$. In this subsection we assume that $\mathbf{\Sigma}$ is known. Suppose that the observation $\mathbf{x}$ of p repeated measurements is from π_1, the conditional distribution of $L_0 = L(\mathbf{x}; \widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \mathbf{\Sigma})$ given $(\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2)$ is $N(-U_0, V_0)$, that is, $\mathbb{E}[L_0|\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2] = -U_0$ and $\mathbb{V}ar(L_0|\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2) = V_0$, where

$$U_0 = (\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)'\mathbf{A}'\mathbf{\Sigma}^{-1}\mathbf{A}(\widehat{\mathbf{b}}_1 - \mathbf{b}_1) - \frac{1}{2}V_0, \tag{10}$$

$$V_0 = (\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)'\mathbf{A}'\mathbf{\Sigma}^{-1}\mathbf{A}(\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2). \tag{11}$$

Hence, L_0 given $\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2$ can now be expressed as a location and scale mixture of the standard normal distribution given by

$$L_0 = V_0^{1/2} Z_0 - U_0, \tag{12}$$

where

$$Z_0 = V_0^{-1/2}(\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)'\mathbf{A}'\mathbf{\Sigma}^{-1}(\mathbf{x} - \mathbf{A}\mathbf{b}_1).$$

Given $\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2$, Z_0 is obviously independent of (U_0, V_0) and is distributed as $\mathrm{N}(0, 1)$. The probability of misclassification where $\mathbf{x}$ is assigned to π_2, when it actually belongs to π_1 can be expressed using (12) as

$$\begin{aligned} e_0(2|1) &= \Pr(L_0 \leq 0|\mathbf{x} \in \pi_1) = \mathbb{E}_{U_0,V_0}[\mathbb{E}[\chi_{\{L_0 \leq 0\}}|U_0, V_0]] \\ &= \mathbb{E}_{U_0,V_0}[\Pr(\mathrm{L}_0 \leq 0|U_0, V_0)] = \mathbb{E}_{U_0,V_0}[\Pr(V_0^{1/2} Z_0 - U_0 \leq 0|U_0, V_0)] \\ &= \mathbb{E}_{U_0,V_0}[\Phi(V_0^{-1/2} U_0)], \end{aligned} \tag{13}$$

where $\chi_{\{\cdot\}}$ denotes the indicator function. As an approximation of (13), we propose

$$e_0(2|1) \simeq \Phi\big((\mathbb{E}[V_0])^{-1/2}\mathbb{E}[U_0]\big),$$

obtained by replacing U_0 and V_0 with $\mathbb{E}[U_0]$ and $\mathbb{E}[V_0]$ in a similar manner as was performed by [5] for the classical case.

Theorem 1. *The expectations of U_0 and V_0 defined in* (10) *and* (11) *equal*

$$\mathbb{E}[V_0] = \Delta^2 + \frac{n_1 + n_2}{n_1 n_2} q,$$

$$\mathbb{E}[U_0] = -\frac{1}{2}\left(\Delta^2 + \frac{n_1 - n_2}{n_1 n_2} q\right),$$

where Δ^2 *is the squared Mahalanobis distance*

$$\Delta^2 = (\mathbf{b}_1 - \mathbf{b}_2)' \mathbf{A}' \mathbf{\Sigma}^{-1} \mathbf{A} (\mathbf{b}_1 - \mathbf{b}_2). \tag{14}$$

Proof. Firstly, $\mathbb{E}[V_0]$ is derived. It is utilized that $\widehat{\mathbf{b}}_1$ and $\widehat{\mathbf{b}}_2$ are independently distributed. Since $\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2 \sim N_q\left(\mathbf{b}_1 - \mathbf{b}_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)(\mathbf{A}'\mathbf{\Sigma}^{-1}\mathbf{A})^{-1}\right)$, which follows from (9), we have

$$\begin{aligned}
\mathbb{E}[V_0] &= \mathbb{E}[(\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)' \mathbf{A}' \mathbf{\Sigma}^{-1} \mathbf{A} (\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)] \\
&= \operatorname{tr}\left(\mathbf{A}'\mathbf{\Sigma}^{-1}\mathbf{A}\mathbb{E}[(\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)(\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)']\right) \\
&= \operatorname{tr}\left(\mathbf{A}'\mathbf{\Sigma}^{-1}\mathbf{A}\left(\left(\frac{1}{n_1} + \frac{1}{n_2}\right)(\mathbf{A}'\mathbf{\Sigma}^{-1}\mathbf{A})^{-1} + (\mathbf{b}_1 - \mathbf{b}_2)(\mathbf{b}_1 - \mathbf{b}_2)'\right)\right) \\
&= (\mathbf{b}_1 - \mathbf{b}_2)' \mathbf{A}' \mathbf{\Sigma}^{-1} \mathbf{A} (\mathbf{b}_1 - \mathbf{b}_2) + \frac{n_1 + n_2}{n_1 n_2} q. \qquad (15)
\end{aligned}$$

Moreover, $\mathbb{E}[U_0]$ is calculated as

$$\mathbb{E}[U_0] = \mathbb{E}\Big[(\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)' \mathbf{A}' \mathbf{\Sigma}^{-1} \mathbf{A} (\widehat{\mathbf{b}}_1 - \mathbf{b}_1)\Big] - \frac{1}{2}\mathbb{E}[V_0],$$

where

$$\begin{aligned}
&\mathbb{E}[(\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)' \mathbf{A}' \mathbf{\Sigma}^{-1} \mathbf{A} (\widehat{\mathbf{b}}_1 - \mathbf{b}_1)] \\
&\quad = \mathbb{E}[\widehat{\mathbf{b}}_1' \mathbf{A}' \mathbf{\Sigma}^{-1} \mathbf{A} \widehat{\mathbf{b}}_1] - \mathbf{b}_1' \mathbf{A}' \mathbf{\Sigma}^{-1} \mathbf{A} \mathbf{b}_1 + \mathbf{b}_2' \mathbf{A}' \mathbf{\Sigma}^{-1} \mathbf{A} \mathbf{b}_1 - \mathbb{E}[\widehat{\mathbf{b}}_2' \mathbf{A}' \mathbf{\Sigma}^{-1} \mathbf{A} \widehat{\mathbf{b}}_1] \\
&\quad = \mathbb{E}[\widehat{\mathbf{b}}_1' \mathbf{A}' \mathbf{\Sigma}^{-1} \mathbf{A} \widehat{\mathbf{b}}_1] - \mathbf{b}_1' \mathbf{A}' \mathbf{\Sigma}^{-1} \mathbf{A} \mathbf{b}_1 = \frac{q}{n_1},
\end{aligned}$$

where the last equality follows from similar derivations as in (15) and where (9) has been used. Then,

$$\mathbb{E}[U_0] = \frac{q}{n_1} - \frac{1}{2}\mathbb{E}[V_0] = -\frac{1}{2}\left(\Delta^2 + \frac{n_1 - n_2}{n_1 n_2} q\right).$$

□

The following theorem appears.

Theorem 2. *For the linear classification function based on* (7) *with unknown* $\mathbf{b}_1, \mathbf{b}_2$ *and known* $\mathbf{\Sigma}$*, the misclassification errors can approximately be evaluated via*

$$e_0(2|1) \simeq \Phi\big(\gamma_0\big),$$

where

$$\gamma_0 = -\frac{1}{2}\frac{\Delta^2 + \frac{n_1 - n_2}{n_1 n_2} q}{\sqrt{\Delta^2 + \frac{n_1 + n_2}{n_1 n_2} q}},$$

with the squared Mahalanobis distance given in (14).

Similarly as in (2) one can see that if n_1 and n_2 tend to infinity, then $e_0(2|1) \simeq \Phi(-\Delta/2)$. Although we can obtain the approximation for the misclassification errors, we can not use $\mathbf{b}_1$ and $\mathbf{b}_2$ directly in the distance measure Δ^2 since they are usually unknown. Thus, when utilizing $e_0(2|1), \mathbf{b}_1$ and $\mathbf{b}_2$ are replaced by $\widehat{\mathbf{b}}_1$ and $\widehat{\mathbf{b}}_2$, respectively.

3.2. Asymptotic approximation of misclassification errors with unknown $\mathbf{\Sigma}$. In Subsection 3.1, we assumed that $\mathbf{\Sigma}$ was known. In this subsection we shall assume that all parameters of the populations $\pi_i,\ i = 1, 2$, are unknown. Hence, we want to study $L(\mathbf{x}; \widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \widehat{\mathbf{\Sigma}})$. However, from Lemma 1 we know that $\mathbf{A}'(n\widehat{\mathbf{\Sigma}})^{-1} = \mathbf{A}'\mathbf{S}^{-1}$, hence $L(\mathbf{x}; \widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \widehat{\mathbf{\Sigma}}) = nL(\mathbf{x}; \widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \mathbf{S})$, and since we are only interested in the sign of L in the classification rule (6) we will study $L(\mathbf{x}; \widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \mathbf{S})$ instead of $L(\mathbf{x}; \widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \widehat{\mathbf{\Sigma}})$ for simplicity. Moreover, the the sum of squares matrix $\mathbf{S}$ is Wishart distributed, whereas the distribution of the maximum likelihood-based estimator is unknown [23].

Theorem 3. *Assume that observation* $\mathbf{x}$ *comes from* π_1. *Then the statistic* $L(\mathbf{x}; \widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \mathbf{S})$ *can be expressed as*

$$L = V^{1/2}Z - U,$$

where

$$V = (\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)'\mathbf{A}'\mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1}\mathbf{A}(\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2), \tag{16}$$

$$Z = V^{-1/2}(\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)'\mathbf{A}'\mathbf{S}^{-1}(\mathbf{x} - \mathbf{A}\mathbf{b}_1),$$

$$U = (\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)'\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}(\widehat{\mathbf{b}}_1 - \mathbf{b}_1) - \frac{1}{2}\widetilde{V}, \tag{17}$$

and $\widetilde{V} = (\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)'\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}(\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)$ *is the squared sample Mahalanobis distance between two populations.*

The result (16) is obtained by noting that the conditional distribution of $(\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)'\mathbf{A}'\mathbf{S}^{-1}(\mathbf{x} - \mathbf{A}\mathbf{b}_1)$ given $\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \mathbf{S}$, is $N(0, V)$ when $\mathbf{x}$ comes from π_1. Given $\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \mathbf{S}$, we can see that Z follows a standard normal distribution, i.e., $Z \sim \mathrm{N}(0, 1)$, which is also conditionally true. Moreover, Z and (U, V) are independent. Analogously to (13), the probability of misclassification when $\mathbf{x}$ comes from π_1 is given by

$$e(2|1) = \Pr(L(\mathbf{x}; \widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \mathbf{S}) \leq 0 | \mathbf{x} \in \pi_1) = \mathbb{E}_{(U,V)}[\Phi(V^{-1/2}U)].$$

Again we consider a similar approach as in [5] to find the asymptotic approximation of the probability of misclassification:

$$e(2|1) \simeq \Phi\big((\mathbb{E}[V])^{-1/2}\mathbb{E}[U]\big). \tag{18}$$

To find the expectations in the probability of misclassification (18), we need the following lemma.

Lemma 2. *Let* $\mathbf{S} : p \times p$ *be a random matrix distributed according to* $W_p(n-2, \mathbf{\Sigma})$*, where* $\mathbf{\Sigma}$ *is positive definite. Assume that* $n-p-5>0$*, then*

$$
\begin{aligned}
&(i)\ \mathbb{E}\big[\mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1}\big] = (n-3)d_1\mathbf{\Sigma}^{-1},\\
&(ii)\ \mathbb{E}\big[\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\big] = d_2\mathbf{\Sigma}^{-1}\\
&\qquad - d_3(\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{\Sigma}\mathbf{A})^{-1}\mathbf{A}'\mathbf{\Sigma}^{-1}),\\
&(iii)\ \mathbb{E}\big[\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\big]\\
&\qquad = (n-3)d_1\mathbf{\Sigma}^{-1}\\
&\qquad\quad + \big[(n-3)d_4 - (n+p-2q-3)d_5\big](\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{\Sigma}\mathbf{A})^{-1}\mathbf{A}'\mathbf{\Sigma}^{-1}),
\end{aligned}
$$

where

$$
\begin{aligned}
d_1 &= \frac{1}{(n-p-2)(n-p-3)(n-p-5)},\\
d_2 &= \frac{1}{n-m-p-3},\\
d_3 &= \frac{1}{n-m-(p-q)-3},\\
d_4 &= \frac{1}{(n-(p-q)-2)(n-(p-q)-3)(n-(p-q)-5)},\\
d_5 &= \frac{1}{(n-q-2)(n-(p-q)-3)(n-q-5)},
\end{aligned}
$$

and all constants are supposed to exist.

Proof. For the proof of (i) see [2], p. 388 and for the proofs of (ii) and (iii) see, for example, [23], p. 447. □

We are now ready to derive the expectations in (18).

Theorem 4. *The expectations of* V *and* U *defined in* (16) *and* (17) *are as follows:*

$$
\mathbb{E}[V] = c_1\Delta^2 + \frac{n_1+n_2}{n_1n_2}(pc_1 + (p-q)c_2),
$$

and

$$
\mathbb{E}[U] = -\frac{1}{2}\Big(c_3\Delta^2 + \frac{n_1-n_2}{n_1n_2}((c_4-c_5)p + c_5q)\Big),
$$

respectively, where Δ^2 *is the squared Mahalanobis distance* (14)*,*

$$
\begin{aligned}
c_1 &= \frac{f-1}{(f-p)(f-p-1)(f-p-3)},\\
c_2 &= \frac{1}{f-(p-q)-1}\left(\frac{f-1}{(f-(p-q))(f-(p-q)-3)} - \frac{f+p-2q-1}{(f-q)(f-q-3)}\right),
\end{aligned}
$$

$$c_3 = \frac{1}{f-p-1}, \qquad c_4 = \frac{1}{f-m-p-1}, \qquad c_5 = \frac{1}{f-m-(p-q)-1},$$

with $f = n_1 + n_2 - 2$*, and all constants are supposed to exist.*

Proof. Note that $\mathbf{S}$ and $\widehat{\mathbf{B}} = \begin{pmatrix} \widehat{\mathbf{b}}_1 & \widehat{\mathbf{b}}_2 \end{pmatrix}$ are not independent. However, we have

$$\begin{aligned}\mathbb{E}[V] &= \mathbb{E}_{\mathbf{S}}\big[\mathbb{E}[V|\mathbf{S}]\big] = \mathbb{E}_{\mathbf{S}}\big[\mathbb{E}[(\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)'\mathbf{A}'\mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1}\mathbf{A}(\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)|\mathbf{S}]\big] \\ &= \mathbb{E}_{\mathbf{S}}\big[\mathrm{tr}\big(\mathbf{A}'\mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1}\mathbf{A}\mathbb{E}[(\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)(\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2)'|\mathbf{S}]\big)\big]. \qquad (19)\end{aligned}$$

The conditional covariance matrix is

$$\mathbb{C}ov(\widehat{\mathbf{B}}|\mathbf{S}) = (\mathbf{C}\mathbf{C}')^{-1} \otimes (\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}, \qquad (20)$$

where $\otimes$ denotes the Kronecker product and $\mathbb{C}ov(\cdot)$ denotes the covariance matrix. Result (20) can be found, for example, in [10]. For our choice of $\mathbf{C}$ in (4) we have

$$(\mathbf{C}\mathbf{C}')^{-1} = \begin{pmatrix} \frac{1}{n_1} & 0 \\ 0 & \frac{1}{n_2} \end{pmatrix},$$

and it follows that

$$\mathbb{C}ov(\widehat{\mathbf{b}}_i|\mathbf{S}) = \frac{1}{n_i}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}, \quad i = 1, 2.$$

From (19) and since

$$\begin{aligned}\widehat{\mathbf{b}}_1 - \widehat{\mathbf{b}}_2|\mathbf{S} \sim N_q\Big(&\mathbf{b}_1 - \mathbf{b}_2, \\ &\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\Big), \qquad (21)\end{aligned}$$

we have

$$\begin{aligned}&\mathbb{E}_{\mathbf{S}}\big[\mathbb{E}[V|\mathbf{S}]\big] \\ &= \mathbb{E}_{\mathbf{S}}\Big[\mathrm{tr}\Big(\mathbf{A}'\mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1}\mathbf{A}\Big(\Big(\frac{1}{n_1}+\frac{1}{n_2}\Big)(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1} \\ &\quad + (\mathbf{b}_1 - \mathbf{b}_2)(\mathbf{b}_1 - \mathbf{b}_2)'\Big)\Big)\Big] \\ &= \frac{n_1 + n_2}{n_1 n_2}\mathrm{tr}\big(\mathbf{\Sigma}\mathbb{E}_{\mathbf{S}}\big[\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\big]\big) \\ &\quad + (\mathbf{b}_1 - \mathbf{b}_2)'\mathbf{A}'\mathbb{E}_{\mathbf{S}}\big[\mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1}\big]\mathbf{A}(\mathbf{b}_1 - \mathbf{b}_2)\big]. \qquad (22)\end{aligned}$$

Using Lemma 2 (i) and (iii), with $\mathbf{S} \sim W_p(n_1 + n_2 - 2, \mathbf{\Sigma})$, we obtain

$$\begin{aligned}&\mathbb{E}_{\mathbf{S}}\big[\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\big] \\ &\quad = c_1\mathbf{\Sigma}^{-1} + c_2(\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{\Sigma}\mathbf{A})^{-1}\mathbf{A}'\mathbf{\Sigma}^{-1}),\end{aligned}$$

where $c_1 = (f-1)d_1, c_2 = (f-1)d_4 - (f+p-2q-1)d_5$ with d_1, d_4 and d_5 given in Lemma 2 and $f = n_1 + n_2 - 2$.

Hence, from (22) we have

$$\begin{aligned}\mathbb{E}[V] &= \mathbb{E}_{\mathbf{S}}\big[\mathbb{E}[V|\mathbf{S}]\big] \\ &= \frac{n_1+n_2}{n_1 n_2}\mathrm{tr}\left(\boldsymbol{\Sigma}(c_1\boldsymbol{\Sigma}^{-1} + c_2(\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{A}(\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A})^{-1}\mathbf{A}'\boldsymbol{\Sigma}^{-1})\right) \\ &\quad + c_1(\mathbf{b}_1-\mathbf{b}_2)'\mathbf{A}'\boldsymbol{\Sigma}^{-1}\mathbf{A}(\mathbf{b}_1-\mathbf{b}_2) \\ &= \frac{n_1+n_2}{n_1 n_2}(pc_1 + (p-q)c_2) + c_1(\mathbf{b}_1-\mathbf{b}_2)'\mathbf{A}'\boldsymbol{\Sigma}^{-1}\mathbf{A}(\mathbf{b}_1-\mathbf{b}_2).\end{aligned}$$

Next $\mathbb{E}[U]$ can be derived as

$$\begin{aligned}\mathbb{E}[U] = \mathbb{E}_{\mathbf{S}}\big[\mathbb{E}[U|\mathbf{S}]\big] &= \mathbb{E}_{\mathbf{S}}\Big[\mathbb{E}\Big[(\widehat{\mathbf{b}}_1-\widehat{\mathbf{b}}_2)'\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}(\widehat{\mathbf{b}}_1-\mathbf{b}_1) - \frac{1}{2}\widetilde{V}|\mathbf{S}\Big]\Big] \\ &= \mathbb{E}_{\mathbf{S}}\big[\mathbb{E}\big[(\widehat{\mathbf{b}}_1-\widehat{\mathbf{b}}_2)'\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}(\widehat{\mathbf{b}}_1-\mathbf{b}_1)|\mathbf{S}\big]\big] - \frac{1}{2}\mathbb{E}_{\mathbf{S}}\big[\mathbb{E}[\widetilde{V}|\mathbf{S}]\big],\end{aligned}$$

where $\widetilde{V} = (\widehat{\mathbf{b}}_1-\widehat{\mathbf{b}}_2)'\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}(\widehat{\mathbf{b}}_1-\widehat{\mathbf{b}}_2)$. Note that $\widehat{\mathbf{b}}_1$ and $\widehat{\mathbf{b}}_2$ are unbiased and given $\mathbf{S}$ they are independently normally distributed. Consider

$$\begin{aligned}&\mathbb{E}_{\mathbf{S}}\big[\mathbb{E}[(\widehat{\mathbf{b}}_1-\widehat{\mathbf{b}}_2)'\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}(\widehat{\mathbf{b}}_1-\mathbf{b}_1)|\mathbf{S}]\big] \\ &= \mathbb{E}_{\mathbf{S}}\big[\mathbb{E}[\widehat{\mathbf{b}}_1'\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}\widehat{\mathbf{b}}_1|\mathbf{S}]\big] - \mathbb{E}_{\mathbf{S}}\big[\mathbb{E}[\widehat{\mathbf{b}}_1'\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}\mathbf{b}_1|\mathbf{S}]\big] \\ &= \mathbb{E}_{\mathbf{S}}\big[\mathrm{tr}\big(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}\mathbb{E}[\widehat{\mathbf{b}}_1\widehat{\mathbf{b}}_1'|\mathbf{S}]\big] - \mathbb{E}_{\mathbf{S}}[\mathbf{b}_1'\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}\mathbf{b}_1] \\ &= \mathbb{E}_{\mathbf{S}}\Big[\mathrm{tr}\Big(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}\Big(\frac{1}{n_1}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\boldsymbol{\Sigma}\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1} + \mathbf{b}_1\mathbf{b}_1'\Big)\Big)\Big] \\ &\quad - \mathbb{E}_{\mathbf{S}}[\mathbf{b}_1'\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}\mathbf{b}_1] \\ &= \frac{1}{n_1}\mathrm{tr}\big(\boldsymbol{\Sigma}\mathbb{E}_{\mathbf{S}}\big[\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\big]\big) = \frac{1}{n_1}((c_4-c_5)p + c_5 q), \qquad (23)\end{aligned}$$

since $\widehat{\mathbf{b}}_1|\mathbf{S} \sim N_q\Big(\mathbf{b}_1, \frac{1}{n_1}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\boldsymbol{\Sigma}\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\Big)$, and where we have used Lemma 2 (ii) in the last equality. Now using (21), one can see that $\mathbb{E}_{\mathbf{S}}\big[\mathbb{E}[\widetilde{V}|\mathbf{S}]\big]$ equals

$$\begin{aligned}&\mathbb{E}_{\mathbf{S}}\big[\mathbb{E}[\widetilde{V}|\mathbf{S}]\big] = \mathbb{E}_{\mathbf{S}}\big[\mathrm{tr}\big(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}\mathbb{E}[(\widehat{\mathbf{b}}_1-\widehat{\mathbf{b}}_2)(\widehat{\mathbf{b}}_1-\widehat{\mathbf{b}}_2)'|\mathbf{S}]\big)\big] \\ &= \mathbb{E}_{\mathbf{S}}\Big[\mathrm{tr}\Big(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A}\Big(\Big(\frac{1}{n_1}+\frac{1}{n_2}\Big)(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\boldsymbol{\Sigma}\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1} \\ &\quad + (\mathbf{b}_1-\mathbf{b}_2)(\mathbf{b}_1-\mathbf{b}_2)'\Big)\Big)\Big] \\ &= \frac{n_1+n_2}{n_1 n_2}\mathrm{tr}\big(\boldsymbol{\Sigma}\mathbb{E}_{\mathbf{S}}\big[\mathbf{S}^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^{-1}\big]\big) \\ &\quad + (\mathbf{b}_1-\mathbf{b}_2)'\mathbf{A}'\mathbb{E}_{\mathbf{S}}\big[\mathbf{S}^{-1}\big]\mathbf{A}(\mathbf{b}_1-\mathbf{b}_2)\big]\end{aligned}$$

$$= c_3(\mathbf{b}_1 - \mathbf{b}_2)'\mathbf{A}'\boldsymbol{\Sigma}^{-1}\mathbf{A}(\mathbf{b}_1 - \mathbf{b}_2) + \frac{n_1 + n_2}{n_1 n_2}((c_4 - c_5)p + c_5 q), \tag{24}$$

where $c_3 = \dfrac{1}{f - p - 1}$ and where Lemma 2 has been used in the last equality. Combining (23) and (24) we get

$$\mathbb{E}[U] = -\frac{1}{2}\Big(c_3(\mathbf{b}_1 - \mathbf{b}_2)'\mathbf{A}'\boldsymbol{\Sigma}^{-1}\mathbf{A}(\mathbf{b}_1 - \mathbf{b}_2) + \frac{n_1 - n_2}{n_1 n_2}((c_4 - c_5)p + c_5 q)\Big).$$

□

From the results above we are now ready to give the following theorem and the main result of this paper.

Theorem 5. *For the linear classification function based on* (7) *with unknown* $\mathbf{b}_1, \mathbf{b}_2$ *and* $\boldsymbol{\Sigma}$*, the misclassification errors can approximately be evaluated via*

$$e(2|1) \simeq \Phi\big(\gamma\big),$$

where

$$\gamma = -\frac{1}{2}\frac{c_3\Delta^2 + \frac{n_1 - n_2}{n_1 n_2}((c_4 - c_5)p + c_5 q)}{\sqrt{c_1\Delta^2 + \frac{n_1 + n_2}{n_1 n_2}(pc_1 + (p - q)c_2)}},$$

with $c_1, \ldots, c_5$ *defined in Theorem 4 and where* Δ^2 *is the squared Mahalanobis distance* (14)*.*

Again one can compare with (2) and also Theorem 2. Even if it is not so straightforward, we can similarly see that if n_1 and n_2 tend to infinity, then $e(2|1) \simeq \Phi(-\Delta/2)$ as expected.

4. Simulation study

The approximation of misclassification errors in case of repeated measures observations that follow a growth curve structure [19] on the means have not been proposed before. In Theorems 2 and 5 we gave approximations for these misclassification errors first when we have assumed $\boldsymbol{\Sigma}$ to be known and then when it is unknown.

In this section a simulation study is performed to examine the reliability of the approximation of the misclassification errors proposed in Theorems 2 and 5. We compare the errors given by the approximations with the relative frequencies, which are the number of times an observation is misclassified to π_2 while it actually comes from π_1, using classification rule (6) both for unknown and known $\boldsymbol{\Sigma}$. As earlier the mean parameters $\mathbf{B}$ is always unknown, i.e., estimated. In our Monte Carlo simulations, we let $n = 80$ and assume for simplicity $n_1 = n_2 = \frac{n}{2}$. Further, let $t_i = t_1 + (i - 1)\frac{t_p - t_1}{p-1}$, with $t_1 = 0.50$ and $t_p = 2.50$, i.e., $(t_1, \ldots, t_p)$ are evenly spread out in the interval $[0.50,\ 2.50]$. Also, let $p \in \{10, 20, 30, 40, 50, 60, 70, 72, 74, 100, 120\}$

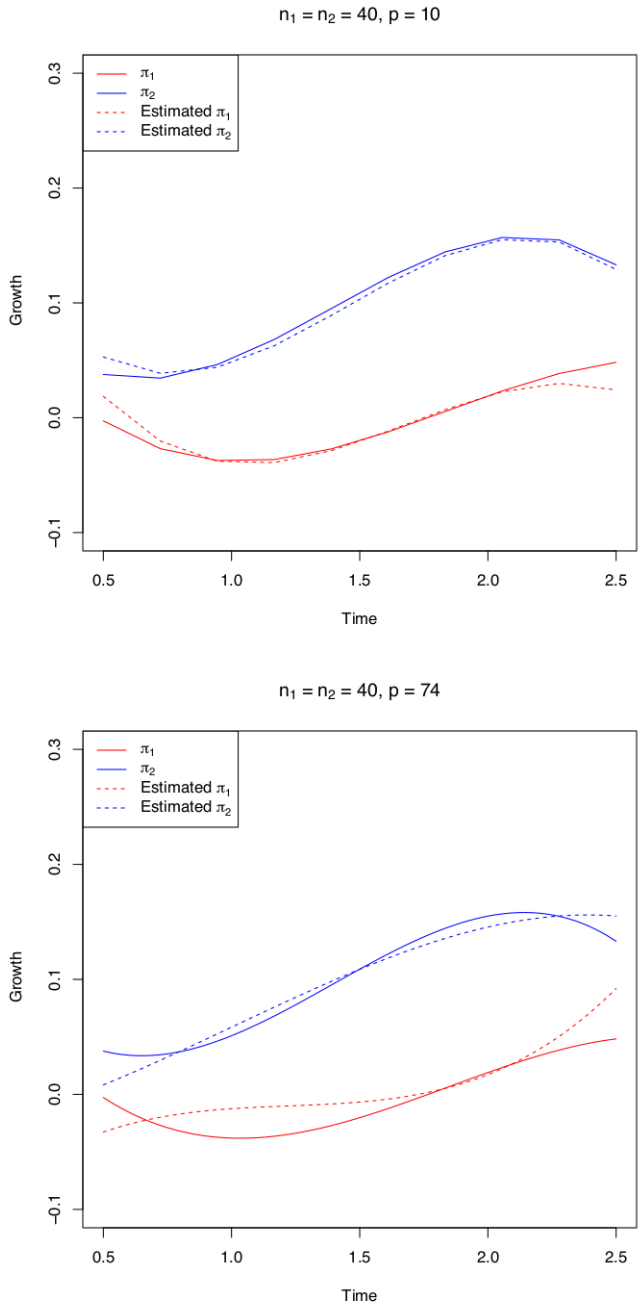

FIGURE 1. The third order growth curves describe the sample mean per group (solid lines) and the estimated mean growth curves (dashed lines) for the populations π_1 and π_2. In the upper plot $p = 10$ whereas in the lower plot $p = 74$.

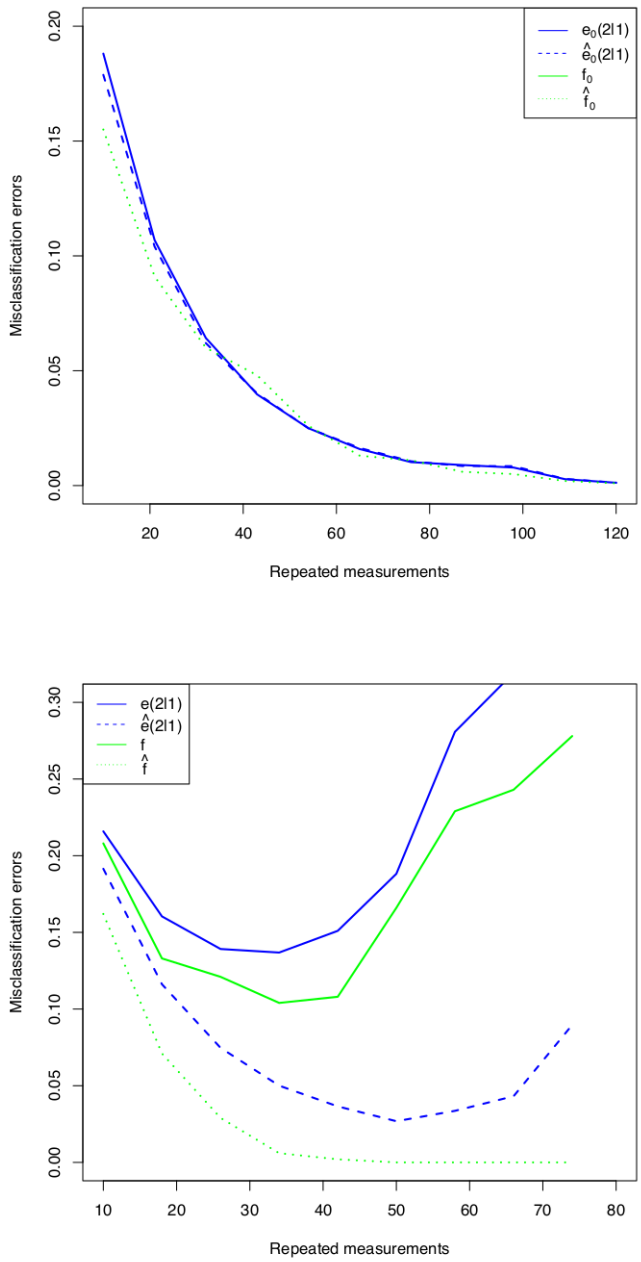

FIGURE 2. (a) True and estimated approximations for misclassification errors computed using Theorem 2 and the relative frequencies f_0 and $\widehat{f_0}$ for known $\mathbf{\Sigma}$ calculated using classification rule (6) (b) True and estimated approximations for misclassification errors calculated using Theorem 5 and the relative frequencies f and $\widehat{f}$, computed using classification rule (6) when $\mathbf{\Sigma}$ is unknown.

p	$e_0(2\|1)$	f_0	$\widehat{e}_0(2\|1)$	$\widehat{f}_0$
10	0.185	0.181	0.180	0.155
20	0.104	0.104	0.102	0.091
30	0.062	0.057	0.061	0.060
40	0.038	0.024	0.037	0.048
50	0.023	0.020	0.024	0.026
60	0.015	0.013	0.015	0.013
70	0.009	0.007	0.010	0.011
72	0.008	0.006	0.008	0.006
74	0.007	0.006	0.007	0.005
100	0.002	0.003	0.003	0.002
120	0.001	0.003	0.001	0.001

TABLE 1. In the table $e_0(2|1)$ and $\widehat{e}_0(2|1)$ are the values of the "true" and estimated approximations of misclassification errors, computed using Theorem 2 for $\mathbf{b}_1, \mathbf{b}_2, \mathbf{\Sigma}$ and $\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \mathbf{\Sigma}$, respectively. The relative frequencies f_0 and $\widehat{f}_0$ are calculated as the relative number of events $\{L(\mathbf{x}; \widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \mathbf{\Sigma}) \leq 0\}$ with observation $\mathbf{x}$ generated as $\mathbf{x} \sim N_p(\mathbf{A}\mathbf{b}_1, \mathbf{\Sigma})$ or $\mathbf{x} \sim N_p(\mathbf{A}\widehat{\mathbf{b}}_1, \mathbf{\Sigma})$, respectively.

for unknown $\mathbf{b}_1, \mathbf{b}_2$ and known $\mathbf{\Sigma}$ and $p \in \{10, 20, 30, 40, 50, 60, 70, 72, 74\}$ for unknown $\mathbf{b}_1, \mathbf{b}_2, \mathbf{\Sigma}$, since we must have $p \leq n - 2$. Data $\mathbf{X} : p \times n$ are generated using the Growth Curve model $\mathbf{X} = \mathbf{ABC}+\mathbf{E}$, $\mathbf{E} \sim N_{p,n}(\mathbf{0}, \mathbf{\Sigma}, \mathbf{I})$, where the design and parameter matrices are respectively given as

$$\mathbf{A} = \begin{pmatrix} 1 & t_1 & t_1^2 & t_1^3 \\ 1 & t_2 & t_2^2 & t_2^3 \\ \vdots & \vdots & & \vdots \\ 1 & t_p & t_p^2 & t_p^3 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} \mathbf{1}'_{40} & \mathbf{0}'_{40} \\ \mathbf{0}'_{40} & \mathbf{1}'_{40} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 0.117 & 0.125 \\ -0.345 & -0.313 \\ 0.232 & 0.314 \\ -0.042 & -0.075 \end{pmatrix}.$$

Furthermore, for $\mathbf{\Sigma}$ we have $\mathbf{\Sigma} = \mathbf{DRD}$, where $\mathbf{D} = diag(\sigma_1, \ldots, \sigma_p), \sigma_i = \sqrt{0.1 + (i-1)d}$, with $d = \frac{1.9}{p-1}$ for $i = 1, \ldots, p$, and $\mathbf{R} = (\rho_{ij})$, where $\rho_{ij} = (-1)^{i+j} r^{|i-j|^\gamma}$, with $r = 0.2$, $\gamma = 0.1$ for $j = 1, \ldots, p$.

In Figure 1, the third order growth curves are given which show the true mean growth profiles which were used when data was simulated and the estimated growth profiles for two populations π_1 and π_2. Note here that Figure 1 is produced when $\mathbf{b}_1, \mathbf{b}_2, \mathbf{\Sigma}$ are unknown. In the upper plot, the dashed lines seem to be close to the solid lines for $p = 10$. This means that the mean growth profile is well estimated with $p = 10$ repeated measurements. However, on the lower plot, the discrepancy between lines (dashed and solid lines) become considerably large as the number of repeated measurements

increases. This means that the mean growth profile is poorly estimated with a large number of repeated measurements, i.e., p close to n.

p	$e(2\|1)$	f	$\widehat{e}(2\|1)$	$\widehat{f}$
10	0.216	0.208	0.188	0.162
20	0.160	0.133	0.117	0.071
30	0.139	0.121	0.074	0.029
40	0.137	0.104	0.051	0.006
50	0.151	0.108	0.035	0.002
60	0.188	0.166	0.028	0.000
70	0.281	0.229	0.032	0.000
72	0.319	0.243	0.041	0.000
74	0.385	0.278	0.090	0.000

TABLE 2. In the table $e(2|1)$ and $\widehat{e}(2|1)$ are the values of the "true" and estimated approximations of misclassification errors, computed using Theorem 5 for $\mathbf{b}_1, \mathbf{b}_2, \mathbf{\Sigma}$ and $\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \widehat{\mathbf{\Sigma}}$, respectively. The relative frequencies f and $\widehat{f}$ are calculated as the relative number of events $\{L(\mathbf{x}; \widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \widehat{\mathbf{\Sigma}}) \leq 0\}$ with observation $\mathbf{x}$ generated as $\mathbf{x} \sim N_p(\mathbf{A}\mathbf{b}_1, \mathbf{\Sigma})$ or $\mathbf{x} \sim N_p(\mathbf{A}\widehat{\mathbf{b}}_1, \widehat{\mathbf{\Sigma}})$, respectively.

In Table 1 the approximations and the relative frequencies of the misclassification errors are investigated for p repeated measurements in the case when $\mathbf{\Sigma}$ is known. As before, $e_0(2|1)$ denotes the "true" approximation of the misclassification errors and $\widehat{e}_0(2|1)$ denotes the estimated approximation of misclassification errors, where we have plugged in the estimates for the mean parameters. Furthermore, f_0 and $\widehat{f}_0$ stand for relative frequencies of misclassifications computed using classification rule (6) for known $\mathbf{\Sigma}$. These relative frequencies are calculated, based on 10,000 simulations, as the relative number of events $\{L(\mathbf{x}; \widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \mathbf{\Sigma}) \leq 0\}$ with observation $\mathbf{x}$ generated as $\mathbf{x} \sim N_p(\mathbf{A}\mathbf{b}_1, \mathbf{\Sigma})$ or $\mathbf{x} \sim N_p(\mathbf{A}\widehat{\mathbf{b}}_1, \mathbf{\Sigma})$, respectively. In real life we do not know $\mathbf{b}_1$ and must then trust $\widehat{\mathbf{b}}_1$ if any.

In Table 1 the values of the estimated misclassification errors $\widehat{e}_0(2|1)$ are closer to the true misclassification errors $e_0(2|1)$ when more information (repeated measurements) were included, and the misclassification errors become smaller for larger p. This means that the more information the smaller are the misclassification errors. This can be seen in Figure 2 (upper plot). It can be noted that there are values of misclassification errors for $p > n - 2$, which we can have because $\mathbf{\Sigma}$ is known.

In Table 2 we present results for the approximation of misclassification errors for the case when $\mathbf{\Sigma}$ is unknown. As earlier, $e(2|1)$ denotes the "true" approximation for the misclassification errors and $\widehat{e}(2|1)$ denotes the estimated approximation for the misclassification errors. Now the relative frequencies f and $\widehat{f}$ are calculated as the relative number of events $\{L(\mathbf{x}; \widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_2, \widehat{\mathbf{\Sigma}}) \leq 0\}$ with observation $\mathbf{x}$ generated as $\mathbf{x} \sim N_p(\mathbf{A}\mathbf{b}_1, \mathbf{\Sigma})$ or $\mathbf{x} \sim N_p(\mathbf{A}\widehat{\mathbf{b}}_1, \widehat{\mathbf{\Sigma}})$, respectively.

In Table 2 the values of the estimated misclassification errors decrease when numbers of repeated measurements are relatively small as for $p = 10$ through $p = 60$. For large numbers of repeated measurements which are closer to the sample size, the misclassification errors increase, see for example when $p \in \{70, 72, 74\}$. This is due to the sample variance-covariance matrix which is not a good estimator when the number of repeated measurements gets larger since it becomes unstable. Remember that the classification function (7) includes the inverse of the variance-covariance matrix. Thus, the misclassification errors increase. The consequence when the number of repeated measurements is close to the sample size can also be seen in Figure 1, where the fitted growths are poor for $p = 74$[1].

Also, in the last column in Table 2, the relative frequency $\widehat{f}$ has zero values for example from $p = 60$ through $p = 74$, it is because a new observation in the simulation is generated based on the estimated parameters instead of the true $\mathbf{b}_1$ and $\mathbf{\Sigma}$, i.e., the new observation $\mathbf{x}$ to be classified is generated using the *wrong* model which also has the same values as used in the classification rule (6).

We conclude that, the proposed approximation can be suggested for use when the number of repeated measurements is not too close to the sample size. The simulation results pave the way for one to propose new estimators and investigate the case when the number of repeated measurements is comparable to sample size or exceeds it.

5. Summary

In this paper we have considered the linear classification function when the means follow the Growth Curve model given by [19]. The linear classification function can assign a new observation of p repeated measurements to one of two specified groups. Given a classification rule it is natural to enquiry how well the decision rule can appropriately classify a new observation of p repeated measurements. In general, it is hard to obtain the exact expression for the probability of misclassification. We express the linear discriminant

[1]Remark: This is a known phenomena but not easy to overcome. One way could be by regularization, i.e., Tikhonov regularization. However, with regularization the distributional properties are hard to derive and will be considered in future research

function as a location and scale mixture of the standard normal distribution and derive approximations for the probability of misclassification.

It seems that larger p is better for classification when $\mathbf{\Sigma}$ is known, but when $\mathbf{\Sigma}$ is unknown and p is close to n we have a problem with the instability for the sum of squares matrix $\mathbf{S}$. If $p > n-2$, then $\mathbf{S}$ is singular and a regular inverse cannot be taken.

Acknowledgements

The research of Edward K. Ngailo has been supported by the Sida-funded bilateral sub-program 'Capacity Building of Mathematics in Higher Education Institutions in Tanzania' and Dietrich von Rosen has been supported by the Swedish Research Council (2017-03003).

The authors would also like to thank the anonymous reviewer of this paper for many valuable and helpful comments and suggestions.

References

[1] T. P. Burnaby, *Growth-invariant discriminant functions and generalized distances*, Biometrics, **22** (1966), 96–110.
[2] S. Das Gupta, *Some aspects of discrimination function coefficients*, Sankhyā Ser. A **30** (1968), 387–400.
[3] R. A. Fisher, *The use of multiple measurements in taxonomic problems*, Ann. Eugenics **7** (1936), 179–188.
[4] R. A. Fisher, *The statistical utilization of multiple measurements*, Ann. Eugenics **8** (1938), 376–386.
[5] Y. Fujikoshi, *Error bounds for asymptotic approximations of the linear discriminant function when the sample sizes and dimensionality are large*, J. Multivariate Anal. **73** (2000), 1–17.
[6] H. Hotelling, *The generalization of Students ratio*, Ann. Eugenics **2** (1931), 360–378.
[7] M. Hyodo, T. Mitani, T. Himeno, and T. Seo, *Approximate interval estimation for EPMC for improved linear discriminant rule under high dimensional frame work*. SUT J. Math. **51** (2015), 145–179.
[8] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, Springer Science and Business Media, New York, 2013.
[9] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, Upper Saddle River, New Jersey, 2007.
[10] T. Kollo and D. von Rosen, *Advanced Multivariate Statistics with Matrices*, Springer, Dordrecht, 2005.
[11] J. C. Lee, *Bayesian classification of data from growth curves*, South African Statist. J. **11** (1977), 155–166.
[12] J. C. Lee, *Classification of growth curves*, Handbook of Statistics **2** (1982), 121–137.
[13] L. Lix and T. Sajobi, *Discriminant analysis for repeated measures data: a review*, Frontiers in Psychology **1** (2010), 1–9.
[14] D. J. Nagel, *Bayesian classification estimation and prediction of growth curves*, South African Statist. J. **13** (1979), 127–137.
[15] G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley & Sons, New Jersey, 2004.

[16] P. C. Mahalanobis, *On the generalized distance in statistics*, Proc. Nat. Inst. Sci. India **2** (1936), 49–55.
[17] G. B. Mentz and A. M. Kshirsagar, *Classification using growth curves*, Comm. Statist. Theory Methods **33** (2005), 2487–2502.
[18] M. Okamoto, *An asymptotic expansion for the distribution of the linear discriminant function*, Ann. Math. Statist. **34** (1963), 1286–1301.
[19] R. F. Potthoff and S. N. Roy, *A generalized multivariate analysis of variance model useful especially for Growth Curve problems*, Biometrika **51** (1964), 313–326.
[20] C. R. Rao, *Discriminant function between composite hypotheses and related problems*, Biometrika **53** (1966), 339–345.
[21] C. R. Rao, *Linear Statistical Inference and its Applications*, Wiley, New York, 1973.
[22] A. C. Rencher, *Multivariate Statistical Inference and Applications*, Wiley, New York, 1998.
[23] D. von Rosen, *Bilinear Regression Analysis: An Introduction*, Springer, New York, 2018.
[24] A. Roy and R. Khattree, *Discrimination and classification with repeated measures data under different covariance structures*, Comm. Statist. Simulation Comput. **34** (2005), 167–178.
[25] A. Roy and R. Khattree, *On discrimination and classification with multivariate repeated measures data*, J. Statist. Plann. Inference **134** (2005), 462–485.
[26] M. Siotani, *Large sample approximations and asymptotic expansions of classification statistics*, Handbook of Statistics **2** (1982), 61–100.
[27] M. S. Srivastava and C. G. Khatri, *An Introduction to Multivariate Statistics*, North Holland, New York, 1979.
[28] H. Watanabe, M. Hyodo, T. Seo, and T. Pavlenko, *Asymptotic properties of the misclassification rates for euclidean distance discriminant rule in high-dimensional data*, J. Multivariate Anal. **140** (2015), 234–244.

Department of Mathematics, University of Dar Es Salaam, Box 2329, Dar Es Salaam, Tanzania.
E-mail address: edward.ngailo@duce.ac.tz

Department of Energy and Technology, Swedish University of Agricultural Sciences, Box 7032, SE–750 07 Uppsala, Sweden, and Department of Mathematics, Linköping University, SE–581 83 Linköping, Sweden.
E-mail address: dietrich.von.rosen@slu.se

Department of Mathematics, Linköping University, SE–581 83 Linköping, Sweden.
E-mail address: martin.singull@liu.se