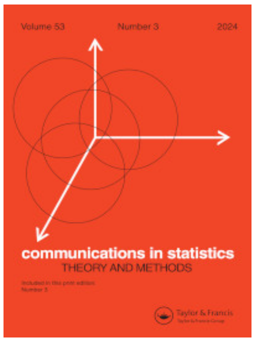

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/lsta20

# Moments of the likelihood-based discriminant function

Emelyne Umunoza Gasana, Dietrich von Rosen & Martin Singull

**To cite this article:** Emelyne Umunoza Gasana, Dietrich von Rosen & Martin Singull (2024) Moments of the likelihood-based discriminant function, Communications in Statistics - Theory and Methods, 53:3, 1122-1134, DOI: 10.1080/03610926.2022.2100909

**To link to this article:** https://doi.org/10.1080/03610926.2022.2100909

Published online: 22 Jul 2022.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2024, VOL. 53, NO. 3, 1122–1134
https://doi.org/10.1080/03610926.2022.2100909

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS Check for updates

# Moments of the likelihood-based discriminant function

Emelyne Umunoza Gasana[a,b], Dietrich von Rosen[a,c], and Martin Singull[a]

[a]Department of Mathematics, Linköping University, Linköping, Sweden; [b]Department of Mathematics, University of Rwanda, Kigali, Rwanda; [c]Department of Energy and Technology, Swedish University of Agricultural Sciences, Uppsala, Sweden

**ABSTRACT**
The likelihood approach used in this paper leads to quadratic discriminant functions. Classification into one of two known multivariate normal populations with a known and unknown covariance matrix are separately considered, where the two cases depend on the sample size and an unknown squared Mahalanobis distance. Their exact distributions are complicated to obtain. Therefore, moments for the likelihood based discriminant functions are established to express the basic characteristics of respective distribution.

**ARTICLE HISTORY**
Received 30 October 2021
Accepted 6 July 2022

**KEYWORDS**
classification; discriminant function; maximum likelihood; moments

## 1. Introduction

Over the years many 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. The distribution of the proposed discriminant functions are usually 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 paper, two alternative classification functions are considered under assumption about multivariate normality of the populations and adopting the likelihood approach. For one approach a known covariance matrix is supposed to hold and for the second approach an unknown covariance matrix is handled. Assume that we have $\pi_i$, $i \in \{1, \ldots, q\}$, populations, in classification methodology the aim is to allocate a $p$-dimensional random vector $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ to one of these populations by minimizing the error of misclassification. Pearson (1915, 1926), Mahalanobis (1925, 1930), Barnard (1935), Fisher (1936, 1938) and Rao (1948, 1966) introduced a classification rule for discriminating between two normal multivariate populations $\pi_1$ or $\pi_2$, the covariance matrix being the same, but with different mean vectors. See McLachlan (1992) for a general reference to discriminant analysis.

There exists many different techniques of deriving discriminant functions. The most widely used is the plug-in approach, that is, to derive the classification function by simply replacing the unknown parameters in the classification function with their

**CONTACT** Emelyne Umunoza Gasana emelyne.umunoza.gasana@liu.se; e.umunoza@ur.ac.rw Department of Mathematics, Linköping University, Linköping, Sweden.

estimators. This approach leads to a linear discriminant function. The most well known linear classification rule is the Wald-Anderson's rule, proposed by Wald (1944) and Anderson (1951), commonly known as the W-rule:

$$W(x) = (\bar{x}_1 - \bar{x}_2)^{\mathsf{T}} S^{-1}\boldsymbol{x} - \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^{\mathsf{T}} S^{-1}(\bar{x}_1 + \bar{x}_2), \tag{1.1}$$

where $S \sim W_p(n, \boldsymbol{\Sigma})$, the $p$-variate central Wishart distribution with $n$ degrees of freedom and covariance matrix $\boldsymbol{\Sigma}$ and the rule is that the new observation $x$ is classified as coming from $\pi_1$ if $W(x) > 0$ and from $\pi_2$ otherwise (Anderson 1951).

An alternative approach to derive a discriminant function when the populations are normally distributed and have unknown parameters is a likelihood approach as presented by Srivastava and Khatri (1979). The maximum likelihood discriminant rule consist of allocating an observation $\boldsymbol{x}$ into population $\pi_i$ if $f_i = \max_{1 \le j \le q} f_j(\boldsymbol{x})$ where $f_j(x)$ stands for the density connected to the $j^{th}$ population, Day and Kerridge (1967). This approach was introduced by Kudo (1959, 1960) as the $Z$-rule; a maximum likelihood criterion as an alternative to Wald-Anderson's $W$-rule. In this paper, only two populations $\pi_1$ and $\pi_2$ will be considered.

Unfortunately, the exact distribution of the classification function is often too complicated to allow for easy numerical calculations of obtaining misclassification probabilities. This has been pointed out by Sitgreaves (1961), see also Fujikoshi, Ulyanov, and Shimizu (2011). The use of asymptotic expansions is one way to address this problem; see Bowker and Sitgreaves (1961) and Okamoto (1963). Moments and cumulants of a function play an important role for expressing characteristic properties of its distribution as they can be used to approximate the distribution for example via an Edgeworth expansion. Critchley and Ford (1984), for example, obtained the variance of the estimated linear discriminant function and Davis (1987) derived the first four central moments and carried out asymptotic expansions for the cumulants of Wald-Anderson's linear discriminant function. The expected value and variance of $W$ are given by

$$E[W] = \frac{1}{2}\frac{n}{m}\left\{\Delta^2 + p\left(\frac{1}{n_2} - \frac{1}{n_1}\right)\right\},$$

$$\begin{aligned}\left(\frac{n}{m}\right)^2 (m-2)(m+1)Var[W] &= \frac{1}{2}(m+1)\Delta^4 + \Delta^2(m+1)\left\{m\left(1 + \frac{1}{n_2}\right) + \left(\frac{1}{n_1} - \frac{1}{n_2}\right)\right\} \\ &\quad + p(m+p)\left\{m\left(\frac{1}{n_2} + \frac{1}{n_1}\right) + \frac{1}{2}(m+1)\left(\frac{1}{n_1} + \frac{1}{n_2}\right) - \frac{1}{n_1 n_2}\right\},\end{aligned}$$

where $m = n_1 + n_2 - p - 3$ and $\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ is the Mahalanobis squared distance. These results have been obtained by Schaafsma (1982) and Davis (1987).

The maximum likelihood-based discriminant functions are derived in Section 3 and the first two moments of the discriminant functions are given explicitly in Section 4.

All vectors are column vectors. Bold lower case letters are used to denote vector valued variables and bold upper case letters are used for matrix valued variables.

## 2. Useful definitions and technical results

In this section, we will define some concepts and theorems applied in derivations of moments of the likelihood-based discriminant function.

**Definition 2.1.**

(i) The vectorization of a $m \times n$ matrix $\boldsymbol{A} = (a_{ij})$ is the $mn \times 1$ column vector

$$\mathrm{vec}\boldsymbol{A} = (a_{11}, \ldots, a_{m1}, a_{12}, \ldots, a_{m2}, \ldots, a_{1n}, \ldots, a_{mn})^{\mathsf{T}}.$$

(ii) The Kronecker product of $\boldsymbol{A} = (a_{ij})$ and $\boldsymbol{B} = (b_{kl})$ is defined by $\boldsymbol{A} \otimes \boldsymbol{B} = (a_{ij}\boldsymbol{B})$.

**Theorem 2.2.** *Let $\boldsymbol{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$, and let $\boldsymbol{A}$ be a $p \times p$ constant matrix, the $h^{th}$ cumulant $(\psi_h)$ of the quadratic form $\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}$ equals*

$$\psi_h(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}) = 2^{h-1}h!\left\{\frac{\mathrm{tr}\{(\boldsymbol{A\Sigma})^h\}}{h} + \boldsymbol{\mu}^{\mathsf{T}}(\boldsymbol{A\Sigma})^{h-1}\boldsymbol{A}\boldsymbol{\mu}\right\}.$$

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

From Theorem 2.2, the first and second cumulants of the quadratic form are directly deduced in the new corollary.

**Corollary 2.3.** *Let $\boldsymbol{x} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\boldsymbol{A}$ be a constant matrix. Then*

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

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

**Definition 2.4.**

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

(ii) The random matrix $\boldsymbol{W}^{-1}$ is said to follow an inverted Wishart distribution.

**Definition 2.5.** The partitioned matrix $\boldsymbol{K}_{p,q} : pq \times pq$ consisting of $q \times p$-blocks is called commutation matrix, if

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

Definitions 2.4 and 2.5 can be found in Kollo and von Rosen (2005).

**Theorem 2.6.** *Let $\boldsymbol{W} \sim W_p(\Sigma, n)$ and let $\boldsymbol{A} : p \times p$ be a constant matrix. Then*

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

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

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

*where* $c_1 = \frac{1}{n-p-1}$, $n - p - 1 > 0$; $c_2 = \frac{n-p-2}{(n-p)(n-p-1)(n-p-3)}$, $n - p - 3 > 0$ *and* $c_3 = \frac{1}{n-p-2}c_2$, $n - p - 3 > 0$.

The proofs and technical expressions for Theorem 2.6 can be found in Kollo and von Rosen (2005).

**Theorem 2.7.** *Let* $\boldsymbol{\Sigma}$ *and* $\boldsymbol{S}$ *be positive definite matrices of size* $p \times p$*. Then*

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

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

The proof of Theorem 2.7 can for example be found in Srivastava and Khatri (1979).

## 3. The likelihood approach

In classification problems, in some cases we consider that the populations 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).

Assume that a new observation $\boldsymbol{x}$ is to be classified into one of two known multivariate normal populations. We want to calculate the likelihood when $\boldsymbol{x}$ belongs either to population $\pi_1$ or $\pi_2$. Let $\boldsymbol{y} \in \pi_1$, $\boldsymbol{z} \in \pi_2$ and $\boldsymbol{x}$ be the new observation to be classified. If $\boldsymbol{x} \in \pi_i$, $i \in \{1, 2\}$, the likelihood is given by

$$\begin{aligned} L_i(\boldsymbol{y}, \boldsymbol{z} \,|\, \boldsymbol{x} \in \pi_i) = (2\pi)^{-\frac{3}{2}p} |\boldsymbol{\Sigma}|^{-\frac{3}{2}} \exp\bigg\{ -\frac{1}{2}\Big( (\boldsymbol{y} - \boldsymbol{\mu}_1)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}_1) \\ + (\boldsymbol{z} - \boldsymbol{\mu}_2)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_2) + (\boldsymbol{x} - \boldsymbol{\mu}_i)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) \Big) \bigg\}, \quad i \in \{1, 2\}. \end{aligned}$$

The ratio of the two likelihood functions equals

$$\frac{L_1(\boldsymbol{y}, \boldsymbol{z} \,|\, \boldsymbol{x} \in \pi_1)}{L_2(\boldsymbol{y}, \boldsymbol{z} \,|\, \boldsymbol{x} \in \pi_2)} = e^{-\frac{1}{2}\left[(\boldsymbol{x}-\boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right]}. \tag{3.1}$$

The observation $\boldsymbol{x}$ is classified into population $\pi_1$ if the ratio (3.1) is greater than or equal to 1 and otherwise into $\pi_2$. Taking the logarithm of the ratio yields

$$\begin{aligned} D = \ln\left(\frac{L_1(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x} \in \pi_1)}{L_2(\boldsymbol{y}, \boldsymbol{z}|\boldsymbol{x} \in \pi_2)}\right) &= -\frac{1}{2}\left[(\boldsymbol{x} - \boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) - (\boldsymbol{x} - \boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)\right] \\ &= (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)), \end{aligned} \tag{3.2}$$

which is a linear function in $\boldsymbol{x}$ and hence is normally distributed. The decision of classifying $\boldsymbol{x}$ depends on whether $D \geq 0$ or $D < 0$.

**Theorem 3.1.** *Consider the discriminant function* $D$ *in* (3.2)*. Then*

$$\begin{cases} D \sim \mathcal{N}\left(\frac{1}{2}\Delta^2, \Delta^2\right), & \textit{if} \quad \boldsymbol{x} \in \pi_1, \\ D \sim \mathcal{N}\left(-\frac{1}{2}\Delta^2, \Delta^2\right), & \textit{if} \quad \boldsymbol{x} \in \pi_2, \end{cases}$$

*where*

$$\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{\top}\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{3.3}$$

*is the Mahalanobis squared distance (*Anderson 2003*).*

### 3.1. Classification with known dispersion matrix

Suppose that there are $\boldsymbol{y}_i$, $i \in \{1, 2, \ldots, n_1\}$, observations selected from $\pi_1$ and $\boldsymbol{z}_j$, $j \in \{1, 2, \ldots, n_2\}$, observations from $\pi_2$. Let $\bar{\boldsymbol{y}} = \frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{y}_i$, and $\bar{\boldsymbol{z}} = \frac{1}{n_2}\sum_{j=1}^{n_2}\boldsymbol{z}_j$. In the next proposition, since the mean is supposed to be unknown, instead of (3.2), the following classification rule is presented.

**Proposition 3.2.** *Let $\boldsymbol{y}_i \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \Sigma)$, $i \in \{1, \ldots, n_1\}$, be a sample from $\pi_1$ collected in $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n_1})$ and $\boldsymbol{z}_j \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \Sigma)$, $j \in \{1, \ldots, n_2\}$ be a sample from $\pi_2$ collected in $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{n_2})$. Assume that an observation $\boldsymbol{x}$ is to be classified. Put*

$$\tilde{D} = \frac{1}{2}\frac{n_2}{n_2+1}(\bar{\boldsymbol{z}} - \boldsymbol{x})^{\top}\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{z}} - \boldsymbol{x}) - \frac{1}{2}\frac{n_1}{n_1+1}(\bar{\boldsymbol{y}} - \boldsymbol{x})^{\top}\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{x}). \tag{3.4}$$

*If $\tilde{D} \geq 0$ then $\boldsymbol{x}$ is classified to $\pi_1$ and to $\pi_2$ otherwise.*

In case of $n_1 = n_2$, definition of the classification function given in (3.4) is identical to Fisher's linear discriminant function, what motivates factor $\frac{1}{2}$ in the formula. The classification function $\tilde{D}$ has same distribution as the difference between two non-central $\chi^2$ distributions. The distributions for $\boldsymbol{\Sigma}^{-\frac{1}{2}}(\bar{\boldsymbol{z}} - \boldsymbol{x})$ and $\boldsymbol{\Sigma}^{-\frac{1}{2}}(\bar{\boldsymbol{y}} - \boldsymbol{x})$ do not depend on $\boldsymbol{\Sigma}$. Thus, the distribution for $\tilde{D}$ does not depend on the covariance, which is reasonable since we have assumed a common variance for both populations and therefore it is reasonable that $\boldsymbol{\Sigma}$ is not involved in classification of $\boldsymbol{x}$ to either $\pi_1$ or $\pi_2$.

### 3.2. Classification with unknown dispersion matrix

One approach which can be used in classification to handle unknown parameters is the maximum likelihood approach with the observation which is to be classified, $\boldsymbol{x}$, involved in the estimation, e.g., see Kudo (1959). Assume observations to be normally distributed. Now, let $\boldsymbol{y}_i$ be a sample from $\pi_1$ collected in $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{n_1})$, and $\boldsymbol{z}_j$ be a sample from $\pi_2$ collected in $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{n_2})$. Moreover $\boldsymbol{Y}$, $\boldsymbol{Z}$, and $\boldsymbol{x}$ are supposed to be jointly independent. If $\boldsymbol{x} \in \pi_1$ the following two models emerge in terms of $\boldsymbol{Y}$, $\boldsymbol{Z}$ and $\boldsymbol{x}$,

$$\begin{aligned} (\boldsymbol{Y} : \boldsymbol{x}) &= \boldsymbol{\mu}_1\mathbf{1}_{n_1+1}^{\top} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_{p,n_1+1}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_{n_1+1}), \\ \boldsymbol{Z} &= \boldsymbol{\mu}_2\mathbf{1}_{n_2}^{\top} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_{p,n_2}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_{n_2}), \end{aligned} \tag{3.5}$$

where $\mathbf{1}_n$ is a column vector of $n$ ones. If $\boldsymbol{x} \in \pi_2$ the models are specified as

$$\begin{aligned} \boldsymbol{Y} &= \tilde{\boldsymbol{\mu}}_1\mathbf{1}^{\top}{}_{n_1} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}_{p,n_1}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_{n_1}), \\ (\boldsymbol{Z} : \boldsymbol{x}) &= \tilde{\boldsymbol{\mu}}_2\mathbf{1}^{\top}{}_{n_2+1} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_{p,n_2+1}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_{n_2+1}), \end{aligned} \tag{3.6}$$

Equations (3.5) and (3.6) will be used to obtain the likelihood function when $\boldsymbol{x} \in \pi_1$ and $\boldsymbol{x} \in \pi_2$, respectively. Put

$$\begin{gathered}\boldsymbol{X}=(\boldsymbol{Y}:\boldsymbol{x}:\boldsymbol{Z}),\quad \boldsymbol{C}=\begin{pmatrix}\mathbf{1}_{n_1+1}^{\mathsf{T}} & \mathbf{0}_{n_2}^{\mathsf{T}}\\ \mathbf{0}_{n_1+1}^{\mathsf{T}} & \mathbf{1}_{n_2}^{\mathsf{T}}\end{pmatrix},\quad \tilde{\boldsymbol{C}}=\begin{pmatrix}\mathbf{1}_{n_1}^{\mathsf{T}} & \mathbf{0}_{n_2+1}^{\mathsf{T}}\\ \mathbf{0}_{n_1}^{\mathsf{T}} & \mathbf{1}_{n_2+1}^{\mathsf{T}}\end{pmatrix},\\ \boldsymbol{S}=\boldsymbol{X}(\mathbf{I}-\boldsymbol{C}^{\mathsf{T}}(\boldsymbol{C}\boldsymbol{C}^{\mathsf{T}})^{-1}\boldsymbol{C})\boldsymbol{X}^{\mathsf{T}},\quad \tilde{\boldsymbol{S}}=\boldsymbol{X}(\mathbf{I}-\tilde{\boldsymbol{C}}^{\mathsf{T}}(\tilde{\boldsymbol{C}}\tilde{\boldsymbol{C}}^{\mathsf{T}})^{-1}\tilde{\boldsymbol{C}})\boldsymbol{X}^{\mathsf{T}},\end{gathered}\tag{3.7}$$

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

**Lemma 3.3.** *Let $\hat{\boldsymbol{\Sigma}}_i$ be the estimated covariance matrix when $\boldsymbol{x}$ belongs to population $i$, $i\in\{1,2\}$ and $\boldsymbol{C}$, $\tilde{\boldsymbol{C}}$, $\boldsymbol{S}$ and $\tilde{\boldsymbol{S}}$ be defined in (3.7), and put $n=n_1+n_2$. Then, if $\boldsymbol{x}\in\pi_1$, the maximum likelihood estimators of the unknown parameters in (3.5) equal*

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

*If $\boldsymbol{x}\in\pi_2$, the maximum estimators of the unknown parameters in (3.6) equal*

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

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

$$L_{(\boldsymbol{Y}:\boldsymbol{x})}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma})=(2\pi)^{-(n_1+1)p/2}|\boldsymbol{\Sigma}|^{-(n_1+1)/2}e^{-\frac{1}{2}\sum_{i=1}^{n_1+1}(\boldsymbol{y}_i-\boldsymbol{\mu}_1)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_i-\boldsymbol{\mu}_1)},\tag{3.8}$$

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

$$L_{\boldsymbol{Z}}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma})=(2\pi)^{-n_2p/2}|\boldsymbol{\Sigma}|^{-n_2/2}e^{-\frac{1}{2}\sum_{j=1}^{n_2}(\boldsymbol{z}_j-\boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_j-\boldsymbol{\mu}_2)}.\tag{3.9}$$

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

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

Using MANOVA results and Theorem 2.7 yield the proof. □

Replacing the unknown parameters with their estimators in the likelihood functions, the likelihood ratio for classification of $\boldsymbol{x}$ into $\pi_1$ or $\pi_2$ is given by

$$\frac{(2\pi)^{-\frac{p}{2}(n+1)}|\hat{\boldsymbol{\Sigma}}_1|^{-\frac{1}{2}(n+1)}e^{-\frac{1}{2}\mathrm{tr}\{\hat{\boldsymbol{\Sigma}}_1^{-1}(\boldsymbol{X}-(\hat{\boldsymbol{\mu}}_1,\hat{\boldsymbol{\mu}}_2)\boldsymbol{C})(\boldsymbol{X}-(\hat{\boldsymbol{\mu}}_1,\hat{\boldsymbol{\mu}}_2)\boldsymbol{C})^{\mathsf{T}}\}}}{(2\pi)^{-\frac{p}{2}(n+1)}|\hat{\boldsymbol{\Sigma}}_2|^{-\frac{1}{2}(n+1)}e^{-\frac{1}{2}\mathrm{tr}\{\hat{\boldsymbol{\Sigma}}_2^{-1}(\boldsymbol{X}-(\hat{\hat{\boldsymbol{\mu}}}_1,\hat{\hat{\boldsymbol{\mu}}}_2)\boldsymbol{C})(\boldsymbol{X}-(\hat{\hat{\boldsymbol{\mu}}}_1,\hat{\hat{\boldsymbol{\mu}}}_2)\boldsymbol{C})^{\mathsf{T}}\}}}=\left(\frac{|\hat{\boldsymbol{\Sigma}}_2|}{|\hat{\boldsymbol{\Sigma}}_1|}\right)^{\frac{n+1}{2}},\tag{3.10}$$

where $\hat{\boldsymbol{\Sigma}}_i$ denotes the maximum likelihood estimator, given in Lemma 3.3, when $\boldsymbol{x}\in\pi_i$, $i=1$, 2. Moreover, $\hat{\boldsymbol{\mu}}_1$, $\hat{\boldsymbol{\mu}}_2$ are the maximum likelihood estimators of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ when $\boldsymbol{x}\in\pi_1$, and $\hat{\hat{\boldsymbol{\mu}}}_1$, $\hat{\hat{\boldsymbol{\mu}}}_2$ are respective estimators when $\boldsymbol{x}\in\pi_2$. The new observation $\boldsymbol{x}$ is classified into $\pi_1$ if the ratio (3.10) is larger or equal to 1 and into $\pi_2$ if the ratio (3.10) is smaller than 1. Furthermore, let

$$\boldsymbol{S}_y=\boldsymbol{Y}\boldsymbol{Y}^{\mathsf{T}}-n_1\bar{\boldsymbol{y}}\bar{\boldsymbol{y}}^{\mathsf{T}},\quad \boldsymbol{S}_z=\boldsymbol{Z}\boldsymbol{Z}^{\mathsf{T}}-n_2\bar{\boldsymbol{z}}\bar{\boldsymbol{z}}^{\mathsf{T}}$$

and put $\boldsymbol{S}_{yz}=\boldsymbol{S}_y+\boldsymbol{S}_z$, i.e., the sum of squares matrix based on the joint sample from populations $\pi_1$ and $\pi_2$, without taking into consideration the new observation. Moreover, $\boldsymbol{S}_{yz}\sim W_p(\boldsymbol{\Sigma},n_1+n_2-p-3)$. Consequently, $\boldsymbol{S}_y$ and $\boldsymbol{S}_z$ can be expressed as

$$S_y \quad = X(\mathbf{I} - C^{\mathsf{T}}(CC^{\mathsf{T}})^{-1})X^{\mathsf{T}} = S_{yz} + \frac{n_1}{n_1+1}(\bar{y} - x)(\bar{y} - x)^{\mathsf{T}},$$

and similarly,

$$S_z = S_{yz} + \frac{n_2}{n_2+1}(\bar{z} - x)(\bar{z} - x)^{\mathsf{T}}.$$

As a result, the likelihood ratio in (3.10) is equivalent to

$$\frac{|S_z|}{|S_y|} = \frac{|S_{yz} + \frac{n_2}{n_2+1}(\bar{z} - x)(\bar{z} - x)^{\mathsf{T}}|}{|S_{yz} + \frac{n_1}{n_1+1}(\bar{y} - x)(\bar{y} - x)^{\mathsf{T}}|} = \frac{1 + \frac{n_2}{n_2+1}(\bar{z} - x)^{\mathsf{T}} S_{yz}^{-1}(\bar{z} - x)}{1 + \frac{n_1}{n_1+1}(\bar{y} - x)^{\mathsf{T}} S_{yz}^{-1}(\bar{y} - x)}, \quad n_1 + n_2 - 3 \geq p. \tag{3.11}$$

Note that $S_{yz}^{-1}$ exits if $n_1 + n_2 - 3 \geq p$. A large likelihood suggests that $x$ is to be classified into $\pi_1$. Therefore, $x$ is classified into $\pi_1$ if

$$\frac{n_2}{n_2+1}(\bar{z} - x)^{\mathsf{T}} S_{yz}^{-1}(\bar{z} - x) \geq \frac{n_1}{n_1+1}(\bar{y} - x)^{\mathsf{T}} S_{yz}^{-1}(\bar{y} - x). \tag{3.12}$$

Note, since $S_{yz} \sim W_p(\Sigma, n_1 + n_2 - p - 3)$, Theorem 2.6 (i) implies $E[S_{yz}^{-1}] = m^{-1}\Sigma^{-1}$.

**Proposition 3.4.** *Consider the models given in (3.5) and (3.6), suppose $m = n_1 + n_2 - p - 3 > 0$, and put*

$$\hat{D} = \frac{n_2 m}{n_2 + 1}\frac{1}{2}(\bar{z} - x)^{\mathsf{T}} S_{yz}^{-1}(\bar{z} - x) - \frac{n_1 m}{n_1 + 1}\frac{1}{2}(\bar{y} - x)^{\mathsf{T}} S_{yz}^{-1}(\bar{y} - x). \tag{3.13}$$

*The observation $x$ is classified to $\pi_1$ if $\hat{D} \geq 0$ and to $\pi_2$ otherwise.*

Note that the classification rule (3.13) does not have to include $\frac{m}{2}$. However, the factor $\frac{1}{2}$ is used because in Proposition 3.2 it was also used and the constant $m$ is reasonable since $E[S_{yz}^{-1}] = m^{-1}\Sigma^{-1}$. The distribution function for $\hat{D}$ given in Proposition 3.4 is more complicated than the distribution of $\tilde{D}$ given in Proposition 3.2 since it is expressed as a difference of two non-central F-distributions. Its basic properties are given in the next section.

## 4. Moments of $\tilde{D}$ and $\hat{D}$

The two classification functions $\tilde{D}$ and $\hat{D}$ both depend on sample sizes and the unknown $\Delta^2$, given in (3.3), and their distributions are a difference between non-central $\chi^2$ distributions and a difference between two non-central F-distributions, respectively, which are complicated to derive. Therefore, we derive moments of the classification functions $\tilde{D}$ and $\hat{D}$. It follows that $\hat{D}$ and $\tilde{D}$ are asymptotically equivalently distributed since for any $\varepsilon$,

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

**Theorem 4.1.** *Let $\tilde{D}$ be defined in Proposition 3.2. Then*

$$\text{(i)}\quad \textit{if } \boldsymbol{x}\in\pi_1,\ \begin{cases} E[\tilde{D}] = \dfrac{1}{2}\dfrac{n_2}{n_2+1}\Delta^2, \\ Var[\tilde{D}] = p(1-q^2) + \dfrac{n_2}{n_2+1}\Delta^2; \end{cases}$$

$$\text{(ii)}\quad \textit{if } \boldsymbol{x}\in\pi_2,\ \begin{cases} E[\tilde{D}] = -\dfrac{1}{2}\dfrac{n_1}{n_1+1}\Delta^2, \\ Var[\tilde{D}] = p(1-q^2) + \dfrac{n_1}{n_1+1}\Delta^2, \end{cases}$$

*where* $\Delta^2 = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ *and*

$$q = \sqrt{\frac{n_1}{n_1+1}}\sqrt{\frac{n_2}{n_2+1}}. \tag{4.1}$$

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

$$\begin{aligned} \boldsymbol{u}_1 &= \sqrt{\frac{n_2}{n_2+1}}\frac{1}{\sqrt{2}}(\bar{\boldsymbol{z}}-\boldsymbol{x}) \sim \mathcal{N}_p\left(\sqrt{\frac{n_2}{n_2+1}}\frac{1}{\sqrt{2}}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_x), \frac{1}{2}\boldsymbol{\Sigma}\right) \\ \boldsymbol{u}_2 &= \sqrt{\frac{n_1}{n_1+1}}\frac{1}{\sqrt{2}}(\bar{\boldsymbol{y}}-\boldsymbol{x}) \sim \mathcal{N}_p\left(\sqrt{\frac{n_1}{n_1+1}}\frac{1}{\sqrt{2}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_x), \frac{1}{2}\boldsymbol{\Sigma}\right). \end{aligned} \tag{4.2}$$

Thus, $\boldsymbol{u} \sim \mathcal{N}_{2p}(\boldsymbol{\mu}, \boldsymbol{\psi})$, $\boldsymbol{\mu}$ is now defined as

$$\begin{aligned} \boldsymbol{\mu}^{\mathsf{T}} &= \left(\sqrt{\frac{n_2}{n_2+1}}\frac{1}{\sqrt{2}}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_x)^{\mathsf{T}}, \sqrt{\frac{n_1}{n_1+1}}\frac{1}{\sqrt{2}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_x)^{\mathsf{T}}\right), \\ \boldsymbol{\psi} &= \begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix} \otimes \frac{1}{2}\boldsymbol{\Sigma}, \end{aligned} \tag{4.3}$$

Let $\boldsymbol{P} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \otimes \boldsymbol{\Sigma}^{-1}$. Then, $\tilde{D}$ in Proposition 3.2 equals $\boldsymbol{u}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{u}$. Therefore, using Corollary 2.3 (i) it follows that

$$E[\tilde{D}] = \frac{1}{2}\frac{n_2}{n_2+1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_x)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_x) - \frac{1}{2}\frac{n_1}{n_1+1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_x)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_x). \tag{4.4}$$

If $\boldsymbol{x}\in\pi_1$, $\boldsymbol{\mu}_x = \boldsymbol{\mu}_1$. Thus, the proof follows. Furthermore, using Corollary 2.3 (ii) we get

$$Var[\tilde{D}] = 2\mathrm{tr}\{\boldsymbol{\psi}\boldsymbol{P}\boldsymbol{\psi}\boldsymbol{P}\} + 4\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{\psi}\boldsymbol{P}\boldsymbol{\psi}, \tag{4.5}$$

where we have

$$\boldsymbol{P}\boldsymbol{\psi}\boldsymbol{P} = \begin{pmatrix} 1 & -q \\ -q & 1 \end{pmatrix} \otimes \frac{1}{2}\boldsymbol{\Sigma}^{-1}. \tag{4.6}$$

Therefore,

$$\boldsymbol{\psi}\boldsymbol{P}\boldsymbol{\psi}\boldsymbol{P} = \begin{pmatrix} 1-q^2 & 0 \\ 0 & 1-q^2 \end{pmatrix} \otimes \frac{1}{4}\mathbf{I}_p. \tag{4.7}$$

Hence, the terms in (4.5) are given by

$$\operatorname{tr}\{\boldsymbol{\psi}\boldsymbol{P}\boldsymbol{\psi}\boldsymbol{P}\} = \frac{1}{2}p(1-q^2) \tag{4.8}$$

and

$$\begin{aligned}\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{\psi}\boldsymbol{P}\boldsymbol{\psi} = \frac{1}{4}\bigg[&\frac{n_2}{n_2+1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_x)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_x) - q^2(\boldsymbol{\mu}_1-\boldsymbol{\mu}_x)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_x)\\ &- q^2(\boldsymbol{\mu}_2-\boldsymbol{\mu}_x)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_x) + \frac{n_1}{n_1+1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_x)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_x)\bigg],\end{aligned} \tag{4.9}$$

where $\boldsymbol{\mu}_x$ equals either $\boldsymbol{\mu}_1$ or $\boldsymbol{\mu}_2$. Moreover, the two middle terms will always be zero. If $\boldsymbol{x} \in \pi_1$, replacing (4.8) and (4.9) in (4.5) and considering $\Delta^2$ defined by (3.3) yields

$$Var[\tilde{D}] = p(1-q^2) + \frac{n_2}{n_2+1}\Delta^2. \tag{4.10}$$

The proof of (ii) can be obtained using similar calculations. □

**Theorem 4.2.** *Consider the discriminant function $\hat{D}$ in Proposition 3.4. The expected value of the discriminant function is given by*

$$E[\hat{D}] = \begin{cases} \dfrac{1}{2}\dfrac{n_2}{n_2+1}\Delta^2, & \textit{if} \quad \boldsymbol{x} \in \pi_1,\\ -\dfrac{1}{2}\dfrac{n_1}{n_1+1}\Delta^2, & \textit{if} \quad \boldsymbol{x} \in \pi_2, \end{cases} \tag{4.11}$$

*where $\Delta^2$ is the Mahalanobis squared distance given in (3.3).*

*Proof.* Let $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ be defined in (4.2) and hence, $\boldsymbol{u} = (\boldsymbol{u}_1, \boldsymbol{u}_2) \sim \mathcal{N}_{2p}(\boldsymbol{\mu}, \boldsymbol{\psi})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\psi}$ are defined in (4.3) and $q$ in (4.1). Let $\boldsymbol{Q} = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \otimes m\boldsymbol{S}_{yz}^{-1}$. Hence, $\hat{D} = \boldsymbol{u}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{u}$. The vectors $\boldsymbol{u}$ and $\boldsymbol{Q}$ are independently distributed. Then, since $\boldsymbol{u} \sim \mathcal{N}_{2p}(\boldsymbol{\mu}, \boldsymbol{\psi})$, using Corollary 2.3 (i) and the fact that $\operatorname{tr}(\boldsymbol{\psi}\boldsymbol{Q}) = \operatorname{tr}\left\{\begin{pmatrix} 1 & -q \\ q & -1 \end{pmatrix} \otimes \frac{1}{2}m\boldsymbol{\Sigma}\boldsymbol{S}_{yz}^{-1}\right\} = 0$, we get

$$E\left[\hat{D}|\boldsymbol{S}_{yz}\right] = E\left[\boldsymbol{u}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{u}|\boldsymbol{S}_{yz}\right] = \underbrace{\operatorname{tr}(\boldsymbol{\psi}\boldsymbol{Q})}_{=0} + \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{\mu} = \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{\mu},$$

where $E[\cdot|\cdot]$ denotes conditional expectation. As a result,

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

As depending on population which $\boldsymbol{x}$ belongs to, $\boldsymbol{\mu}_x$ is equal $\boldsymbol{\mu}_1$ or $\boldsymbol{\mu}_2$, both middle terms are equal to 0. Thus, if $\boldsymbol{x} \in \pi_1$, then $\boldsymbol{\mu}_x = \boldsymbol{\mu}_1$. Therefore, since from Theorem 2.6 (i), $E[m\boldsymbol{S}_{yz}^{-1}] = \boldsymbol{\Sigma}^{-1}$, we get

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

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

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

□

**Theorem 4.3.** *Consider the discriminant function $\hat{D}$ in Proposition 3.4 and let $c_0 = \frac{m+p}{m(m-2)(m+1)}$. For $m = n_1 + n_2 - p - 3 > 2$, the variance of the discriminant function is given by*

$$Var[\hat{D}] = \begin{cases} (1-q^2)m^2c_0p + m^2\dfrac{n_2}{n_2+1}c_0\Delta^2 + \left(\dfrac{n_2}{n_2+1}\right)^2\dfrac{1}{2(m-2)}(\Delta^2)^2, & \text{if} \quad \boldsymbol{x} \in \pi_1, \\ (1-q^2)m^2c_0p + m^2\dfrac{n_1}{n_1+1}c_0\Delta^2 + \left(\dfrac{n_1}{n_1+1}\right)^2\dfrac{1}{2(m-2)}(\Delta^2)^2, & \text{if} \quad \boldsymbol{x} \in \pi_2, \end{cases} \tag{4.12}$$

*where $\Delta^2$ is the Mahalanobis squared distance given by (3.3) and $q$ is given by (4.1).*

*Proof.* It will be utilized that

$$Var[\hat{D}] = Var\left[E\left[\hat{D}|\mathbf{S}_{yz}\right]\right] + E\left[Var\left[\hat{D}|\mathbf{S}_{yz}\right]\right], \tag{4.13}$$

where $Var[\cdot|\cdot]$ denotes conditional variance. Moreover, $\hat{D} = \boldsymbol{u}^\top\mathbf{Q}\boldsymbol{u}$, $\boldsymbol{u} \sim \mathcal{N}_{2p}(\boldsymbol{\mu}, \boldsymbol{\psi})$, where $\boldsymbol{u}$, $\mathbf{Q}$ and $\boldsymbol{\psi}$ are defined in the proof of Theorem 4.2. Using Theorem 2.2 (ii) we get,

$$Var\left[\hat{D}|\mathbf{S}_{yz}\right] = 2\text{tr}\{\boldsymbol{\psi}\mathbf{Q}\boldsymbol{\psi}\mathbf{Q}\} + 4\boldsymbol{\mu}^\top\mathbf{Q}\boldsymbol{\psi}\mathbf{Q}\boldsymbol{\psi}. \tag{4.14}$$

Evaluating (4.14) step by step:

$$\begin{aligned} \mathbf{Q}\boldsymbol{\psi}\mathbf{Q} &= \left[\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \otimes m\mathbf{S}_{yz}^{-1}\right]\left[\begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix} \otimes \frac{1}{2}\boldsymbol{\Sigma}\right]\left[\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \otimes m\mathbf{S}_{yz}^{-1}\right] \\ &= \begin{pmatrix} 1 & -q \\ -q & 1 \end{pmatrix} \otimes \frac{1}{2}m^2\mathbf{S}_{yz}^{-1}\boldsymbol{\Sigma}\mathbf{S}_{yz}^{-1}, \end{aligned} \tag{4.15}$$

$$\begin{aligned} \boldsymbol{\psi}\mathbf{Q}\boldsymbol{\psi}\mathbf{Q} &= \left[\begin{pmatrix} 1 & q \\ q & 1 \end{pmatrix} \otimes \frac{1}{2}\boldsymbol{\Sigma}\right]\begin{pmatrix} 1 & -q \\ -q & 1 \end{pmatrix} \otimes \frac{1}{2}m^2\mathbf{S}_{yz}^{-1}\boldsymbol{\Sigma}\mathbf{S}_{yz}^{-1} \\ &= \begin{pmatrix} 1-q^2 & 0 \\ 0 & 1-q^2 \end{pmatrix} \otimes \frac{1}{4}m^2\boldsymbol{\Sigma}\mathbf{S}_{yz}^{-1}\boldsymbol{\Sigma}\mathbf{S}_{yz}^{-1}. \end{aligned} \tag{4.16}$$

Then the first term on the right side in (4.14) becomes

$$\mathrm{tr}\{\boldsymbol{\psi}\boldsymbol{Q}\boldsymbol{\psi}\boldsymbol{Q}\} = \frac{1}{2}(1-q^2)m^2\mathrm{tr}\{\boldsymbol{\Sigma}\boldsymbol{S}_{yz}^{-1}\boldsymbol{\Sigma}\boldsymbol{S}_{yz}^{-1}\}.$$

Furthermore, for $q = \sqrt{\frac{n_1}{n_1+1}}\sqrt{\frac{n_2}{n_2+1}}$, the second term of (4.14) equals

$$\begin{aligned}\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{\psi}\boldsymbol{Q}\boldsymbol{\psi} &= \frac{1}{4}m^2\frac{n_2}{n_2+1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{S}_{yz}^{-1}\boldsymbol{\Sigma}\boldsymbol{S}_{yz}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_{\boldsymbol{x}}) - \underbrace{\frac{1}{2}q^2m^2(\boldsymbol{\mu}_2-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{S}_{yz}^{-1}\boldsymbol{\Sigma}\boldsymbol{S}_{yz}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_{\boldsymbol{x}})}_{=0} \\ &+ \frac{1}{4}m^2\frac{n_1}{n_1+1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{S}_{yz}^{-1}\boldsymbol{\Sigma}\boldsymbol{S}_{yz}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_{\boldsymbol{x}}).\end{aligned} \tag{4.17}$$

The term $\frac{1}{2}q^2m^2(\boldsymbol{\mu}_2-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{S}_{yz}^{-1}\boldsymbol{\Sigma}\boldsymbol{S}_{yz}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_{\boldsymbol{x}})$ is always zero because either $\boldsymbol{x}\in\pi_1$ or $\boldsymbol{x}\in\pi_2$.

Moreover, for $m>2$, Theorem 2.6 (ii) yields $E[\boldsymbol{S}_{yz}^{-1}\boldsymbol{\Sigma}\boldsymbol{S}_{yz}^{-1}] = c_0\boldsymbol{\Sigma}^{-1}$. Then, we can evaluate

$$\begin{aligned}E\left[Var\left[\hat{D}|\boldsymbol{S}_{yz}\right]\right] &= (1-q^2)m^2\mathrm{tr}\left\{E\left[\boldsymbol{\Sigma}\boldsymbol{S}_{yz}^{-1}\boldsymbol{\Sigma}\boldsymbol{S}_{yz}^{-1}\right]\right\} + m^2\frac{n_2}{n_2+1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathsf{T}}E\left[\boldsymbol{S}_{yz}^{-1}\boldsymbol{\Sigma}\boldsymbol{S}_{yz}^{-1}\right](\boldsymbol{\mu}_2-\boldsymbol{\mu}_{\boldsymbol{x}}) \\ &+ m^2\frac{n_1}{n_1+1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathsf{T}}E\left[\boldsymbol{S}_{yz}^{-1}\boldsymbol{\Sigma}\boldsymbol{S}_{yz}^{-1}\right](\boldsymbol{\mu}_1-\boldsymbol{\mu}_{\boldsymbol{x}}) \\ &= (1-q^2)m^2c_0p + m^2\frac{n_2}{n_2+1}c_0(\boldsymbol{\mu}_2-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_{\boldsymbol{x}}) \\ &+ m^2\frac{n_1}{n_1+1}c_0(\boldsymbol{\mu}_1-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_{\boldsymbol{x}}).\end{aligned}$$

Hence, if $\boldsymbol{x}\in\pi_1$ then

$$E\left[Var\left[\hat{D}|\boldsymbol{S}_{yz}\right]\right] = (1-q^2)m^2c_0p + m^2\frac{n_2}{n_2+1}c_0\Delta^2.$$

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

$$E\left[\hat{D}|\boldsymbol{S}_{yz}\right] = \frac{1}{2}\frac{n_2}{n_2+1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathsf{T}}m\boldsymbol{S}_{yz}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_{\boldsymbol{x}}) - \frac{1}{2}\frac{n_1}{n_1+1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathsf{T}}m\boldsymbol{S}_{yz}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_{\boldsymbol{x}}).$$

Put $\boldsymbol{M} = \frac{1}{2}\frac{n_2}{n_2+1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{\mu}_2-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathsf{T}} - \frac{1}{2}\frac{n_1}{n_1+1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{\mu}_1-\boldsymbol{\mu}_{\boldsymbol{x}})^{\mathsf{T}}$. Then, $E[\hat{D}|\boldsymbol{S}_{yz}] = \mathrm{tr}\{m\ \boldsymbol{S}_{yz}^{-1}\boldsymbol{M}\}$.

Let $m = n_1+n_2-p-3$ and put

$$c_1 = \frac{1}{m}, \quad c_2 = \frac{m-1}{m(m+1)(m-2)} \quad \textit{and} \quad c_3 = \frac{1}{m-1}c_2. \tag{4.18}$$

Hence, using Theorem 2.6 (vii), with $c_1$, $c_2$ and $c_3$ defined as in (4.18) we get

$$\begin{aligned}Var\left[E\left[\hat{D}|\boldsymbol{S}_{yz}\right]\right] &= Var\left[\mathrm{tr}\{m\boldsymbol{S}_{yz}^{-1}\boldsymbol{M}\}\right] = m^2\mathrm{vec}^{\mathsf{T}}\boldsymbol{M}Var\left[\boldsymbol{S}_{yz}^{-1}\right]\mathrm{vec}\boldsymbol{M} \\ &= 2c_3m^2\mathrm{vec}^{\mathsf{T}}\boldsymbol{M}(\boldsymbol{\Sigma}^{-1}\otimes\boldsymbol{\Sigma}^{-1})\mathrm{vec}\boldsymbol{M} + (c_2-c_1^2)(\mathrm{tr}\{\boldsymbol{\Sigma}^{-1}\boldsymbol{M}\})^2.\end{aligned}$$

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

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

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

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

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

Though they are are asymptotically equivalent, the distribution of $\tilde{D}$ defined in Proposition 3.2 is comparably simpler than the distribution of $\hat{D}$ in Proposition 3.4. In addition, the expectations of $\hat{D}$ and $\tilde{D}$ are the same.

## 5. Concluding remarks

In this paper, two asymptotically equivalent classification rules were considered using a likelihood approach, for a known and unknown covariance matrix, under the assumption about multivariate normality of the populations. We regard an observation $\boldsymbol{x}$ as coming from $\pi_1$ or $\pi_2$ populations according to whether the observed value of the discriminant functions $\tilde{D}$ or $\hat{D}$ is positive or negative. Note that when $n_1 = n_2 \equiv n$, $\hat{D} = -\frac{mn}{n+1}W$, where $m = 2n - p - 3$. The discriminant functions $\tilde{D}$ and $\hat{D}$ have smaller variances than the $W$-rule, (Wald 1944; Anderson 1951)

The presence of an inverted Wishart distributed covariance matrix in the estimated discriminant function, $\hat{D}$ makes it difficult to derive moments. The results of this paper can be utilized in for example Edgeworth expansion which gives alternative approximations of the distribution of the misclassification errors.

## ORCID

Emelyne Umunoza Gasana http://orcid.org/0000-0002-5559-4120
Dietrich von Rosen http://orcid.org/0000-0002-3135-4325
Martin Singull http://orcid.org/0000-0001-9896-4438

## References

Anderson, T. W. 1951. Classification by multivariate analysis. *Psychometrika* 16 (1):31–50. doi:10.1007/BF02313425.

Anderson, T. W. 2003. *An introduction to multivariate statistical analysis*. 3rd ed. Hoboken: John Wiley & Sons, Inc.

Barnard, M. M. 1935. The secular variations of skull characters in four series of egyptian skulls. *Annals of Eugenics* 6 (4):352–71. doi:10.1111/j.1469-1809.1935.tb02117.x.

Bowker, A. H., and R. Sitgreaves. 1961. An asymptotic expansion for the distribution of the $W$-classification statistic. Chapter 19 In *Studies in item analysis and prediction*, ed. H. Solomon. California: Stanford University Press.

Critchley, F., and I. Ford. 1984. On the covariance of two noncentral F random variables and the variance of the estimated liner discriminant function. *Biometrika* 71:637–8.

Davis, A. 1987. Moments of linear discriminant functions and an asymptotic confidence interval for the log odds ratio. *Biometrika* 74 (4):829–40. doi:10.1093/biomet/74.4.829.

Day, N. E., and D. F. Kerridge. 1967. A general maximum likelihood discriminant. *Biometrics* 23 (2):313–23. doi:10.2307/2528164.

Fisher, R. A. 1936. The use of multiple measurements in 10 taxonomic problems. *Annals of Eugenics* 7 (2):179–88. doi:10.1111/j.1469-1809.1936.tb02137.x.

Fisher, R. A. 1938. The statistical utilization of multiple measurements. *Annals of Eugenics* 8 (4): 376–86. doi:10.1111/j.1469-1809.1938.tb02189.x.

Fujikoshi, Y., V. V. Ulyanov, and R. Shimizu. 2011. *Multivariate statistics: High-dimensional and large-sample approximations*, vol. 760. Hoboken: John Wiley & Sons, Inc.

Kollo, T., and D. von Rosen. 2005. *Advanced multivariate statistics with matrices*, vol. 579. Dordrecht, The Netherlands: Springer Science & Business Media.

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

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

Mahalanobis, P. C. 1925. Analysis of race-mixture in Bengal. *Journal and Proceedings of Asiatic Society of Bengal New Series* 23:301–33.

Mahalanobis, P. C. 1930. On test and measures of group divergence: Theoretical formulae. *Journal and Proceedings of Asiatic Society of Bengal New Series* 26:541–88.

Mathai, A, and S. B. Provost. 1992. *Quadratic forms in random variables: Theory and applications*, vol. 87. New York: Marcel Dekker Inc.

McLachlan, G. J. 1992. *Discriminant analysis and statistical pattern recognition*. New York: John Wiley & Sons.

Okamoto, M. 1963. An asymptotic expansion for the distribution of the linear discriminant function. *The Annals of Mathematical Statistics* 34 (4):1286–301. doi:10.1214/aoms/1177703864.

Pearson, K. 1915. On the problem of sexing osteometric material. *Biometrika* 10 (4):479–87. doi: 10.1093/biomet/10.4.479.

Pearson, K. 1926. On the coefficient of racial likeness. *Biometrika* 18 (1–2):105–17. doi:10.1093/biomet/18.1-2.105.

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):159–93.

Rao, C. R. 1966. Discriminant function between composite hypotheses and related problems. *Biometrika* 53 (3–4):339–45. doi:10.1093/biomet/53.3-4.339.

Schaafsma, W. 1982. Selecting variables in discriminant analysis for improving upon classical procedures. *Handbook of Statistics* 2:857–81.

Sitgreaves, R. 1961. Some results on the distribution of the *W*-classification statistic. *Chapter 15.* In *Studies in item analysis and prediction*, ed. H. Solomon, 241–61. California: Stanford University Press.

Srivastava, M. S., and C. Khatri. 1979. *An introduction to multivariate statistics*. New York, North-Holland.

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):145–62. doi:10.1214/aoms/1177731280.