# Discriminant analysis in small and large dimensions

Bodnar, T.; Mazur, S.; Ngailo, E.; Parolya, N.

**DOI**
10.1090/tpms/1096

**Publication date**
2019

**Document Version**
Accepted author manuscript

**Published in**
Theory of Probability and Mathematical Statistics

**Citation (APA)**
Bodnar, T., Mazur, S., Ngailo, E., & Parolya, N. (2019). Discriminant analysis in small and large dimensions. *Theory of Probability and Mathematical Statistics*, *100*, 21-41. Advance online publication. https://doi.org/10.1090/tpms/1096

**Important note**
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Теорія ймовірностей
та математична статистика
Вип. 1(100)/2019, с. 24–42

Teoriya Ĭmovirnosteĭ
ta Matematychna Statystyka
No. 1(100)/2019, pp. 24–42

UDC 519.21

# DISCRIMINANT ANALYSIS IN SMALL AND LARGE DIMENSIONS

T. BODNAR, S. MAZUR, E. NGAILO, N. PAROLYA

Abstract. We study the distributional properties of the linear discriminant function under the assumption of normality by comparing two groups with the same covariance matrix but different mean vectors. A stochastic representation for the discriminant function coefficients is derived, which is then used to obtain their asymptotic distribution under the high-dimensional asymptotic regime. We investigate the performance of the classification analysis based on the discriminant function in both small and large dimensions. A stochastic representation is established, which allows to compute the error rate in an efficient way. We further compare the calculated error rate with the optimal one obtained under the assumption that the covariance matrix and the two mean vectors are known. Finally, we present an analytical expression of the error rate calculated in the high-dimensional asymptotic regime. The finite-sample properties of the derived theoretical results are assessed via an extensive Monte Carlo study.

*Key words and phrases.* Discriminant function, stochastic representation, large-dimensional asymptotics, random matrix theory, classification analysis.

2010 *Mathematics Subject Classification.* 62H10, 62E15, 62E20, 60F05, 60B20.

## 1. Introduction

In the modern world of science and technology, high-dimensional data are present in various fields such as finance, environment science and social sciences. In the sense of many complex multivariate dependencies observed in data, formulating correct models and developing inferential procedures are the major challenges. It is usually assumed that the sample size is considerably larger than the process dimension in the traditional multivariate statistical theory. However, this assumption is not longer valid under the high-dimensional setting, where the dimension is comparable to the sample size.

The covariance matrix is one of the most popular approaches to capture the dependence among variables. Although its application is restricted only to linear dependence and more sophisticated methods, like copula, should be applied in the general case. Recently, a number of papers have been published, which deal with estimating the covariance matrix (see, e. g., [1, 8, 9, 16, 18, 22, 23, 32]) and testing its structure (see, e. g., [2, 7, 19, 20, 28, 29, 31]) in large dimension.

In many applications, the covariance matrix and mean vector are utilized together. For example, the product of the inverse sample covariance matrix and the difference of the sample mean vectors is present in the discriminant function, where a linear combination of variables (discriminant function coefficients) is determined such that the standardized distance between the groups of observations is maximized. A second example arises in portfolio theory, where the vector of optimal portfolio weights is proportional to the products of inverse sample covariance matrix and sample mean vector [13].

The discriminant analysis is a multivariate technique concerned with separating distinct sets of objects (or observations) [30]. Its two main tasks are to distinguish distinct sets of observations and to allocate new observations to previously defined groups [37]. The main methods of the discriminant analysis are the linear discriminant and quadratic discriminant functions. The linear discriminant function is a generalization of Fisher linear discriminant analysis, which is used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or

more groups of objects in the best way. The application of the linear discriminant function is restricted to the assumption of the equal covariance matrices of the populations. Although the quadratic discriminant function can be used when the latter assumption is violated, its application is more computational exhaustive which needs to estimate the covariance matrices of each group, and requires more observations than in the case of linear discriminant function [35]. Moreover, the decision boundary is easy to understand and to visualize in high-dimensional settings, if the linear discriminant function is used.

The discriminant analysis is a well established topic in multivariate statistics. Many asymptotic results are available when the sample sizes of groups to be separated are assumed to be large, while the number of variables is fixed and significantly smaller than the sample size (see, e. g., [34, 37]). However, these results cannot automatically be transferred when the number of variables is higher than the sample size, the situation which is known in the statistical literature as the high-dimensional asymptotic regime. It is remarkable that in this case the results obtained under the standard asymptotic regime can deviate significantly from those obtained under the high-dimensional asymptotics (see, e. g., [3]). Fujikoshi [24] provided an asymptotic approximation of the linear discriminant function in high dimension by considering the case of equal sample sizes and compared the results with the classical asymptotic approximation by [41]. For the samples of non-equal sizes, they pointed out that the high-dimensional approximation is extremely accurate. However, [40] showed that the Fisher linear discriminant function performs poorly due to diverging spectra in the case of large-dimensional data and small sample sizes. The papers [6, 39] investigated the asymptotic properties of the linear discriminant function in high dimension, while modifications of the linear discriminant function can be found in [17, 38]. The asymptotic results for the discriminant function coefficients in matrix-variate skew models can be found in [11].

We contribute to the statistical literature by deriving a stochastic representation of the discriminant function coefficient and the classification rule based on the linear discriminant function. These results provide us an efficient way of simulating these random quantities and they are also used in the derivation of their high-dimensional asymptotic distributions, using which the error rate of the classification rule based on the linear discriminant function can be easily assessed and the problem of the increasing dimensionality can be visualized in a simple way. An important challenge, which is not discussed in this paper, is the extension of the derived theoretical results to the case of the quadratic discriminant function, i. e. when two populations have different covariance matrices. These results will require the development of new stochastic representations and are left for future research.

The rest of the paper is organized as follows. The finite-sample properties of the discriminant function are presented in Subsection 2.1, where we derive a stochastic representation for the discriminant function coefficients. In Subsection 2.2, an exact one-sided test for the comparison of the population discriminant function coefficients is suggested, while a stochastic representation for the classification rule is obtained in Subsection 2.3. The asymptotic distributions of the discriminant function coefficients and of the classification rule are derived in Section 3, while finite sample performance of the asymptotic distribution is analysed in Subsection 3.2.

## 2. Finite-sample properties of the discriminant function

Let $\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{n_1}^{(1)}$ and $\mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_{n_2}^{(2)}$ be two independent samples from the multivariate normal distributions which consist of independent and identically distributed random vectors with $\mathbf{x}_i^{(1)} \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ for $i = 1, \ldots, n_1$ and $\mathbf{x}_j^{(2)} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ for $j = 1, \ldots, n_2$ where $\boldsymbol{\Sigma}$ is positive definite. Throughout the paper, $\mathbf{1}_n$ denotes the $n$-dimensional vector

of ones, $\mathbf{I}_n$ is the $n \times n$ identity matrix, and the symbol $\otimes$ stands for the Kronecker product.

Let $\mathbf{X}^{(1)} = \left(\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{n_1}^{(1)}\right)$ and $\mathbf{X}^{(2)} = \left(\mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_{n_2}^{(2)}\right)$ be observation matrices. Then the sample estimators for the mean vectors and the covariance matrices constructed from each sample are given by

$$\bar{\mathbf{x}}^{(j)} = \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{x}_i^{(j)} = \frac{1}{n_j}\mathbf{X}^{(j)}\mathbf{1}_{n_j},$$

$$\mathbf{S}^{(j)} = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}\left(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}^{(j)}\right)\left(\mathbf{x}_i^{(j)} - \bar{\mathbf{x}}^{(j)}\right)^{\top}.$$

The pooled estimator for the covariance matrix, i. e., an estimator for $\boldsymbol{\Sigma}$ obtained from two samples, is then given by

$$\mathbf{S}_{pl} = \frac{1}{n_1 + n_2 - 2}\Big[(n_1 - 1)\mathbf{S}^{(1)} + (n_2 - 1)\mathbf{S}^{(2)}\Big]. \tag{1}$$

The following lemma (see, e. g., [37, Section 5.4.2]) presents the joint distribution of $\bar{\mathbf{x}}^{(1)}$, $\bar{\mathbf{x}}^{(2)}$ and $\mathbf{S}_{pl}$.

**Lemma 1.** *Let* $\mathbf{X}_1 \sim \mathcal{N}_{p,n_1}\left(\boldsymbol{\mu}_1\mathbf{1}_{n_1}^{\top}, \boldsymbol{\Sigma} \otimes \mathbf{I}_{n_1}\right)$ *and* $\mathbf{X}_2 \sim \mathcal{N}_{p,n_2}\left(\boldsymbol{\mu}_2\mathbf{1}_{n_2}^{\top}, \boldsymbol{\Sigma} \otimes \mathbf{I}_{n_2}\right)$ *for* $p < n_1 + n_2 - 2$. *Assume that* $\mathbf{X}_1$ *and* $\mathbf{X}_2$ *are independent. Then*

*(a)* $\bar{\mathbf{x}}^{(1)} \sim \mathcal{N}_p\left(\boldsymbol{\mu}_1, \frac{1}{n_1}\boldsymbol{\Sigma}\right)$;

*(b)* $\bar{\mathbf{x}}^{(2)} \sim \mathcal{N}_p\left(\boldsymbol{\mu}_2, \frac{1}{n_2}\boldsymbol{\Sigma}\right)$;

*(c)* $(n_1 + n_2 - 2)\mathbf{S}_{pl} \sim \mathcal{W}_p(n_1 + n_2 - 2, \boldsymbol{\Sigma})$.

*Moreover,* $\bar{\mathbf{x}}^{(1)}$, $\bar{\mathbf{x}}^{(2)}$ *and* $\mathbf{S}_{pl}$ *are mutually independently distributed.*

The results of Lemma 1, in particular, imply that

$$\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)} \sim \mathcal{N}_p\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{\Sigma}\right), \tag{2}$$

which is independent of $\mathbf{S}_{pl}$.

2.1. **Stochastic representation for the discriminant function coefficients.** The discriminant function coefficients are given by the following vector

$$\hat{\mathbf{a}} = \mathbf{S}_{pl}^{-1}\left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right), \tag{3}$$

which is the sample-based feasible estimator of the population discriminant function coefficient vector expressed as

$$\mathbf{a} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

We consider a more general problem by deriving the distribution of linear combinations of the discriminant function coefficients. This result possesses several practical application: (i) it allows a direct comparison of the population coefficients in the discriminant function by deriving a corresponding statistical test; (ii) it can be used in the classification problem, where, providing a new observation vector, one has to decide to which of two groups the observation vector has to be classified.

Let $\mathbf{L}$ be a $k \times p$ matrix of constants such that $\operatorname{rank}(\mathbf{L}) = k < p$. We are then interested in

$$\hat{\boldsymbol{\theta}} = \mathbf{L}\hat{\mathbf{a}} = \mathbf{L}\mathbf{S}_{pl}^{-1}\left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right). \tag{4}$$

Choosing different matrices $\mathbf{L}$, we are able to provide different inferences about the linear combinations of the discriminant function coefficients. For instance, if $k = 1$ and $\mathbf{L}$ is the

vector with all elements zero except the one on the $j$th position which is one, then we get the distribution of the $j$th coefficient in the discriminant function. If we choose $k = 1$ and $\mathbf{L} = (1, -1, 0, \ldots, 0)^\top$, then we analyse the difference between the first two coefficients in the discriminant function. The corresponding result can be further used to test if the population counterparts to these coefficients are zero or not. For $k > 1$ several linear combinations of the discriminant function coefficients are considered simultaneously.

In the next theorem we derive a stochastic representation for $\hat{\boldsymbol{\theta}}$. The stochastic representation is a very important tool in analysing the distributional properties of random quantities. It is widely spread in the computation statistics (e. g., [25]), in the theory of elliptical distributions [27] as well as in Bayesian statistics (cf. [4, 5, 10]). Later on, we use the symbol $\stackrel{d}{=}$ to denote the equality in distribution.

**Theorem 1.** *Let* $\mathbf{L}$ *be an arbitrary* $k \times p$ *matrix of constants such that* $\operatorname{rank}(\mathbf{L}) = k < p$. *Then, under the assumption of Lemma 1, the stochastic representation of* $\hat{\boldsymbol{\theta}} = \mathbf{L}\hat{\mathbf{a}}$ *is given by*

$$\hat{\boldsymbol{\theta}} \stackrel{d}{=} (n_1 + n_2 - 2)\xi^{-1}\left(\mathbf{L}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}} + \sqrt{\frac{\breve{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}}{n_1 + n_2 - p}}\left(\mathbf{L}\mathbf{R}_{\breve{\mathbf{x}}}\mathbf{L}^\top\right)^{1/2}\mathbf{t}_0\right),$$

*where* $\mathbf{R}_{\breve{\mathbf{x}}} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}\breve{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}/\breve{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}$; $\xi \sim \chi^2_{n_1+n_2-p-1}$, $\breve{\mathbf{x}} \sim \mathcal{N}_p\Big(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\boldsymbol{\Sigma}\Big)$, *and* $\mathbf{t}_0 \sim t_k(n_1 + n_2 - p, \mathbf{0}_k, \mathbf{I}_k)$. *Moreover,* $\xi$, $\breve{\mathbf{x}}$ *and* $\mathbf{t}_0$ *are mutually independent.*

*Proof.* From Lemma 1 (c) and Theorem 3.4.1 of [26], we obtain that

$$\frac{1}{n_1 + n_2 - 2}\mathbf{S}_{pl}^{-1} \sim \mathcal{IW}_p(n_1 + n_2 + p - 1, \boldsymbol{\Sigma}^{-1}).$$

Also, since $\breve{\mathbf{x}} = \bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}$ and $\mathbf{S}_{pl}$ are independent, the conditional distribution of $\hat{\boldsymbol{\theta}} = \mathbf{L}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}$ given $\breve{\mathbf{x}} = \breve{\mathbf{x}}^*$ equals to the distribution of $\boldsymbol{\theta}^* = \mathbf{L}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*$ and it can be rewritten in the following form

$$\boldsymbol{\theta}^* \stackrel{d}{=} (n_1 + n_2 - 2)\breve{\mathbf{x}}^{*\top}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}^*\frac{\mathbf{L}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*}{\breve{\mathbf{x}}^{*\top}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*}\frac{\breve{\mathbf{x}}^{*\top}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*}{(n_1 + n_2 - 2)\breve{\mathbf{x}}^{*\top}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}^*}.$$

Applying Theorem 3.2.12 of [34] we obtain that

$$\xi^* = (n_1 + n_2 - 2)\frac{\breve{\mathbf{x}}^{*\top}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}^*}{\breve{\mathbf{x}}^{*\top}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*} \sim \chi^2_{n_1+n_2-p-1}$$

and its distribution is independent of $\breve{\mathbf{x}}^*$. Hence,

$$\xi = (n_1 + n_2 - 2)\frac{\breve{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}}{\breve{\mathbf{x}}^\top\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}} \sim \chi^2_{n_1+n_2-p-1}$$

and $\xi$, $\breve{\mathbf{x}}$ are independent.

Using Theorem 3 of [12], we obtain that $\breve{\mathbf{x}}^{*\top}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*$ is independent of $\frac{\mathbf{L}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*}{\breve{\mathbf{x}}^{*\top}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*}$ for given $\breve{\mathbf{x}}^*$. Therefore, $\xi^*$ is independent of $\breve{\mathbf{x}}^{*\top}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}^* \cdot \mathbf{L}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*/\breve{\mathbf{x}}^{*\top}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*$ and $\xi$ is independent of $\breve{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}} \cdot \mathbf{L}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}/\breve{\mathbf{x}}^\top\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}$. Furthermore, it holds from the proof of Theorem 1 of [15] that

$$\breve{\mathbf{x}}^{*\top}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}^*\frac{\mathbf{L}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*}{\breve{\mathbf{x}}^{*\top}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*} \sim t_k\left(n_1 + n_2 - p; \mathbf{L}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}^*, \frac{\breve{\mathbf{x}}^{*\top}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}^*}{n_1 + n_2 - p}\mathbf{L}\mathbf{R}_{\breve{\mathbf{x}}^*}\mathbf{L}^\top\right)$$

with $\mathbf{R}_{\breve{\mathbf{x}}^*} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}^*\breve{\mathbf{x}}^{*\top}\boldsymbol{\Sigma}^{-1}/\breve{\mathbf{x}}^{*\top}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}^*$.

Thus, we obtain the following stochastic representation of $\hat{\boldsymbol{\theta}}$ which is given by

$$\hat{\boldsymbol{\theta}} \stackrel{d}{=} (n_1+n_2-2)\xi^{-1}\left(\mathbf{L}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}} + \sqrt{\frac{\breve{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}}{n_1+n_2-p}}\big(\mathbf{L}\mathbf{R}_{\breve{\mathbf{x}}}\mathbf{L}^\top\big)^{1/2}\mathbf{t}_0\right),$$

where $\mathbf{R}_{\breve{\mathbf{x}}} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}\breve{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}/\breve{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}$, $\xi \sim \chi^2_{n_1+n_2-p-1}$, $\breve{\mathbf{x}} \sim \mathcal{N}_p\Big(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \Big(\frac{1}{n_1}+\frac{1}{n_2}\Big)\boldsymbol{\Sigma}\Big)$, and $\mathbf{t}_0 \sim t_k(n_1+n_2-p, \mathbf{0}_k, \mathbf{I}_k)$. Moreover, $\xi$, $\breve{\mathbf{x}}$ and $\mathbf{t}_0$ are mutually independent. The theorem is proved. □

In the next corollary we consider the special case when $k = 1$, i. e., $\mathbf{L} = \mathbf{l}^\top$ is a $p$-dimensional vector of constants.

**Corollary 1.** *Let $\lambda = 1/n_1 + 1/n_2$ and let $\mathbf{l}$ be a $p$-dimensional vector of constants. Then, under the condition of Theorem 1, the stochastic representation of $\hat{\theta} = \mathbf{l}^\top\hat{\mathbf{a}}$ is given by*

$$\hat{\theta} \stackrel{d}{=} (n_1+n_2-2)\xi^{-1}\left(\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) + \sqrt{\left(\lambda + \frac{\lambda(p-1)}{n_1+n_2-p}u\right)\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}}z_0\right),$$

*where*

$$\xi \sim \chi^2_{n_1+n_2-p-1},\ z_0 \sim \mathcal{N}(0,1),\ u \sim \mathcal{F}\big(p-1, n_1+n_2-p, (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\mathbf{R}_{\mathbf{l}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)/\lambda\big)$$

*(non-central $\mathcal{F}$-distribution with $p-1$ and $n_1+n_2-p$ degrees of freedom and non-centrality parameter $(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\mathbf{R}_{\mathbf{l}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)/\lambda$) with $\mathbf{R}_{\mathbf{l}} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{l}\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}/\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}$; $\xi$, $z_0$ and $u$ are mutually independently distributed.*

*Proof.* We get from Theorem 1 that

$$\hat{\theta} \stackrel{d}{=} (n_1+n_2-2)\xi^{-1}\left(\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}} + t_0\sqrt{\frac{\breve{\mathbf{x}}^\top\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}}{n_1+n_2-p}\cdot\mathbf{l}^\top\mathbf{R}_{\breve{\mathbf{x}}}\mathbf{l}}\right) =$$

$$= (n_1+n_2-2)\xi^{-1}\left(\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}} + \frac{t_0}{\sqrt{n_1+n_2-p}}\sqrt{\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}}\sqrt{\breve{\mathbf{x}}^\top\mathbf{R}_{\mathbf{l}}\breve{\mathbf{x}}}\right),$$

where $\mathbf{R}_{\mathbf{l}} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{l}\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}/\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}$; $\xi \sim \chi^2_{n_1+n_2-p-1}$, $t_0 \sim t(n_1+n_2-p, 0, 1)$, and $\breve{\mathbf{x}} \sim \mathcal{N}_p(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2, \lambda\boldsymbol{\Sigma})$ with $\lambda = 1/n_1 + 1/n_2$; $\xi$, $t_0$ and $\breve{\mathbf{x}}$ are mutually independent.

Because of $\breve{\mathbf{x}} \sim \mathcal{N}_p(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2, \lambda\boldsymbol{\Sigma})$, $\mathbf{R}_{\mathbf{l}}\boldsymbol{\Sigma}\mathbf{R}_{\mathbf{l}} = \mathbf{R}_{\mathbf{l}}$, and $\mathrm{tr}[\mathbf{R}_{\mathbf{l}}\boldsymbol{\Sigma}] = p-1$, the application of Corollary 5.1.3a of [33] leads to

$$\zeta = \lambda^{-1}\breve{\mathbf{x}}^\top\mathbf{R}_{\mathbf{l}}\breve{\mathbf{x}} \sim \chi^2_{p-1}\big(\delta^2\big),$$

where $\delta^2 = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\mathbf{R}_{\mathbf{l}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)/\lambda$. Moreover, since $\mathbf{R}_{\mathbf{l}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\mathbf{l} = \mathbf{0}$, the application of Theorem 5.5.1 of [33] proves that $\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}$ and $\zeta$ are independently distributed.

Finally, we note that the random variable $t_0 \sim t(n_1+n_2-p, 0, 1)$ has the following stochastic representation

$$t_0 \stackrel{d}{=} z_0\sqrt{\frac{n_1+n_2-p}{w}},$$

where $z_0 \sim \mathcal{N}(0,1)$ and $w \sim \chi^2_{n_1+n_2-p}$; $z_0$ and $w$ are independent. Hence,

$$\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}} + t_0\sqrt{\frac{\lambda\zeta\cdot\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}}{n_1+n_2-p}}\Bigg|\zeta, w \sim \mathcal{N}\left(\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \lambda\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}\left(1+\frac{\zeta}{w}\right)\right) =$$

$$= \mathcal{N}\left(\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, \lambda\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}\left(1+\frac{p-1}{n_1+n_2-p}u\right)\right),$$

where

$$u = \frac{\zeta/(p-1)}{w/(n_1+n_2-p)} \sim \mathcal{F}\big(p-1, n_1+n_2-p, (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \mathbf{R}_{\mathbf{l}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)/\lambda\big).$$

Putting all above together we get the statement of the corollary. □

2.2. **Test for the population discriminant function coefficients.** An important question, when the discriminant analysis is performed, is to decide which coefficients are the most influential in making the decision. The most popular methods in the literature are (cf. [37, Section 5.5]): (i) standardized coefficients; (ii) partial $F$-values; (iii) correlations between the variables and the discriminant function. In Theorem 5.7A of [36] it is argued that each of these three methods has several drawbacks. For instance, the correlations between the variables and the discriminant function do not show the multivariate contribution of each variable, but provide only univariate information how each variable separates the groups, ignoring the presence of other variables.

In this subsection, we propose an alternative approach based on the statistical hypothesis test. The exact statistical tests will be derived under the null hypothesis that two population discriminant function coefficients are equal (two-sided test) versus the alternative hypothesis that a coefficient in the discriminant function is larger than another one (one-sided test). The testing hypothesis for the equality of the $i$-th and the $j$-th coefficients in the population discriminant function is described by

$$H_0: \ a_i = a_j \quad \text{against} \quad H_1: \ a_i \neq a_j\,, \tag{5}$$

while in the case of one-sided test we test

$$H_0: \ a_i \leq a_j \quad \text{against} \quad H_1: \ a_i > a_j\,. \tag{6}$$

The following test statistic is suggested in both the cases:

$$T = \sqrt{n_1+n_2-p-1}\times \\ \times \frac{\mathbf{l}^\top \mathbf{S}_{pl}^{-1}(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)})}{\sqrt{\mathbf{l}^\top \mathbf{S}_{pl}^{-1}\mathbf{l}}\sqrt{(n_1+n_2-2)(\frac{1}{n_1}+\frac{1}{n_2})+(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)})^\top \hat{\mathbf{R}}_{\mathbf{l}}(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)})}}$$

with

$$\hat{\mathbf{R}}_{\mathbf{l}} = \mathbf{S}_{pl}^{-1} - \frac{\mathbf{S}_{pl}^{-1}\mathbf{l}\mathbf{l}^\top \mathbf{S}_{pl}^{-1}}{\mathbf{l}^\top \mathbf{S}_{pl}^{-1}\mathbf{l}} \quad \text{and} \quad \mathbf{l} = (0,\ldots,0,\underbrace{1}_{i},0\ldots,0,\underbrace{-1}_{j},0,\ldots,0)^\top.$$

The distribution of $T$ follows from [13, Theorem 6] and it is summarized in Theorem 2.

**Theorem 2.** *Let $\lambda = 1/n_1+1/n_2$ and let $\mathbf{l}$ be a $p$-dimensional vector of constants. Then, under the condition of Theorem 1,*

*(a) the density of $T$ is given by*

$$f_T(x) = \frac{n_1+n_2-p}{\lambda(p-1)}\int_0^\infty f_{t_{n_1+n_2-p-1,\delta_1(y)}}(x)\times \\ \times f_{\mathcal{F}_{p-1,n_1+n_2-p,s/\lambda}}\left(\frac{n_1+n_2-p}{\lambda(p-1)}y\right)dy$$

*with $\delta_1(y) = \eta/\sqrt{\lambda+y}$, $\eta = \frac{\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{\sqrt{\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}}}$, and $s = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\mathbf{R}_{\mathbf{l}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$; the symbol $f_G(\cdot)$ denotes the density of the distribution $G$.*

*(b) Under the null hypothesis it holds that $T \sim t_{n_1+n_2-p-1}$ and $T$ is independent of $\big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\big)^\top\hat{\mathbf{R}}_{\mathbf{l}}\big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\big)$.*

Theorem 2 shows that the test statistics $T$ has a standard $t$-distribution under the null hypothesis. As a result, the suggested test will reject the null hypothesis of the two-sided test (5) as soon as $|T| > t_{n_1+n_2-p-1;1-\alpha/2}$.

The situation is more complicated in the case of the one-sided test (6). In this case the maximal probability of the type I error has to be controlled. For that reason, we first calculate the probability of rejection of the null hypothesis for all possible parameter values and after that we calculate its maximum for the parameters, which correspond to the null hypothesis in (6). Since the distribution of $T$ depends on $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ only over $\eta$ and $s$ (see Theorem 2), the task of finding the maximum is significantly simplified. Let $F_G(\cdot)$ denote the distribution function of the distribution $G$. For any constant $q$, we get

$$\begin{aligned}
\mathbb{P}(T > q) &= \int_q^{+\infty} f_T(x)dx = \\
&= \int_q^{+\infty} \frac{n_1+n_2-p}{\lambda(p-1)} \int_0^{\infty} f_{t_{n_1+n_2-p-1,\delta_1(y)}}(x) f_{\mathcal{F}_{p-1,n_1+n_2-p,s/\lambda}}\left(\frac{n_1+n_2-p}{\lambda(p-1)}y\right)dydx = \\
&= \frac{n_1+n_2-p}{\lambda(p-1)} \int_0^{\infty} f_{\mathcal{F}_{p-1,n_1+n_2-p,s/\lambda}}\left(\frac{n_1+n_2-p}{\lambda(p-1)}y\right) \int_q^{+\infty} f_{t_{n_1+n_2-p-1,\delta_1(y)}}(x)dxdy = \\
&= \frac{n_1+n_2-p}{\lambda(p-1)} \int_0^{\infty} \Big(1 - F_{t_{n_1+n_2-p-1,\delta_1(y)}}(q)\Big) f_{\mathcal{F}_{p-1,n_1+n_2-p,s/\lambda}}\left(\frac{n_1+n_2-p}{\lambda(p-1)}y\right)dy \leq \\
&\leq \frac{n_1+n_2-p}{\lambda(p-1)} \int_0^{\infty} \big(1 - F_{t_{n_1+n_2-p-1,0}}(q)\big) f_{\mathcal{F}_{p-1,n_1+n_2-p,s/\lambda}}\left(\frac{n_1+n_2-p}{\lambda(p-1)}y\right)dy = \\
&= 1 - F_{t_{n_1+n_2-p-1,0}}(q),
\end{aligned}$$

where the last equality follows from the fact that the distribution function of the non-central $t$-distribution is a decreasing function in non-centrality parameter and $\delta_1(y) \leq 0$. Consequently, we get $q = t_{n_1+n_2-p-1;1-\alpha}$ and the one-sided test rejects the null hypothesis in (6) as soon as $T > t_{n_1+n_2-p-1;1-\alpha}$.

2.3. **Classification analysis.** Having a new observation vector $\mathbf{x}$, we classify it to one of the two groups under consideration. Assuming that no prior information is available about the classification result, i. e. the prior probability of each group is $1/2$, the decision, which is based on the optimal rule, is to assign the observation vector $\mathbf{x}$ to the first group as soon as the following inequality holds (cf. Section 6.2 of [36]):

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} > \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \tag{7}$$

and to the second group otherwise. The error rate is defined as the probability of classifying the observation $\mathbf{x}$ into one group, while it comes from another one. The book [36] presented the expression of the error rate expressed as

$$\begin{aligned}
ER_p(\Delta) &= \frac{1}{2}\mathbb{P}(\text{classify to the first group} \mid \text{second group is true}) + \\
&\quad + \frac{1}{2}\mathbb{P}(\text{classify to the second group} \mid \text{first group is true}) = \\
&= \Phi\left(-\frac{\Delta}{2}\right) \quad \text{with} \quad \Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,,
\end{aligned}$$

where $\Phi(\cdot)$ denotes the distribution function of the standard normal distribution.

In practice, however, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ are unknown quantities and the decision is based on the inequality

$$\left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right)^\top \mathbf{S}_{pl}^{-1}\mathbf{x} > \frac{1}{2}\left(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\right)^\top \mathbf{S}_{pl}^{-1}\left(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}\right) \tag{8}$$

instead. Next, we derive the error rate of the decision rule (8). Let

$$\hat{d} = \Big(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\Big)^\top \mathbf{S}_{pl}^{-1}\mathbf{x} - \frac{1}{2}\Big(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\Big)^\top \mathbf{S}_{pl}^{-1}\Big(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}\Big) = \\ = \Big(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\Big)^\top \mathbf{S}_{pl}^{-1}\Big(\mathbf{x} - \frac{1}{2}\Big(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}\Big)\Big).$$

In Theorem 3 we present the stochastic representation of $\hat{d}$.

**Theorem 3.** *Let $\lambda = 1/n_1 + 1/n_2$. Then, under the condition of Theorem 1, the stochastic representation of $\hat{d}$ is given by*

$$\hat{d} \stackrel{d}{=} \frac{n_1+n_2-2}{\xi}\Bigg((-1)^{i-1}\frac{\lambda n_i - 2}{2\lambda n_i}\Big(\lambda\xi_2 + (\Delta+\sqrt{\lambda}w_0)^2\Big) + \frac{(-1)^{i-1}}{\lambda n_i}\Big(\Delta^2 + \sqrt{\lambda}\Delta w_0\Big) + \\ + \sqrt{\Big(1 + \frac{1}{n_1+n_2} + \frac{p-1}{n_1+n_2-p}u\Big)\lambda\xi_2 + (\Delta+\sqrt{\lambda}w_0)^2}\, z_0\Bigg) \quad \textit{for} \quad i = 1,2, \quad (9)$$

*where $u|\xi_1,\xi_2,w_0 \sim \mathcal{F}\big(p-1, n_1+n_2-p, (n_1+n_2)^{-1}\xi_1\big)$ with $\xi_1|\xi_2,w_0 \sim \chi_{p-1,\delta^2_{\xi_2,w_0,i}}$ and $\delta^2_{\xi_2,w_0,i} = \frac{n_1 n_2}{n_i^2}\frac{\Delta^2\xi_2}{\lambda\xi_2+(\Delta+\sqrt{\lambda}w_0)^2}$, $z_0, w_0 \sim \mathcal{N}(0,1)$, $\xi \sim \chi^2_{n_1+n_2-p-1}$, $\xi_2 \sim \chi^2_{p-1}$; $\xi$, $z_0$ are independent of $u, \xi_1, \xi_2, w_0$, where $\xi_2$ and $w_0$ are independent as well.*

*Proof.* Let $\mathbf{x} \sim \mathcal{N}_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$. Since $\bar{\mathbf{x}}^{(1)}$, $\bar{\mathbf{x}}^{(2)}$, $\mathbf{x}$, and $\mathbf{S}_{pl}$ are independently distributed, we get that the conditional distribution of $\hat{d}$ given $\bar{\mathbf{x}}^{(1)} = \mathbf{x}_0^{(1)}$ and $\bar{\mathbf{x}}^{(2)} = \mathbf{x}_0^{(2)}$ is equal to the distribution of $d_0$ defined by

$$d_0 = \Big(\bar{\mathbf{x}}_0^{(1)} - \bar{\mathbf{x}}_0^{(2)}\Big)^\top \mathbf{S}_{pl}^{-1}\tilde{\mathbf{x}},$$

where $\tilde{\mathbf{x}} = \mathbf{x} - \frac{1}{2}\Big(\bar{\mathbf{x}}_0^{(1)} + \bar{\mathbf{x}}_0^{(2)}\Big) \sim \mathcal{N}_p\Big(\boldsymbol{\mu}_i - \frac{1}{2}\Big(\bar{\mathbf{x}}_0^{(1)} + \bar{\mathbf{x}}_0^{(2)}\Big), \boldsymbol{\Sigma}\Big)$, $(n_1+n_2-2)\mathbf{S}_{pl} \sim \sim \mathcal{W}_p(n_1+n_2-2, \boldsymbol{\Sigma})$, $\tilde{\mathbf{x}}$ and $\mathbf{S}_{pl}$ are independent.

Following the proof of Corollary 1, we get

$$d_0 \stackrel{d}{=} (n_1+n_2-2)\xi^{-1}\Bigg(\Big(\bar{\mathbf{x}}_0^{(1)} - \bar{\mathbf{x}}_0^{(2)}\Big)^\top \boldsymbol{\Sigma}^{-1}\Big(\boldsymbol{\mu}_i - \frac{1}{2}\Big(\bar{\mathbf{x}}_0^{(1)} + \bar{\mathbf{x}}_0^{(2)}\Big)\Big) + \\ + \sqrt{\Big(1 + \frac{(p-1)}{n_1+n_2-p}u\Big)\Big(\bar{\mathbf{x}}_0^{(1)} - \bar{\mathbf{x}}_0^{(2)}\Big)^\top \boldsymbol{\Sigma}^{-1}\Big(\bar{\mathbf{x}}_0^{(1)} - \bar{\mathbf{x}}_0^{(2)}\Big)}\, z_0\Bigg),$$

where $u \sim \mathcal{F}\Big(p-1, n_1+n_2-p, \Big(\boldsymbol{\mu}_i - \frac{1}{2}\big(\bar{\mathbf{x}}_0^{(1)} + \bar{\mathbf{x}}_0^{(2)}\big)\Big)^\top \mathbf{R}_0\Big(\boldsymbol{\mu}_i - \frac{1}{2}\big(\bar{\mathbf{x}}_0^{(1)} + \bar{\mathbf{x}}_0^{(2)}\big)\Big)\Big)$ with $\mathbf{R}_0 = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\Big(\bar{\mathbf{x}}_0^{(1)} - \bar{\mathbf{x}}_0^{(2)}\Big)\Big(\bar{\mathbf{x}}_0^{(1)} - \bar{\mathbf{x}}_0^{(2)}\Big)^\top \boldsymbol{\Sigma}^{-1} / \Big(\bar{\mathbf{x}}_0^{(1)} - \bar{\mathbf{x}}_0^{(2)}\Big)^\top \boldsymbol{\Sigma}^{-1}\Big(\bar{\mathbf{x}}_0^{(1)} - \bar{\mathbf{x}}_0^{(2)}\Big)$, $z_0 \sim \mathcal{N}(0,1)$, and $\xi \sim \chi^2_{n_1+n_2-p-1}$; $\xi$, $z_0$ and $u$ are mutually independently distributed.

By using that

$$\boldsymbol{\mu}_i - \frac{1}{2}\Big(\bar{\mathbf{x}}_0^{(1)} + \bar{\mathbf{x}}_0^{(2)}\Big) = \boldsymbol{\mu}_i - \bar{\mathbf{x}}_0^{(i)} + (-1)^{i-1}\frac{1}{2}\Big(\bar{\mathbf{x}}_0^{(1)} - \bar{\mathbf{x}}_0^{(2)}\Big)$$

and $\Big(\bar{\mathbf{x}}_0^{(1)} - \bar{\mathbf{x}}_0^{(2)}\Big)^\top \mathbf{R}_0 = \mathbf{0}$, we get

$$\hat{d} \stackrel{d}{=} \frac{n_1+n_2-2}{\xi}\Bigg(\frac{(-1)^{i-1}}{2}\Big(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\Big)^\top \boldsymbol{\Sigma}^{-1}\Big(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\Big) - \\ - \Big(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\Big)^\top \boldsymbol{\Sigma}^{-1}\Big(\bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}_i\Big) + \\ + \sqrt{\Big(1 + \frac{p-1}{n_1+n_2-p}u\Big)\big(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\big)^\top \boldsymbol{\Sigma}^{-1}\big(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}\big)}\, z_0\Bigg),$$

where $u|\bar{\mathbf{x}}^{(1)}, \bar{\mathbf{x}}^{(2)} \sim \mathcal{F}\Big(p-1, n_1+n_2-p, \big(\bar{\mathbf{x}}^{(i)}-\boldsymbol{\mu}_i\big)^\top \mathbf{R}_{\mathbf{x}} \big(\bar{\mathbf{x}}^{(i)}-\boldsymbol{\mu}_i\big)\Big)$ with $\mathbf{R}_{\mathbf{x}} = \boldsymbol{\Sigma}^{-1} - -\boldsymbol{\Sigma}^{-1}\big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\big)\big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\big)^\top \boldsymbol{\Sigma}^{-1} / \big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\big)^\top \boldsymbol{\Sigma}^{-1}\big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\big)$, $z_0 \sim \mathcal{N}(0,1)$, and $\xi \sim \chi^2_{n_1+n_2-p-1}$; $\xi$, $z_0$ are independent of $u, \bar{\mathbf{x}}^{(1)}, \bar{\mathbf{x}}^{(2)}$.

Since $\bar{\mathbf{x}}^{(1)}$ and $\bar{\mathbf{x}}^{(2)}$ are independent and normally distributed, we get that

$$\begin{pmatrix} \bar{\mathbf{x}}^{(i)}-\boldsymbol{\mu}_i \\ \bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)} \end{pmatrix} \sim \mathcal{N}_{2p}\left( \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\mu}_1-\boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \frac{1}{n_i}\boldsymbol{\Sigma} & \frac{(-1)^{i-1}}{n_i}\boldsymbol{\Sigma} \\ \frac{(-1)^{i-1}}{n_i}\boldsymbol{\Sigma} & \lambda\boldsymbol{\Sigma} \end{pmatrix} \right)$$

and, consequently,

$$\bar{\mathbf{x}}^{(i)}-\boldsymbol{\mu}_i \big| \big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\big) \sim \mathcal{N}_p\left( \frac{(-1)^{i-1}}{\lambda n_i}\big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big), \frac{1}{n_1+n_2}\boldsymbol{\Sigma}\right),$$

where we used that $\frac{1}{n_i} - \frac{1}{\lambda n_i^2} = \frac{1}{n_1+n_2}$.

The application of Theorem 5.5.1 in [33] shows that given $(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)})$ the random variables $(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)})^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}^{(i)}-\boldsymbol{\mu}_i)$ and $(\bar{\mathbf{x}}^{(i)}-\boldsymbol{\mu}_i)^\top\mathbf{R}_{\mathbf{x}}(\bar{\mathbf{x}}^{(i)}-\boldsymbol{\mu}_i)$ are independently distributed with

$$\begin{aligned} &\Big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\Big)^\top \boldsymbol{\Sigma}^{-1}\Big(\bar{\mathbf{x}}^{(i)}-\boldsymbol{\mu}_i\Big) \Big| \Big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\Big) \sim \\ &\quad \sim \mathcal{N}\Bigg( \frac{(-1)^{i-1}}{\lambda n_i}\Big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\Big)^\top \boldsymbol{\Sigma}^{-1}\Big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\Big), \\ &\quad \frac{1}{n_1+n_2}\Big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\Big)^\top \boldsymbol{\Sigma}^{-1}\Big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\Big)\Bigg) \end{aligned}$$

and, by using Corollary 5.1.3a of [33],

$$(n_1+n_2)\Big(\bar{\mathbf{x}}^{(i)}-\boldsymbol{\mu}_i\Big)^\top \mathbf{R}_{\mathbf{x}}\Big(\bar{\mathbf{x}}^{(i)}-\boldsymbol{\mu}_i\Big) \Big| \Big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\Big) \sim \chi_{p-1,\delta^2_{\mathbf{x}}}$$

with

$$\begin{aligned} \delta^2_{\mathbf{x}} &= \frac{n_1+n_2}{\lambda^2 n_i^2}\Big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\Big)^\top \mathbf{R}_{\mathbf{x}}\Big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}-(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\Big) = \\ &= \frac{n_1+n_2}{\lambda^2 n_i^2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top \mathbf{R}_{\mathbf{x}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) = \\ &= \frac{n_1+n_2}{\lambda^2 n_i^2}\frac{(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)})^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)})}\Big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\Big)^\top \mathbf{R}_{\boldsymbol{\mu}}\Big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\Big), \end{aligned}$$

where we use that $(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)})^\top\mathbf{R}_{\mathbf{x}} = \mathbf{0}$ and

$$\mathbf{R}_{\boldsymbol{\mu}} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}/(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2).$$

As a result, we get

$$\begin{aligned} \hat{d} \stackrel{d}{=} \frac{n_1+n_2-2}{\xi}\Bigg( &(-1)^{i-1}\frac{\lambda n_i-2}{2\lambda n_i}\Delta^2_{\mathbf{x}} + \frac{(-1)^{i-1}}{\lambda n_i}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}) + \\ &+ \sqrt{\left(1+\frac{1}{n_1+n_2}+\frac{p-1}{n_1+n_2-p}u\right)}\Delta_{\mathbf{x}} z_0\Bigg), \end{aligned}$$

where $\Delta^2_{\mathbf{x}} = (\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)})^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)})$, $u|\bar{\mathbf{x}}^{(1)}, \bar{\mathbf{x}}^{(2)} \sim \mathcal{F}\big(p-1, n_1+n_2-p, (n_1+n_2)^{-1}\xi_1\big)$ with $\xi_1 \sim \chi_{p-1,\delta^2_{\mathbf{x}}}$, $z_0 \sim \mathcal{N}(0,1)$, and $\xi \sim \chi^2_{n_1+n_2-p-1}$; $\xi$, $z_0$ are independent of $u, \xi_1, \bar{\mathbf{x}}^{(1)}, \bar{\mathbf{x}}^{(2)}$.

Finally, it holds with $\Delta^2 = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ that

$$\Delta^2_{\mathbf{x}} = \Big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\Big)^\top \mathbf{R}_{\boldsymbol{\mu}}\Big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\Big) + \frac{\big((\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)})\big)^2}{\Delta^2},$$

where both summands are independent following Theorem 5.5.1 in [33]. The application of Corollary 5.1.3a in [33] leads to

$$\lambda^{-1}\big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\big)^{\top}\mathbf{R}_{\boldsymbol{\mu}}\big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\big)\sim\chi^2_{p-1}$$

and

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^{\top}\boldsymbol{\Sigma}^{-1}\big(\bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}\big)\sim\mathcal{N}(\Delta^2,\lambda\Delta^2).$$

We get the stochastic representation of $\hat{d}$ from the last statement expressed as

$$\hat{d}\overset{d}{=}\frac{n_1+n_2-2}{\xi}\Bigg((-1)^{i-1}\frac{\lambda n_i-2}{2\lambda n_i}\Big(\lambda\xi_2+(\Delta+\sqrt{\lambda}w_0)^2\Big)+\frac{(-1)^{i-1}}{\lambda n_i}\Big(\Delta^2+\sqrt{\lambda}\Delta w_0\Big)+$$
$$+\sqrt{\left(1+\frac{1}{n_1+n_2}+\frac{p-1}{n_1+n_2-p}u\right)\lambda\xi_2+(\Delta+\sqrt{\lambda}w_0)^2}\,z_0\Bigg),$$

where $u|\xi_1,\xi_2,w_0\sim\mathcal{F}\big(p-1,n_1+n_2-p,(n_1+n_2)^{-1}\xi_1\big)$ with $\xi_1|\xi_2,w_0\sim\chi_{p-1,\delta^2_{\xi_2,w_0,i}}$ and $\delta^2_{\xi_2,w_0,i}=\frac{n_1+n_2}{\lambda n_i^2}\frac{\Delta^2}{\lambda\xi_2+(\Delta+\sqrt{\lambda}w_0)^2}\xi_2$, $z_0,w_0\sim\mathcal{N}(0,1)$, $\xi\sim\chi^2_{n_1+n_2-p-1}$, $\xi_2\sim\chi^2_{p-1}$; $\xi$, $z_0$ are independent of $u,\xi_1,\xi_2,w_0$, where $\xi_2$ and $w_0$ are independent as well. □

Theorem 3 shows that the distribution of $\hat{d}$ is determined by six random variables $\xi,\xi_1,\xi_2,\ z_0,w_0$, and $u$. Moreover, it depends on $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ only via the quadratic form $\Delta$. As a result, the error rate based on the decision rule (8) is a function of $\Delta$ only and it is calculated by

$$ER_s(\Delta)=\frac{1}{2}\mathbb{P}(\hat{d}>0|\text{ second group is true})+\frac{1}{2}\mathbb{P}(\hat{d}\le 0|\text{ first group is true})\,. \tag{10}$$

The two probabilities in (10) can easily be approximated for all $\Delta$, $p$, $n_1$, and $n_2$ with high precision by applying the results of Theorem 3 via the following simulation study

(i) Fix $\Delta$ and $i\in\{1,2\}$.
(ii) Generate four independent random variables $\xi_b\sim\chi^2_{n_1+n_2-p-1}$, $\xi_{2;b}\sim\chi^2_{p-1}$, $z_{0;b}\sim\mathcal{N}(0,1)$, and $w_{0;b}\sim\mathcal{N}(0,1)$.
(iii) Generate $\xi_{1,b}\sim\chi_{p-1,\delta^2_{\xi_2,w_0,i}}$ with $\delta^2_{\xi_{2,b},w_{0,b},i}=\frac{n_1n_2}{n_i^2}\frac{\Delta^2\xi_{2;b}}{\lambda\xi_{2;b}+(\Delta+\sqrt{\lambda}w_{0;b})^2}$.
(iv) Generate $u\sim\mathcal{F}\big(p-1,n_1+n_2-p,(n_1+n_2)^{-1}\xi_{1,b}\big)$.
(v) Calculate $\hat{d}_b^{(i)}$ following the stochastic representation (9) of Theorem 3.
(vi) Repeat steps (ii)–(v) for $b=1,\dots,B$ leading to the sample $\hat{d}_1^{(i)},\dots,\hat{d}_B^{(i)}$.

The procedure has to be performed for both values of $i=1,2$ where for $i=1$ the relative number of events $\{\hat{d}>0\}$ will approximate the first summand in (10), while for $i=2$ the relative number of events $\{\hat{d}\le 0\}$ will approximate the second summand in (10).

It is important to note that the difference between the error rates calculated for the two decision rules (7) and (8) could be very large as shown in Figure 1, where $ER_p(\Delta)$ and $ER_s(\Delta)$ are calculated for several values of $n_1=n_2\in\{50,100,150,250\}$ with fixed values of $p\in\{10,25,50,75\}$. If $p=10$, we do not observe large differences between $ER_p(\Delta)$ and $ER_s(\Delta)$ computed for different sample sizes. However, this statement does not hold any longer when $p$ becomes comparable to both $n_1$ and $n_2$ as documented for $p=50$ and $p=75$. This case is known in the literature as a large-dimensional asymptotic regime and it is investigated in detail in Section 3.

## 3. Discriminant analysis under large-dimensional asymptotics

In this section we derive the asymptotic distribution of the discriminant function coefficients under the high-dimensional asymptotic regime, i. e., when the dimension increases together with the sample sizes and they all tend to infinity. More precisely, we assume that $p/(n_1+n_2)\to c\in[0,1)$ as $n_1+n_2\to\infty$.

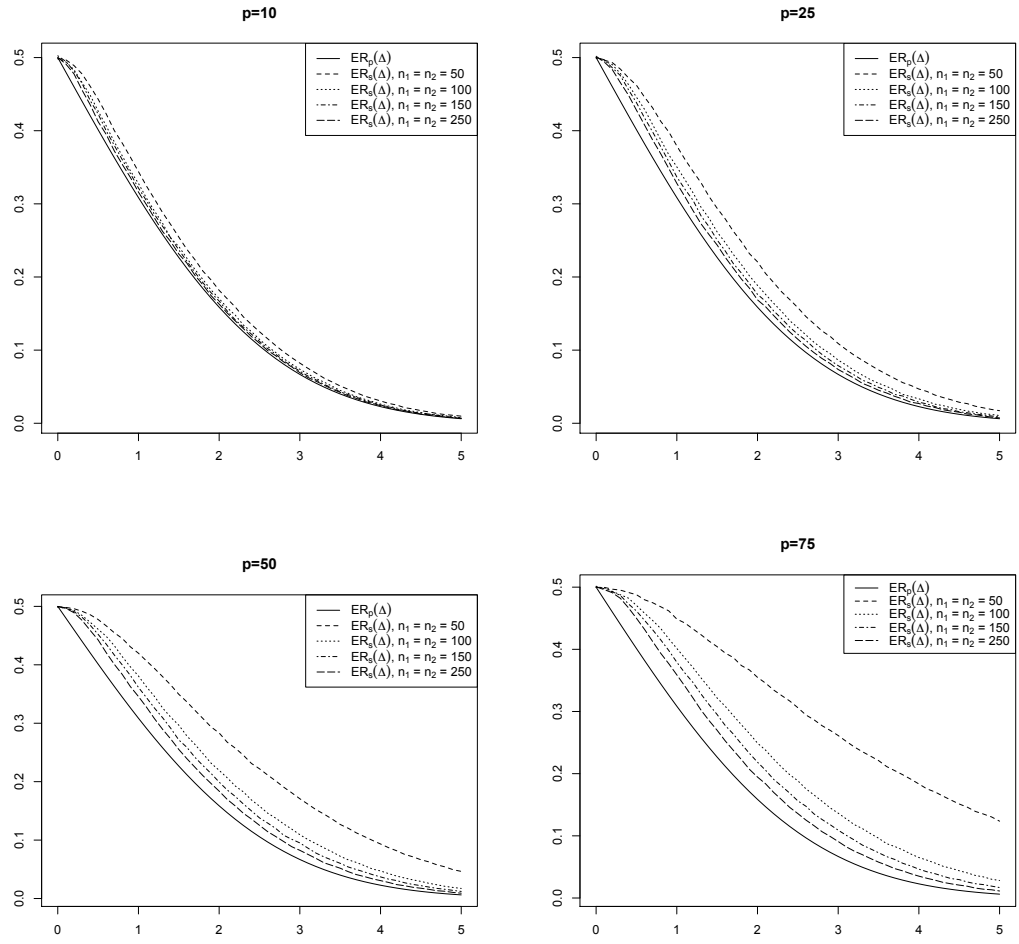



FIGURE 1. Error rates $ER_p(\Delta)$ and $ER_s(\Delta)$ as functions of $\Delta$ for $p \in \{10, 25, 50, 75\}$ and $ER_s(\Delta)$.

The following conditions are needed for the validity of the asymptotic results:

(A1) There exists $\gamma \geq 0$ such that $p^{-\gamma}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < \infty$ uniformly on $p$.

(A2) $0 < \lim\limits_{(n_1,n_2)\to\infty} (n_1/n_2) < \infty$.

It is important to note, that no assumption on the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$ is imposed like they are uniformly bounded on $p$. The asymptotic results are also valid when $\boldsymbol{\Sigma}$ possesses unbounded spectrum, as well as when its smallest eigenvalue tends to zero as $p \to \infty$. The constant $\gamma$ is a technical one and it controls the growth rate of the quadratic form. In Theorem 4 the asymptotic distribution of linear combinations of the discriminant function coefficients is provided.

**Theorem 4.** *Assume (A1) and (A2). Let* $\mathbf{l}$ *be a* $p$*-dimensional vector of constants such that* $p^{-\gamma}\mathbf{l}^\top \boldsymbol{\Sigma}^{-1}\mathbf{l} < \infty$ *is uniformly on* $p$*,* $\gamma \geq 0$*. Then, under the conditions of Theorem 1, the asymptotic distribution of* $\hat{\theta} = \mathbf{l}^\top \hat{\mathbf{a}}$ *is given by*

$$\sqrt{n_1 + n_2}\sigma_\gamma^{-1}\left(\hat{\theta} - \frac{1}{1-c}\mathbf{l}^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

*for* $p/(n_1+n_2) \to c \in [0,1)$ *as* $n_1+n_2 \to \infty$ *with*

$$\sigma_\gamma^2 = \frac{1}{(1-c)^3}\Big(\big(\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big)^2 + \mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) + \\ + \lambda(n_1+n_2)\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}\mathbb{1}_{\{0\}}(\gamma)\Big),$$

*where* $\mathbb{1}_{\mathcal{A}}(\cdot)$ *denotes the indicator function of set* $\mathcal{A}$.

*Proof.* Using the stochastic representation (5) of Corollary 1, we get

$$\sqrt{n_1+n_2}\sigma_\gamma^{-1}\left(\hat{\theta} - \frac{1}{1-c}\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right) \overset{d}{=} \\ \overset{d}{=} \sqrt{n_1+n_2}\left((n_1+n_2-2)\xi^{-1} - \frac{1}{1-c}\right)\frac{p^{-\gamma}\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{p^{-\gamma}\sigma_\gamma} + \\ + \sqrt{\lambda(n_1+n_2)}\frac{n_1+n_2-2}{\xi}\sqrt{\left(p^{-\gamma}+p^{-\gamma}\frac{p-1}{n_1+n_2-p}u\right)}\frac{\sqrt{p^{-\gamma}\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}}}{p^{-\gamma}\sigma_\gamma}z_0,$$

where

$$\xi \sim \chi^2_{n_1+n_2-p-1}, \quad z_0 \sim \mathcal{N}(0,1), \quad u \sim \mathcal{F}\big(p-1, n_1+n_2-p, (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\mathbf{R}_{\mathbf{l}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)/\lambda\big)$$

with $\mathbf{R}_{\mathbf{l}} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{l}\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}/\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}$; $\xi$, $z_0$ and $u$ are mutually independently distributed.

Since, $\xi \sim \chi^2_{n_1+n_2-p-1}$, we get that

$$\sqrt{n_1+n_2-p-1}\left(\frac{\xi}{n_1+n_2-p-1} - 1\right) \xrightarrow{\mathcal{D}} \mathcal{N}(0,2)$$

for $p/(n_1+n_2) \to c \in [0,1)$ as $n_1+n_2 \to \infty$ and, consequently,

$$\sqrt{n_1+n_2}\left((n_1+n_2-2)\xi^{-1} - \frac{1}{1-c}\right) = \frac{\sqrt{n_1+n_2}}{\sqrt{n_1+n_2-p-1}}\frac{n_1+n_2-p-1}{\xi}\frac{1}{1-c} \times \\ \times \sqrt{n_1+n_2-p-1}\left((1-c)\frac{n_1+n_2-2}{n_1+n_2-p-1} - \frac{\xi}{n_1+n_2-p-1}\right) \xrightarrow{\mathcal{D}} \\ \xrightarrow{\mathcal{D}} \tilde{z}_0 \sim \mathcal{N}\left(0, \frac{2}{1-c}\right)$$

for $\frac{p}{n_1+n_2} = c + o((n_1+n_2)^{-1/2})$, where $z_0$ and $\tilde{z}_0$ are independent.

Furthermore, we get (see, [14, Lemma 3])

$$p^{-\gamma} + p^{-\gamma}\frac{p-1}{n_1+n_2-p}u - \mathbb{1}_{\{0\}}(\gamma) - \\ - \frac{c}{1-c}\left(\mathbb{1}_{\{0\}}(\gamma) + \frac{p^{-\gamma}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\mathbf{R}_{\mathbf{l}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)}{c\lambda(n_1+n_2)}\right) \xrightarrow{\text{a.s.}} 0.$$

Putting the above results together, we get the statement of the theorem with

$$\sigma_\gamma^2 = \frac{1}{(1-c)^3}\Big(2\big(\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big)^2 + \mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\mathbf{R}_{\mathbf{l}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) + \\ + \lambda(n_1+n_2)\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}\mathbb{1}_{\{0\}}(\gamma)\Big) = \\ = \frac{1}{(1-c)^3}\Big(\big(\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\big)^2 + \mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) + \\ + \lambda(n_1+n_2)\mathbf{l}^\top\boldsymbol{\Sigma}^{-1}\mathbf{l}\mathbb{1}_{\{0\}}(\gamma)\Big).$$

□

The results of Theorem 4 show that the quantity $\gamma$ is present only in the asymptotic variance $\sigma^2_\gamma$. Moreover, if $\gamma > 0$, then the factor $\lambda(n_1+n_2)$ vanishes and therefore the assumption (A2) is no longer needed. However, in the case of $\gamma = 0$ we need (A2) in order to keep the variance bounded. We further investigate this point via simulations in Subsection 3.3, by choosing $\gamma > 0$ and considering small $n_1$ and large $n_2$ such that $n_1/n_2 \to 0$.

3.1. **Classification analysis in high dimension.** The error rate of the classification analysis based on the optimal decision rule (7) remains the same, independent of $p$ and it is always equal to

$$ER_p(\Delta) = \Phi\left(-\frac{\Delta}{2}\right) \text{ with } \Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\,.$$

In practice, however, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ are not known and, consequently, one has to make the decision based on (8) instead of (7). In Theorem 5, we derived the asymptotic distribution of $\hat{d}$ under the large-dimensional asymptotics.

**Theorem 5.** *Assume (A1) and (A2). Let $p^{-\gamma}\Delta^2 \to \widetilde{\Delta}^2$ and $\lambda n_i \to b_i$ for $p/(n_1+n_2) \to$ $\to c \in [0,1)$ as $n_1+n_2 \to \infty$. Then, under the conditions of Theorem 1, it holds that*

$$\begin{aligned}
&p^{\min(\gamma,1)/2}\left(\frac{\hat{d}}{p^\gamma} - \frac{n_1+n_2-2}{n_1+n_2-p-1}\frac{(-1)^{i-1}}{2}p^{-\gamma}\Delta^2\right) \xrightarrow{\mathcal{D}} \\
&\xrightarrow{\mathcal{D}} \mathcal{N}\Bigg((-1)^{i-1}\frac{c}{1-c}\frac{b_i-2}{2b_i}(b_1+b_2)\mathbb{1}_{\{0\}}(\gamma), \\
&\frac{c}{2(1-c)^3}\widetilde{\Delta}^4\mathbb{1}_{[1,+\infty)}(\gamma) + \frac{1}{(1-c)^3}(c(b_1+b_2)\mathbb{1}_{\{0\}}(\gamma) + \widetilde{\Delta}^2\mathbb{1}_{[0,1]}(\gamma))\Bigg)
\end{aligned}$$

*for $p/(n_1+n_2) \to c \in [0,1)$ as $n_1+n_2 \to \infty$.*

*Proof.* The application of Theorem 3 leads to

$$\begin{aligned}
&p^{\min(\gamma,1)/2}\left(\frac{\hat{d}}{p^\gamma} - \frac{n_1+n_2-2}{n_1+n_2-p-1}\frac{(-1)^{i-1}}{2}p^{-\gamma}\Delta^2\right) \stackrel{d}{=} \\
&\stackrel{d}{=} p^{\min(\gamma,1)/2-1/2}\frac{\sqrt{p}}{\sqrt{n_1+n_2-p-1}}\frac{n_1+n_2-2}{\xi}\sqrt{n_1+n_2-p-1}\times \\
&\times\left(1-\frac{\xi}{n_1+n_2-p-1}\right)\frac{(-1)^{i-1}}{2}p^{-\gamma}\Delta^2 + \frac{n_1+n_2-2}{\xi}\Bigg((-1)^{i-1}\frac{\lambda n_i-2}{2\lambda n_i}\times \\
&\times\left(p^{\min(\gamma,1)/2-\gamma}\lambda\xi_2 + 2p^{\min(\gamma,1)/2-\gamma/2}\sqrt{p^{-\gamma}\Delta^2}\sqrt{\lambda}w_0 + p^{\min(\gamma,1)/2-\gamma}\lambda w_0^2\right) + \\
&+\frac{(-1)^{i-1}}{\lambda n_i}p^{\min(\gamma,1)/2-\gamma/2}\sqrt{p^{-\gamma}\Delta^2}\sqrt{\lambda}w_0\Bigg) + \\
&+\frac{n_1+n_2-2}{\xi}\Bigg(\sqrt{\left(1+\frac{1}{n_1+n_2}+\frac{p-1}{n_1+n_2-p}u\right)}\times \\
&\times\sqrt{p^{\min(\gamma,1)-2\gamma}\lambda\xi_2 + (p^{\min(\gamma,1)/2-\gamma/2}\sqrt{p^{-\gamma}\Delta^2} + p^{\min(\gamma,1)/2-\gamma}\sqrt{\lambda}w_0)^2}z_0\Bigg) \xrightarrow{\mathcal{D}}
\end{aligned}$$

$$\xrightarrow{\mathcal{D}} \mathcal{N}\Bigg( (-1)^{i-1}\frac{c}{1-c}\frac{b_i-2}{2b_i}(b_1+b_2)\mathbb{1}_{\{0\}}(\gamma),$$

$$\frac{c}{2(1-c)^3}\widetilde{\Delta}^4\mathbb{1}_{[1,+\infty)}(\gamma)+\frac{1}{(1-c)^3}(c(b_1+b_2)\mathbb{1}_{\{0\}}(\gamma)+\widetilde{\Delta}^2\mathbb{1}_{[0,1]}(\gamma))\Bigg),$$

where the last line follows from Lemma 3 in [14] and Slutsky Theorem (see, [21, Theorem 1.5]). □

The parameters of the limit distribution derived in Theorem 5 can be significantly simplified in the special case of $n_1 = n_2$ because of $\lambda n_1 = \lambda n_2 = 2$. The results of Theorem 5 are also used to derive the approximate error rate for the decision (8). Let $a = \frac{1}{1-c}\frac{1}{2}p^{-\gamma}\Delta$. Then, the error rate is given by

$$\begin{aligned} ER_s(\Delta) &= \frac{1}{2}\mathbb{P}\Big\{\hat{d}>0\Big|i=2\Big\}+\frac{1}{2}\mathbb{P}\Big\{\hat{d}\le 0\Big|i=1\Big\} = \\ &= \frac{1}{2}\mathbb{P}\left\{p^{\min(\gamma,1)/2}\left(\frac{\hat{d}}{p^\gamma}-(-1)^{i-1}a\right)>-p^{\min(\gamma,1)/2}(-1)^{i-1}a\middle|i=2\right\}+ \\ &\quad+\frac{1}{2}\mathbb{P}\left\{p^{\min(\gamma,1)/2}\left(\frac{\hat{d}}{p^\gamma}-(-1)^{i-1}a\right)\le -p^{\min(\gamma,1)/2}(-1)^{i-1}a\middle|i=1\right\}\approx \\ &\approx \frac{1}{2}\left(1-\Phi\left(\frac{ap^{\min(\gamma,1)/2}-m_2}{v}\right)\right)+\frac{1}{2}\Phi\left(\frac{-ap^{\min(\gamma,1)/2}-m_1}{v}\right) \end{aligned}$$

with

$$m_1 = \frac{c}{1-c}\frac{b_1-2}{2b_1}(b_1+b_2)\mathbb{1}_{\{0\}}(\gamma),\quad m_2 = -\frac{c}{1-c}\frac{b_2-2}{2b_2}(b_1+b_2)\mathbb{1}_{\{0\}}(\gamma),$$

$$v^2 = \frac{c}{2(1-c)^3}\big(p^{-\gamma}\Delta^2\big)^2\mathbb{1}_{[1,+\infty)}(\gamma)+\frac{1}{(1-c)^3}\big(c(b_1+b_2)\mathbb{1}_{\{0\}}(\gamma)+p^{-\gamma}\Delta^2\mathbb{1}_{[0,1]}(\gamma)\big),$$

where we approximate $\widetilde{\Delta}^2$ by $p^{-\gamma}\Delta^2$.

In the special case of $n_1 = n_2$ which leads to $b_1 = b_2 = 2$, we get

$$ER_s(\Delta) = \Phi\left(-h_c\frac{\Delta}{2}\right)$$

with

$$h_c = \frac{p^{\min(\gamma,1)/2-\gamma}\sqrt{1-c}\sqrt{p^{-\gamma}\Delta^2}}{\sqrt{c(p^{-\gamma}\Delta^2)^2\mathbb{1}_{[1,+\infty)}(\gamma)/2+4c\mathbb{1}_{\{0\}}(\gamma)+p^{-\gamma}\Delta^2\mathbb{1}_{[0,1]}(\gamma)}},$$

which is always smaller than one. Furthermore, for $\gamma \in (0,1)$ we get $h_c = \sqrt{1-c}$.

In Figure 2, we plot $ER_s(\Delta)$ as a function of $\Delta \in [0,100]$ for $c \in \{0.1, 0.5, 0.8, 0.95\}$. We also add the plot of $ER_p(\Delta)$ in order to compare the error rate of the two decision rules. Since only finite values of $\Delta$ are considered in the figure we put $\gamma = 0$ and also choose $n_1 = n_2$. Finally, the ratio $\frac{n_1+n_2-2}{n_1+n_2-p-1}$ in the definition of $a$ is approximated by $\frac{1}{1-c}$. We observe that $ER_s(\Delta)$ lies very close to $ER_p(\Delta)$ for $c = 0.1$. However, the difference between two curves becomes considerable as $c$ growths, especially for $c = 0.95$ and larger values of $\Delta$.

3.2. **Finite-sample performance.** In this subsection we present the results of the simulation study. The aim is to investigate how good the asymptotic distribution of a linear combination of the discriminant function coefficients $\hat{\theta} = \mathbf{l}^\top\hat{\mathbf{a}}$ performs in the case of the finite dimension and of the finite sample size. For that reason we compare the asymptotic distribution of the standardized $\hat{\theta}$ as given in Theorem 4 to the corresponding exact distribution obtained as a kernel density approximation with the Eppanechnikov

kernel applied to the simulated data from the standardized exact distribution which are generated following the stochastic representation of Corollary 1: (i) first, $\xi_b, z_{0;b}, u_b$ are sampled independently from the corresponding univariate distributions provided in Corollary 1; (ii) second, $\hat{\theta}_b$ is computed by using (5) and standardized after that as in Theorem 4; (iii) finally, the previous two steps are repeated for $b = 1, \ldots, B$ times to obtain a sample of size $B$. It is noted that $B$ could be large to ensure a good performance of the kernel density estimator.

In the simulation study, we take $\mathbf{l} = \mathbf{1}_p$ ($p$-dimensional vector of ones). The elements of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are drawn from the uniform distribution on $[-1, 1]$ when $\gamma > 0$, while the first ten elements of $\boldsymbol{\mu}_1$ and the last ten elements of $\boldsymbol{\mu}_2$ are generated from the uniform distribution on $[-1, 1]$ and the rest of the components are taken to be zero when $\gamma = 0$. We also take $\boldsymbol{\Sigma}$ as a diagonal matrix, where every element is uniformly distributed on $(0, 1]$. The results are compared for several values of $c = \{0.1, 0.5, 0.8, 0.95\}$ and the corresponding values of $p, n_1, n_2$. Simulated data consist of $N = 10^5$ independent repetitions. In both cases $\gamma = 0$ and $\gamma > 0$ we plot two asymptotic density functions to investigate how robust are the obtained results to the choice of $\gamma$.

In Figures 3–4, we present the results in the case of equal and large sample sizes (data are drawn with $\gamma = 0$ in Figure 3 and with $\gamma > 0$ in Figure 4). We observe that the impact of the incorrect specification of $\gamma$ is not large. If $c$ increases, then the difference between the two asymptotic distributions becomes negligible. In contrast, larger differences between the asymptotic distributions and the finite-sample one are observed for $c = 0.8$ and $c = 0.95$ in all figures, although their sizes are relatively small even in such extreme case.

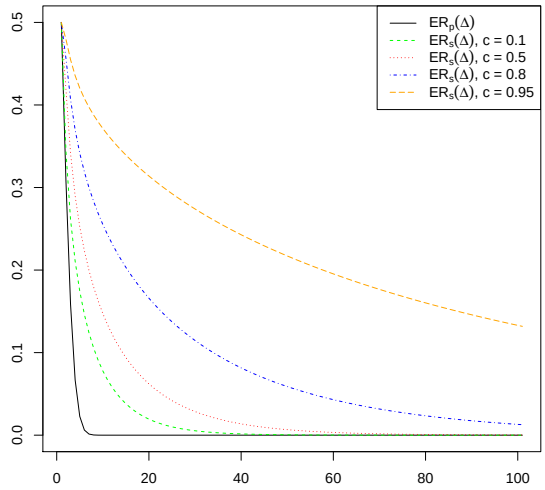
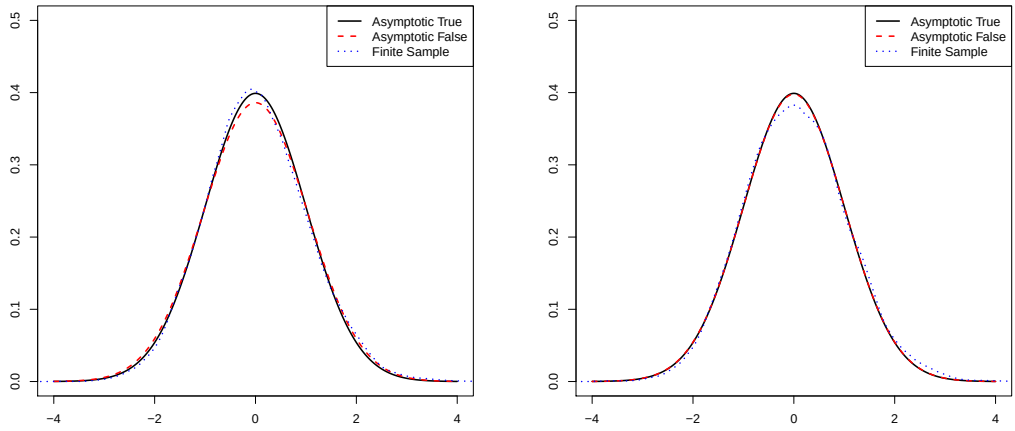



FIGURE 2. Error rates $ER_p(\Delta)$ and $ER_s(\Delta)$ as functions of $\Delta$ for $c \in \{0.1, 0.5, 0.8, 0.95\}$

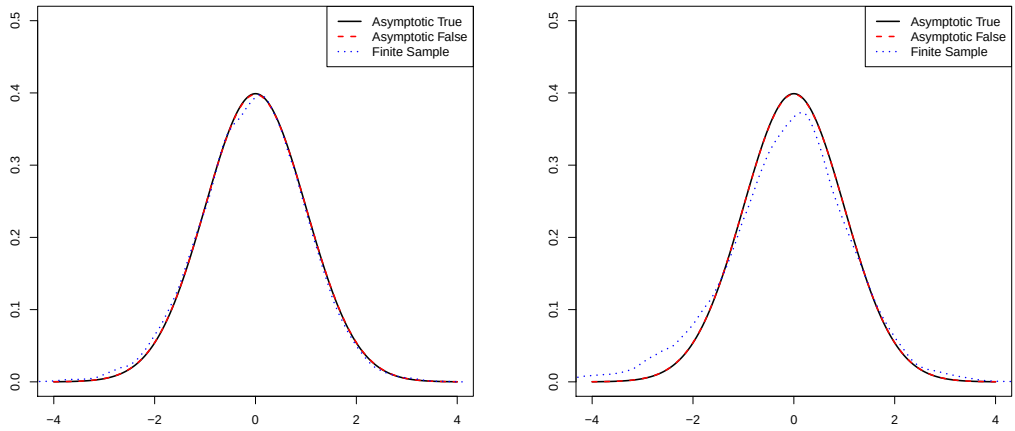
## ACKNOWLEDGEMENTS

This research was partly supported by the Swedish International Development Cooperation Agency (SIDA) through the TZ-Sweden Programme for Research, Higher Education and Institutional Advancement Stepan Mazur acknowledges the financial support from the internal research grants at Örebro University, and from the project "Models for macro and financial economics after the financial crisis" (Dnr: P18-0201) funded by Jan Wallander and Tom Hedelius Foundation.

The authors are grateful to the Editor and the Referee for their suggestions, which have improved the presentation of the paper.




*a*) $p = 50,\ n_1 = n_2 = 250$ *b*) $p = 250,\ n_1 = n_2 = 250$

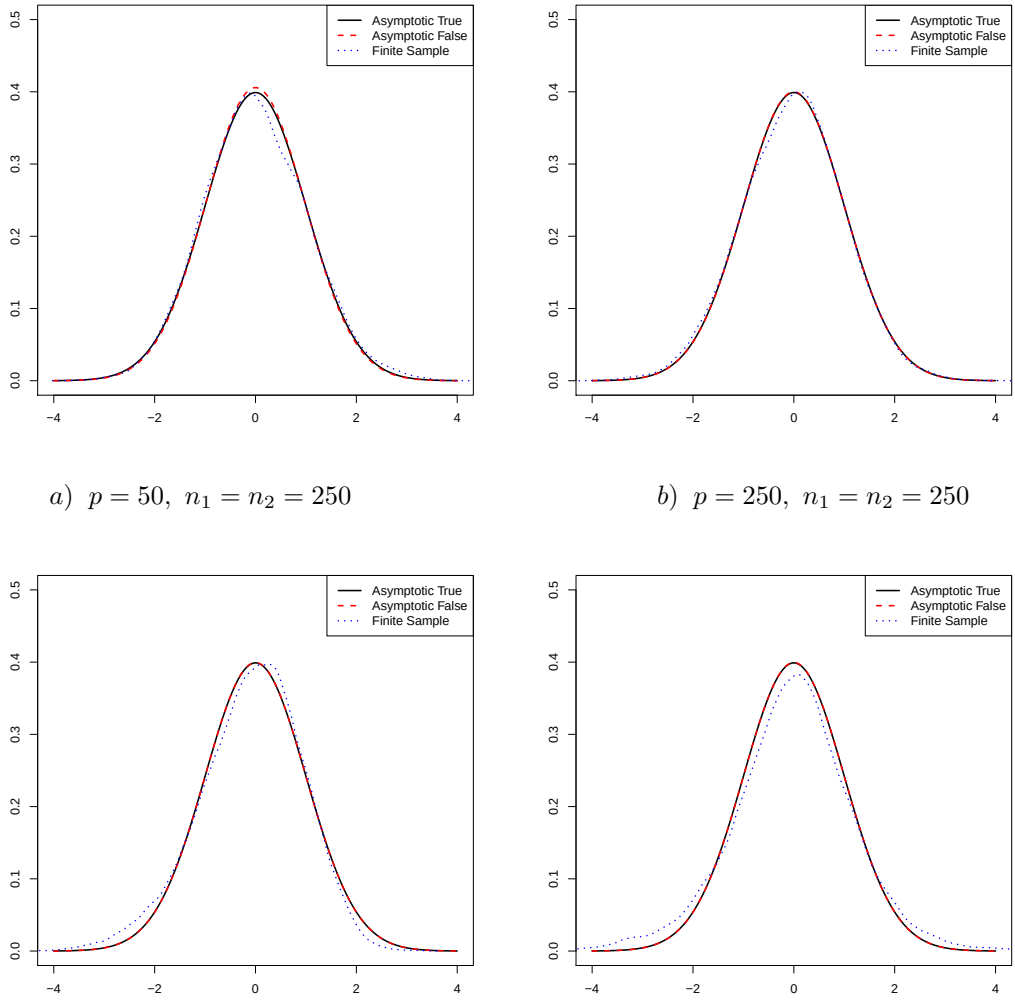



*c*) $p = 400,\ n_1 = n_2 = 250$ *d*) $p = 475,\ n_1 = n_2 = 250$

FIGURE 3. The kernel density estimator of the asymptotic distribution and standard normal distribution for $\hat{\theta}$ as given in Theorem 4 for $\gamma = 0$ and $c = \{0.1, 0.5, 0.8, 0.95\}$.

## REFERENCES

1. A. Agarwal, S. Negahban, M. J. Wainwright, *Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions*, Annals of Statistics, **40** (2012), no. 2, 1171–1197.
2. Z. Bai, D. Jiang, J.-F. Yao, S. Zheng, *Corrections to lrt on large-dimensional covariance matrix by rmt*, Annals of Statistics, **37** (2009), no. 6B, 3822–3840.
3. Z. Bai, J. W. Silverstein, *Spectral analysis of large dimensional random matrices*, New York, NY: Springer Science+ Business Media, LLC, 2010.




*a)* $p = 50,\ n_1 = n_2 = 250$ *b)* $p = 250,\ n_1 = n_2 = 250$

*c)* $p = 400,\ n_1 = n_2 = 250$ *d)* $p = 475,\ n_1 = n_2 = 250$

FIGURE 4. The kernel density estimator of the asymptotic distribution and standard normal distribution for $\hat{\theta}$ as given in Theorem 4 for $\gamma > 0$ and $c = \{0.1, 0.5, 0.8, 0.95\}$.

4. D. Bauder, R. Bodnar, T. Bodnar, W. Schmid, *Bayesian estimation of the efficient frontier*, Scandinavian Journal of Statistics, to appear (2019).
5. D. Bauder, T. Bodnar, S. Mazur, Y. Okhrin, *Bayesian inference for the tangent portfolio*, International Journal of Theoretical and Applied Finance, **21** (2018), no. 8.
6. P. J. Bickel, E. Levina, *Some theory for Fisher's linear discriminant function,'naive bayes', and some alternatives when there are many more variables than observations*, Bernoulli (2004), 989–1010.
7. T. Bodnar, H. Dette, N. Parolya, *Testing for independence of large dimensional vectors*, The Annals of Statistics, to appear (2019).
8. T. Bodnar, A. Gupta, N. Parolya, *Direct shrinkage estimation of large dimensional precision matrix*, Journal of Multivariate Analysis, **146** (2016), 223–236.

9. T. Bodnar, A. K. Gupta, N. Parolya, *On the strong convergence of the optimal linear shrinkage estimator for large dimensional covariance matrix*, Journal of Multivariate Analysis, **132** (2014), 215–228.
10. T. Bodnar, S. Mazur, Y. Okhrin, *Bayesian estimation of the global minimum variance portfolio*, European Journal of Operational Research, **256** (2017), 292–307.
11. T. Bodnar, S. Mazur, N. Parolya, *Central limit theorems for functionals of large sample covariance matrix and mean vector in matrix-variate location mixture of normal distributions*, Scandinavian Journal of Statistics, **46** (2019), 636–660.
12. T. Bodnar, Y. Okhrin, *Properties of the singular, inverse and generalized inverse partitioned wishart distributions*, Journal of Multivariate Analysis, **99** (2008), 2389–2405.
13. T. Bodnar, Y. Okhrin, *On the product of inverse wishart and normal distributions with applications to discriminant analysis and portfolio theory*, Scandinavian Journal of Statistics, **38** (2011), no. 2, 311–331.
14. T. Bodnar, M. Reiß, *Exact and asymptotic tests on a factor model in low and large dimensions with applications*, Journal of Multivariate Analysis, **150** (2016), 125–151.
15. T. Bodnar, W. Schmid, *A test for the weights of the global minimum variance portfolio in an elliptical model*, Metrika, **67** (2008), no. 2, 127–143.
16. T. Cai, W. Liu, *Adaptive thresholding for sparse covariance matrix estimation*, Journal of the American Statistical Association, **106** (2011), no. 494, 672–684.
17. T. Cai, W. Liu, *A direct estimation approach to sparse linear discriminant analysis*, Journal of the American Statistical Association, **106** (2011), no. 496, 1566–1577.
18. T. Cai, W. Liu, X. Luo, *A constrained $l_1$ minimization approach to sparse precision matrix estimation*, Journal of the American Statistical Association, **106** (2011), no. 494, 594–607.
19. T. Cai, T. Jiang, *Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices*, Annals of Statistics, **39** (2011), no. 3, 1496–1525.
20. S. X. Chen, L.-X. Zhang, P.-S. Zhong, *Tests for high-dimensional covariance matrices*, Journal of the American Statistical Association, **105** (2010), no. 490, 810–819.
21. A. DasGupta, *Asymptotic theory of statistics and probability*, Springer Science & Business Media, 2008.
22. J. Fan, Y. Fan, J. Lv, *High dimensional covariance matrix estimation using a factor model*, Journal of Econometrics, **147** (2008), no. 1, 186–197.
23. J. Fan, Y. Liao, M. Mincheva, *Large covariance estimation by thresholding principal orthogonal complements*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), **75** (2013), no. 4, 603–680.
24. Y. Fujikoshi, T. Seo, *Asymptotic aproximations for epmc's of the linear and the quadratic discriminant functions when the sample sizes and the dimension are large*, Random operators and stochastic equations, University of Toronto, 1997.
25. G. H. Givens, J. A. Hoeting, *Computational statistics*, John Wiley & Sons, 2012.
26. A. Gupta, D. Nagar, *Matrix Variate Distributions*, Chapman and Hall/CRC, Boca Raton, 2000.
27. A. Gupta, T. Varga, T. Bodnar, *Elliptically contoured models in statistics and portfolio theory*, second ed., Springer, 2013.
28. A. Gupta, T. Bodnar, *An exact test about the covariance matrix*, Journal of Multivariate Analysis, **125** (2014), 176–189.
29. T. Jiang, F. Yang, *Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions*, Annals of Statistics, **41** (2013), no. 4, 2029–2074.
30. R. A. Johnson, D. W. Wichern et al., *Applied multivariate statistical analysis*, Prentice hall Upper Saddle River, NJ, 2007.
31. I. M. Johnstone, *On the distribution of the largest eigenvalue in principal components analysis*, Annals of Statistics, **29** (2001), no. 2, 295–327.
32. O. Ledoit, M. Wolf, *Improved estimation of the covariance matrix of stock returns with an application to portfolio selection*, Journal of Empirical Finance, **10** (2003), no. 5, 603–621.
33. A. Mathai, S. B. Provost, *Quadratic forms in random variables*, Marcel Dekker, 1992.
34. R. J. Muirhead, *Aspects of multivariate statistical theory*, Wiley, New York, 1982.
35. I. Narsky, F. C. Porter, *Linear and quadratic discriminant analysis, logistic regression, and partial least squares regression*, Statistical Analysis Techniques in Particle Physics: Fits, Density Estimation and Supervised Learning (2013), 221–249.
36. A. C. Rencher, *Multivariate statistical inference and applications*, vol. 338, Wiley-Interscience, 1998.
37. A. C. Rencher, W. F. Christensen, *Methods of multivariate analysis*, John Wiley & Sons, 2012.

38. J. Shao, Y. Wang, X. Deng, S. Wang et al., *Sparse linear discriminant analysis by thresholding for high dimensional data*, The Annals of statistics, **39** (2011), no. 2, 1241–1265.
39. M. S. Srivastava, T. Kubokawa, *Comparison of discrimination methods for high dimensional data*, Journal of the Japan Statistical Society, **37** (2007), 123–134.
40. M. Tamatani, *Asymptotic theory for discriminant analysis in high dimension low sample size*, Memoirs of the Graduate School of Science and Engineering, Shimane University. Series B, Mathematics (2015), 15–26.
41. F. J. Wyman, D. M. Young, D. W. Turner, *A comparison of asymptotic error rate expansions for the sample linear discriminant function*, Pattern Recognition, **23** (1990), 775–783.

Department of Mathematics, Stockholm University, Roslagsvägen 101, SE-10691 Stockholm, Sweden
*E-mail address*: taras.bodnar@math.su.se

Unit of Statistics, School of Business, Örebro University, SE-70182 Örebro, Sweden
*E-mail address*: stepan.mazur@oru.se

Department of Mathematics, Dar es salaam University College of Education, Tanzania
*E-mail address*: edward.ngailo@liu.se

Delft Institute of Applied Mathematics, Delft University of Technology, 2628 CD Delft, Netherlands
*E-mail address*: n.parolya@tudelft.nl

Received 28.01.2019

## ДИСКРИМІНАНТНИЙ АНАЛІЗ У МАЛИХ ТА ВЕЛИКИХ РОЗМІРНОСТЯХ

Т. БОДНАР, С. МАЗУР, Е. НГАЙЛО, Н. ПАРОЛЯ

Анотація. Досліджуються стохастичні властивості лінійної дискримінантної функції за припущення нормальності шляхом порівняння двох груп з однаковою коваріаційною матрицею, але різними векторами середніх. Одержано стохастичне представлення коефіцієнтів дискримінантної функції, яке потім використовується для отримання їх асимптотичного розподілу при багатовимірному асимптотичному режимі. Досліджується ефективність класифікаційного аналізу на основі дискримінантної функції як у малих, так і у великих розмірностях. Установлено стохастичне представлення, яке дозволяє ефективно обчислити коефіцієнт похибки. Далі ми порівнюємо розрахований коефіцієнт похибки з оптимальним, отриманим за припущення, що коваріаційна матриця і два середні вектори є відомими. Нарешті, ми представляємо аналітичний вираз коефіцієнта похибок, одержаного у багатовимірному асимптотичному режимі. Скінченновимірні властивості отриманих теоретичних результатів оцінюються за допомогою обширного методу Монте-Карло.

## ДИСКРИМИНАНТНЫЙ АНАЛИЗ В МАЛЫХ И БОЛЬШИХ РАЗМЕРНОСТЯХ

Т. БОДНАР, С. МАЗУР, Э. НГАЙЛО, Н. ПАРОЛЯ

Аннотация. Исследуются стохастические свойства линейной дискриминантной функции при предположении нормальности путем сравнения двух групп с одинаковой ковариационной матрицей, но разными векторами средних. Получено стохастическое представление коэффициентов дискриминантной функции, которое затем используется для получения их асимптотического распределения при многомерном асимптотическом режиме. Исследуется эффективность классификационного анализа на основе дискриминантной функции как в малых, так и в больших размерностях. Установлено стохастическое представление, которое позволяет эффективно вычислить коэффициент погрешности. Далее мы сравниваем рассчитанный коэффициент погрешности с оптимальным, полученным при предположении, что ковариационная матрица и два средних вектора известны. Наконец, мы представляем аналитическое выражение коэффициента погрешности, полученного в многомерном асимптотическим режиме. Конечномерные свойства полученных теоретических результатов оцениваются с помощью обширного метода Монте-Карло.