WORKING PAPER

6/2017

Discriminant analysis in small and large dimensions

Taras Bodnar, Stepan Mazur, Edward Ngailo, Nestor Parolya
Statistics

Discriminant analysis in small and large dimensions

Taras Bodnar[a1], Stepan Mazur[b], Edward Ngailo[a] and Nestor Parolya[c]

[a] *Department of Mathematics, Stockholm University, Roslagsvägen 101, SE-10691 Stockholm, Sweden*

[b] *Unit of Statistics, School of Business, Örebro University, Fakultetsgatan 1, SE-70182 Örebro, Sweden*

[c] *Institute of Empirical Economics, Leibniz University of Hannover, D-30167 Hannover, Germany*

Abstract

In this article we study the distributional properties of the linear discriminant function under the assumption of the normality by comparing two groups with the same covariance matrix but different mean vectors. A stochastic representation of the discriminant function coefficient is derived which is then used to establish the asymptotic distribution under the high-dimensional asymptotic regime. Moreover, we investigate the classification analysis based on the discriminant function in both small and large dimensions. In the numerical study, a good finite-sample performance of the derived large-dimensional asymptotic distributions is documented.

JEL Classification: C12, C13, C44
ASM Classification: 62H10, 62E15, 62E20, 60F05, 60B20
Keywords: discriminant function, stochastic representation, large-dimensional asymptotics, random matrix theory, classification analysis

---

[1]Corresponding Author: Taras Bodnar. E-Mail: taras.bodnar@math.su.se. Tel: +46 8 164562. Fax: +46 8 612 6717.

1   Introduction

In the modern world of science and technology, high-dimensional data are present in various fields of science 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. The traditional multivariate analysis considers fixed or small sample dimensions, while sample sizes approaching to infinity. However, its methods cannot longer be used in the high-dimensional setting where the dimension is not treated as fixed but it is allowed to be comparable to the sample size.

The covariance matrix is one of the mostly used way to capture the dependence between variables. Although its application is restricted only to linear dependences and more sophisticate methods, like copula modeling, should be applied in the general case, the modeling dynamics in covariance matrix is still very popular subject in both statistics and econometrics. Recently, a number of papers have been published which deal with estimating the covariance matrix (see, e.g., Ledoit and Wolf (2003), Cai and Liu (2011a), Cai et al. (2011), Agarwal et al. (2012), Fan et al. (2008), Fan et al. (2013), Bodnar et al. (2014, 2016)) and testing its structure (see, e.g., Johnstone (2001), Bai et al. (2009), Chen et al. (2010), Cai and Jiang (2011), Jiang and Yang (2013), Gupta and Bodnar (2014)) in large dimension.

In many application, the covariance matrix is accompanied by the mean vector. For example, the product of the inverse covariance matrix and and the difference of the sample mean vectors is present in discriminant analysis 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 the sample mean vector (see Bodnar and Okhrin (2011)).

The discriminant analysis is a multivariate technique concerned with separating distinct sets of objects (or observations) (Johnson et al. (2007)). Its two main tasks are to distinguish distinct sets of observations and to allocate new observations to previously defined groups (Rencher and Christensen (2012)). The main methods of the discriminant analysis are the linear discriminant function and the quadratic discriminant function. The linear discriminant function is a generalization of Fisher's linear discriminant analysis, a method 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 matrix in the groups to be separated. Although the quadratic discriminant function can be applied when the latter assumption is violated, its application is more computational exhaustive, needs to estimate covariance matrices of each group, and

requires more observations than in the case of LDF (Narsky and Porter (2013)). 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 where 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., Muirhead (1982), Rencher and Christensen (2012)). However, these results cannot automatically be transferred when the number of variables is comparable to the sample size 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 high-dimensional asymptotics (see, e.g., Bai and Silverstein (2010)). Fujikoshi and Seo (1997) 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 Wyman et al. (1990). For the case samples non-equal sizes, they pointed out that the high-dimensional approximation is extremely accurate. However, Tamatani (2015) showed that Fisher's linear discriminant function performs poorly due to diverging spectra in the case of large-dimensional data and small sample size. Bickel and Levina (2004), Srivastava and Kubokawa (2007) investigated the asymptotic properties of the linear discriminant function in high dimension, while modifications of the linear discriminant function can be found in Cai and Liu (2011b), Shao et al. (2011). The asymptotic results for the discriminant function coefficients in matrix-variate skew models can be found in Bodnar et al. (2017b).

We contributed 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 accessed and the problem of the increasing dimensionality can be visualized in a simple way.

The rest of the paper is organized as follows. The finite-sample properties of the discriminant function are presented in Section 2, where, in particular we derive a stochastic representation for the discriminant function coefficients. In Section 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 Section 3.2. The finite-sample results are then use to derive the asymptotic distributions of the discriminant function coefficients and of the classification rule in Section 3, while finite sample performance of the asymptotic distribution is analysed in Section 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 $i = 1, .., n_1$ and $\mathbf{x}_i^{(2)} \sim \mathcal{N}_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ for $j = 1, .., n_2$ 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 $\mathbf{X}^{(2)} = \left(\mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_{n_2}^{(2)} \right)$. Then the sample estimators for the mean vectors and the covariance matrices constructed from each sample are given by

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

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

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

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

The following lemma (see, e.g., (Rencher and Christensen, 2012, Section 5.4.2)) presents the joint distribution of $\bar{\mathbf{x}}^{(1)}$, $\bar{\mathbf{x}}^{(2)}$ and $\mathbf{S}_{pl}$.

Lemma 1. *Let* $\mathbf{X}_1 \sim \mathcal{N}_{p,n_1} \left(\boldsymbol{\mu}_1 \mathbf{1}_{n_1}^T, \boldsymbol{\Sigma} \otimes \mathbf{I}_{n_1} \right)$ *and* $\mathbf{X}_2 \sim \mathcal{N}_{p,n_2} \left(\boldsymbol{\mu}_2 \mathbf{1}_{n_2}^T, \boldsymbol{\Sigma} \otimes \mathbf{I}_{n_2} \right)$ *for* $p < n_1 + n_2 - 2$. *Assume that* $\mathbf{X}_1$, $\mathbf{X}_2$ *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, implies 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 estimator of the population discriminant function coefficient vector expressed as

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

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 ordered.

Let $\mathbf{L}$ be a $k \times p$ matrix of constants such that $rank(\mathbf{L}) = k < p$. 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 jth position which is one, then we get the distribution of the jth coefficient in the discriminant function. If we choose $k = 1$ and $\mathbf{L} = (1, -1, 0, \ldots, 0)^T$, 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., Givens and Hoeting (2012)), in the theory of elliptical distributions (see, Gupta et al. (2013)) as well as in Bayesian statistics (cf., Bodnar et al. (2017a)). 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 $rank(\mathbf{L}) = k < p$. Then, under the assumption of Lemma 1 the stochastic representation of $\hat{\boldsymbol{\theta}} = \mathbf{L}\hat{\mathbf{a}}$ is given*

by

$$\hat{\boldsymbol{\theta}} \stackrel{d}{=} (n_1 + n_2 - 2)\xi^{-1}\left(\mathbf{L}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}} + \sqrt{\frac{\breve{\mathbf{x}}^T\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}}{n_1 + n_2 - p}}\left(\mathbf{L}\mathbf{R}_{\breve{\mathbf{x}}}\mathbf{L}^T\right)^{1/2}\mathbf{t}_0\right), \tag{5}$$

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

Proof. From Lemma 1.(c) and Theorem 3.4.1 of Gupta and Nagar (2000) we obtain that

$$\frac{1}{n_1 + n_2 - 2}\mathbf{S}_{pl}^{-1} \sim \mathcal{IW}_p(n_1 + n_2 + p - 1, \boldsymbol{\Sigma}^{-1}). \tag{6}$$

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

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

Applying Theorem 3.2.12 of Muirhead (1982) we obtain that

$$\xi^* = (n_1 + n_2 - 2)\frac{\breve{\mathbf{x}}^{*T}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}^*}{\breve{\mathbf{x}}^{*T}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*} \sim \chi^2_{n_1+n_2-p-1} \tag{7}$$

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

$$\xi = (n_1 + n_2 - 2)\frac{\breve{\mathbf{x}}^T\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}}{\breve{\mathbf{x}}^T\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}} \sim \chi^2_{n_1+n_2-p-1} \tag{8}$$

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

Using Theorem 3 of Bodnar and Okhrin (2008) we get that $\breve{\mathbf{x}}^{*T}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*$ is independent of $\mathbf{L}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*/\breve{\mathbf{x}}^{*T}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*$ for given $\breve{\mathbf{x}}^*$. Therefore, ξ^* is independent of $\breve{\mathbf{x}}^{*T}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}^* \cdot \mathbf{L}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*/\breve{\mathbf{x}}^{*T}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*$ and, respectively, ξ is independent of $\breve{\mathbf{x}}^T\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}} \cdot \mathbf{L}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}/\breve{\mathbf{x}}^T\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}$. Furthermore, from the proof of Theorem 1 of Bodnar and Schmid (2008) it holds that

$$\breve{\mathbf{x}}^{*T}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}^* \frac{\mathbf{L}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*}{\breve{\mathbf{x}}^{*T}\mathbf{S}_{pl}^{-1}\breve{\mathbf{x}}^*} \sim t_k\left(n_1 + n_2 - p; \mathbf{L}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}^*, \frac{\breve{\mathbf{x}}^{*T}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}^*}{n_1 + n_2 - p}\mathbf{L}\mathbf{R}_{\breve{\mathbf{x}}^*}\mathbf{L}^T\right) \tag{9}$$

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

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

$$\hat{\boldsymbol{\theta}} \stackrel{d}{=} (n_1 + n_2 - 2)\xi^{-1}\left(\mathbf{L}\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}} + \sqrt{\frac{\breve{\mathbf{x}}^T\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}}{n_1 + n_2 - p}}\left(\mathbf{L}\mathbf{R}_{\breve{\mathbf{x}}}\mathbf{L}^T\right)^{1/2}\mathbf{t}_0\right), \tag{10}$$

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

In the next corollary we consider the special case when $k = 1$, that is, when $\mathbf{L} = \mathbf{l}^T$ 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 $\hat{\theta} = \mathbf{l}^T\hat{\mathbf{a}}$ is given by*

$$\hat{\theta} \overset{d}{=} (n_1 + n_2 - 2)\xi^{-1}\left(\mathbf{l}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \sqrt{\left(\lambda + \frac{\lambda(p-1)}{n_1 + n_2 - p}u\right)}\mathbf{l}^T\boldsymbol{\Sigma}^{-1}\mathbf{l}z_0\right), \quad (11)$$

where $\xi \sim \chi^2_{n_1+n_2-p-1}$, $z_0 \sim \mathcal{N}(0,1)$, $u \sim \mathcal{F}\left(p-1, n_1+n_2-p, (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{R}_\mathbf{l}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/\lambda\right)$ (non-central $\mathcal{F}$-distribution with $p-1$ and n_1+n_2-p degrees of freedom and non-centrality parameter $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{R}_\mathbf{l}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/\lambda$) with $\mathbf{R}_\mathbf{l} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{l}\mathbf{l}^T\boldsymbol{\Sigma}^{-1}/\mathbf{l}^T\boldsymbol{\Sigma}^{-1}\mathbf{l}$; ξ, z_0 and u are mutually independently distributed.

Proof. From Theorem 1 we get that

$$\hat{\theta} \overset{d}{=} (n_1 + n_2 - 2)\xi^{-1}\left(\mathbf{l}^T\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}} + t_0\sqrt{\frac{\breve{\mathbf{x}}^T\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}}{n_1 + n_2 - p} \cdot \mathbf{l}^T\mathbf{R}_{\breve{\mathbf{x}}}\mathbf{l}}\right) \quad (12)$$

$$= (n_1 + n_2 - 2)\xi^{-1}\left(\mathbf{l}^T\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}} + \frac{t_0}{\sqrt{n_1 + n_2 - p}}\sqrt{\mathbf{l}^T\boldsymbol{\Sigma}^{-1}\mathbf{l}}\sqrt{\breve{\mathbf{x}}^T\mathbf{R}_\mathbf{l}\breve{\mathbf{x}}}\right), \quad (13)$$

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

Because $\breve{\mathbf{x}} \sim \mathcal{N}_p\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2, \lambda\boldsymbol{\Sigma}\right)$, $\mathbf{R}_\mathbf{l}\boldsymbol{\Sigma}\mathbf{R}_\mathbf{l} = \mathbf{R}_\mathbf{l}$, and $\text{tr}\left[\mathbf{R}_\mathbf{l}\boldsymbol{\Sigma}\right] = p - 1$, the application of Corollary 5.1.3a of Mathai and Provost (1992) leads to

$$\zeta = \lambda^{-1}\breve{\mathbf{x}}^T\mathbf{R}_\mathbf{l}\breve{\mathbf{x}} \sim \chi^2_{p-1}\left(\delta^2\right) \quad (14)$$

where $\delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{R}_\mathbf{l}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/\lambda$. Moreover, since $\mathbf{R}_\mathbf{l}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\mathbf{l} = \mathbf{0}$, the application of Theorem 5.5.1 of Mathai and Provost (1992) proves that $\mathbf{l}^T\boldsymbol{\Sigma}^{-1}\breve{\mathbf{x}}$ and ζ 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

$$t_0 \overset{d}{=} z_0\sqrt{\frac{n_1 + n_2 - p}{w}}, \quad (15)$$

7

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,

$$\mathbf{l}^T\mathbf{\Sigma}^{-1}\breve{\mathbf{x}} + t_0\sqrt{\frac{\lambda\zeta \cdot \mathbf{l}^T\mathbf{\Sigma}^{-1}\mathbf{l}}{n_1+n_2-p}}\bigg|\zeta, w \quad \sim \quad \mathcal{N}\left(\mathbf{l}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}, \lambda\mathbf{l}^T\mathbf{\Sigma}^{-1}\mathbf{l}\left(1+\frac{\zeta}{w}\right)\right) \tag{16}$$

$$= \quad \mathcal{N}\left(\mathbf{l}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}, \lambda\mathbf{l}^T\mathbf{\Sigma}^{-1}\mathbf{l}\left(1+\frac{p-1}{n_1+n_2-p}u\right)\right) \tag{17}$$

where

$$u = \frac{\zeta/(p-1)}{w/(n_1+n_2-p)} \sim \mathcal{F}\left(p-1, n_1+n_2-p, (\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\mathbf{R}_\mathbf{l}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)/\lambda\right). \tag{18}$$

Putting all above together we get the statement of the corollary. $\qquad \square$

2.2 Test for the population discriminant function coefficients

One of the most important questions when the discriminant analysis is performed is to decide which coefficients are the most influential in the decision. Several methods exist in the literature with the following three approaches to be the most popular (c.f., (Rencher and Christensen, 2012, Section 5.5)): (i) standardized coefficients; (ii) partial F-values; (iii) correlations between variables and discriminant function. (Rencher, 1998, Theorem 5.7A) argued that each of this three methods has several drawbacks. For instance, the correlations between variables and 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 the other variables.

In this section, we propose an alternative approach based on the statistical hypothesis test. Namely, exact statistical tests will be derived on the null hypothesis that two population discriminant function coefficients are equal (two-sided test) as well as on the alternative hypothesis that a coefficient in the discriminant function is larger than another one (one-sided test). The testing hypothesis for two coefficients in the population discriminant function are given by

$$H_0: \ a_i = a_j \quad \text{against} \quad H_1: \ a_i \neq a_j \quad \text{for } i \neq j\,, \tag{19}$$

while in the case of one-sided test we check if

$$H_0: \ a_i \leq a_j \quad \text{against} \quad H_1: \ a_i > a_j \quad \text{for } i \neq j\,. \tag{20}$$

In both cases the following test statistic will be used

$$T = \sqrt{n_1 + n_2 - p - 1} \frac{\mathbf{l}'\mathbf{S}_{pl}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})}{\sqrt{\mathbf{l}'\mathbf{S}_{pl}^{-1}\mathbf{l}}\sqrt{(n_1 + n_2 - 2)(\frac{1}{n_1} + \frac{1}{n_2}) + (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})'\hat{\mathbf{R}}_{\mathbf{l}}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})}} \tag{21}$$

with

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

The distribution of T follows from (Bodnar and Okhrin, 2011, 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

$$f_T(x) = \frac{n_1 + n_2 - p}{\lambda(p-1)} \int_0^\infty f_{t_{n_1+n_2-p-1,\delta_1(y)}}(x) f_{\mathcal{F}_{p-1,n_1+n_2-p,s/\lambda}}\left(\frac{n_1 + n_2 - p}{\lambda(p-1)}y\right) dy \tag{22}$$

with $\delta_1(y) = \eta/\sqrt{\lambda + y}$, $\eta = \frac{\mathbf{l}'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{\sqrt{\mathbf{l}'\boldsymbol{\Sigma}^{-1}\mathbf{l}}}$, and $s = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\mathbf{R}_{\mathbf{l}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$; the symbol $f_G(.)$ 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 $(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})'\hat{\mathbf{R}}_{\mathbf{l}}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})$.

Theorem 2 shows that the test statistics T has a standard distribution under the null hypothesis. As a result, the suggested test will reject the null hypothesis of the two-sided test (19) 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 (20). In this case the maximal probability of the type I error has to be control. For that reason, we first calculate the probability of rejection of the null hypothesis for all possible parameter values and after that we calculate it maximum for the parameters which corresponds to the null hypothesis in (20). Since the distribution of T depends on $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ only via η and s (see, Theorem 2), the task of finding the maximum is significantly simplified. Let $F_G(.)$ denotes the distribution function of the distribution G. For any constant q, we

get

$$\mathbb{P}(T > q) = \int_{q}^{+\infty} f_T(x)dx$$

$$= \int_{q}^{+\infty} \frac{n_1 + n_2 - p}{\lambda(p-1)} \int_{0}^{\infty} f_{t_{n_1+n_2-p-1,\delta_1(y)}}(x) f_{\mathcal{F}_{p-1,n_1+n_2-p,s/\lambda}}(\frac{n_1 + n_2 - p}{\lambda(p-1)}y)dydx$$

$$= \frac{n_1 + n_2 - p}{\lambda(p-1)} \int_{0}^{\infty} f_{\mathcal{F}_{p-1,n_1+n_2-p,s/\lambda}}(\frac{n_1 + n_2 - p}{\lambda(p-1)}y) \int_{q}^{+\infty} f_{t_{n_1+n_2-p-1,\delta_1(y)}}(x)dxdy$$

$$= \frac{n_1 + n_2 - p}{\lambda(p-1)} \int_{0}^{\infty} (1 - F_{t_{n_1+n_2-p-1,\delta_1(y)}}(q)) f_{\mathcal{F}_{p-1,n_1+n_2-p,s/\lambda}}(\frac{n_1 + n_2 - p}{\lambda(p-1)}y)dy$$

$$\leq \frac{n_1 + n_2 - p}{\lambda(p-1)} \int_{0}^{\infty} (1 - F_{t_{n_1+n_2-p-1,0}}(q)) f_{\mathcal{F}_{p-1,n_1+n_2-p,s/\lambda}}(\frac{n_1 + n_2 - p}{\lambda(p-1)}y)dy$$

$$= (1 - F_{t_{n_1+n_2-p-1,0}}(q)).$$

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 (20) 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 considered two groups. Assuming that no prior information is available about the classification result, i.e. the prior probability of each group is $1/2$, the decision is based on the optimal rule is to assign the observation vector to the first group as soon as the following inequality holds (c.f., (Rencher, 1998, Section 6.2))

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1} \mathbf{x} > \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \tag{23}$$

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. Rencher (1998) presented the expression of the error rate expressed as

$$ER_p(\Delta) = \frac{1}{2}\mathbb{P}(\text{classify to the first group} \mid \text{second group is true})$$

$$+ \frac{1}{2}\mathbb{P}(\text{classify to the second group} \mid \text{first group is true})$$

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

where $\Phi(.)$ 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

$$(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^\top \mathbf{S}_{pl}^{-1}\mathbf{x} > \frac{1}{2}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^\top \mathbf{S}_{pl}^{-1}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}) \tag{24}$$

instead. Next, we derive the error rate of the decision rule (24). Let

$$
\begin{aligned}
\hat{d} &= (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^\top \mathbf{S}_{pl}^{-1}\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^\top \mathbf{S}_{pl}^{-1}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)}) \\
&= (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^\top \mathbf{S}_{pl}^{-1}\left(\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)})\right).
\end{aligned}
\tag{25}
$$

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*

$$
\hat{d} \overset{d}{=} \frac{n_1 + n_2 - 2}{\xi}\left((-1)^{i-1}\frac{\lambda n_i - 2}{2\lambda n_i}\left(\lambda\xi_2 + (\Delta + \sqrt{\lambda}w_0)^2\right) + \frac{(-1)^{i-1}}{\lambda n_i}\left(\Delta^2 + \sqrt{\lambda}\Delta w_0\right)\right.
$$
$$
\left. + \sqrt{\left(1 + \frac{1}{n_1 + n_2} + \frac{p-1}{n_1 + n_2 - p}u\right)}\sqrt{\lambda\xi_2 + (\Delta + \sqrt{\lambda}w_0)^2}z_0\right) \quad \text{for } i = 1, 2, \tag{26}
$$

where $u|\xi_1, \xi_2, w_0 \sim \mathcal{F}(p-1, n_1 + n_2 - p, (n_1 + n_2)^{-1}\xi_1)$ with $\xi_1|\xi_2, w_0 \sim \chi^2_{p-1,\delta^2_{\xi_2,w_0}}$ and $\delta^2_{\xi_2,w_0} = \frac{n_1 n_2}{n_i^2}\frac{\Delta^2\xi_2}{\lambda\xi_2 + (\Delta+\sqrt{\lambda}w_0)^2}$, $z_0, w_0 \sim \mathcal{N}(0,1)$, $\xi \sim \chi^2_{n_1+n_2-p-1}$, $\xi_2 = \chi^2_{p-1}$; ξ, z_0 are independent of u, ξ_1, ξ_2, w_0 where ξ_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 defined by

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

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

Following the proof of Corollary 1, we get

$$
\begin{aligned}
d_0 &\overset{d}{=} (n_1 + n_2 - 2)\xi^{-1}\left((\bar{\mathbf{x}}_0^{(1)} - \bar{\mathbf{x}}_0^{(2)})^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_i - \frac{1}{2}(\bar{\mathbf{x}}_0^{(1)} + \bar{\mathbf{x}}_0^{(2)})\right)\right. \\
&\quad\left. + \sqrt{\left(1 + \frac{(p-1)}{n_1 + n_2 - p}u\right)(\bar{\mathbf{x}}_0^{(1)} - \bar{\mathbf{x}}_0^{(2)})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}_0^{(1)} - \bar{\mathbf{x}}_0^{(2)})}z_0\right),
\end{aligned}
$$

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

In using that

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

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

$$\hat{d} \;\overset{d}{=}\; \frac{n_1 + n_2 - 2}{\xi}\left(\frac{(-1)^{i-1}}{2}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) - (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^T\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}_i\right)\right.$$
$$\left. + \sqrt{\left(1 + \frac{p-1}{n_1 + n_2 - p}u\right)(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})}z_0\right),$$

where $u|\bar{\mathbf{x}}^{(1)}, \bar{\mathbf{x}}^{(2)} \sim \mathcal{F}\left(p - 1, n_1 + n_2 - p, \left(\bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}_i\right)^T \mathbf{R}_{\mathbf{x}}\left(\bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}_i\right)\right)$ with $\mathbf{R}_{\mathbf{x}} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^T\boldsymbol{\Sigma}^{-1}/(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})$, $z_0 \sim \mathcal{N}(0, 1)$, and $\xi \sim \chi^2_{n_1+n_2-p-1}$; ξ, 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,

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

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 Mathai and Provost (1992) shows that given $(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})$ the random variables $(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^T\boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}_i)$ and $(\bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}_i)\mathbf{R}_{\mathbf{x}}(\bar{\mathbf{x}}^{(i)} - \boldsymbol{\mu}_i)$ are independently distributed with

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

and, by using Corollary 5.1.3a of Mathai and Provost (1992),

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

with

$$\delta_{\mathbf{x}}^2 = \frac{n_1 + n_2}{\lambda^2 n_i^2}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)} - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))^T \mathbf{R}_{\mathbf{x}}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)} - (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))$$

$$= \frac{n_1 + n_2}{\lambda^2 n_i^2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{R}_{\mathbf{x}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

$$= \frac{n_1 + n_2}{\lambda^2 n_i^2}\frac{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^T \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^T \mathbf{R}_{\boldsymbol{\mu}}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})$$

where we use that $(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^T \mathbf{R}_{\mathbf{x}} = \mathbf{0}$ and $\mathbf{R}_{\boldsymbol{\mu}} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}/(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

As a result, we get

$$\hat{d} \stackrel{d}{=} \frac{n_1 + n_2 - 2}{\xi}\left((-1)^{i-1}\frac{\lambda n_i - 2}{2\lambda n_i}\Delta_{\mathbf{x}}^2 + \frac{(-1)^{i-1}}{\lambda n_i}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})\right.$$

$$\left. + \sqrt{\left(1 + \frac{1}{n_1 + n_2} + \frac{p-1}{n_1 + n_2 - p}u\right)\Delta_{\mathbf{x}}}z_0\right),$$

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

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

$$\Delta_{\mathbf{x}}^2 = (\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})^T \mathbf{R}_{\boldsymbol{\mu}}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)}) + \frac{\left((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)})\right)^2}{\Delta^2},$$

where both summands are independent following Theorem 5.5.1 in Mathai and Provost (1992). The application of Corollary 5.1.3a in Mathai and Provost (1992) leads to

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

and

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

From the last statement we get the stochastic representation of $\hat{d}$ expressed as

$$\hat{d} \stackrel{d}{=} \frac{n_1 + n_2 - 2}{\xi}\left((-1)^{i-1}\frac{\lambda n_i - 2}{2\lambda n_i}\left(\lambda\xi_2 + (\Delta + \sqrt{\lambda}w_0)^2\right) + \frac{(-1)^{i-1}}{\lambda n_i}\left(\Delta^2 + \sqrt{\lambda}\Delta w_0\right)\right.$$

$$\left. + \sqrt{\left(1 + \frac{1}{n_1 + n_2} + \frac{p-1}{n_1 + n_2 - p}u\right)}\sqrt{\lambda\xi_2 + (\Delta + \sqrt{\lambda}w_0)^2}z_0\right),$$

where $u|\xi_1, \xi_2, w_0 \sim \mathcal{F}(p-1, n_1 + n_2 - p, (n_1 + n_2)^{-1}\xi_1)$ with $\xi_1|\xi_2, w_0 \sim \chi_{p-1,\delta_{\xi_2,w_0}^2}^2$ and $\delta_{\xi_2,w_0}^2 = \frac{n_1+n_2}{\lambda^2 n_i^2}\frac{\Delta^2}{\lambda\xi_2 + (\Delta + \sqrt{\lambda}w_0)^2}\lambda\xi_2$, $z_0, w_0 \sim \mathcal{N}(0, 1)$, $\xi \sim \chi_{n_1+n_2-p-1}^2$, $\xi_2 \sim \chi_{p-1}^2$; ξ, z_0 are

independent of u, ξ_1, ξ_2, w_0 where ξ_2 and w_0 are independent as well. $\quad\square$

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 Δ. As a result, the the error rate based on the decision rule (24) is a function of Δ only and it is calculated by

$$
\begin{aligned}
ER_s(\Delta) \quad = \quad & \frac{1}{2}\mathbb{P}(\hat{d} > 0|\text{ second group is true}) \qquad\qquad (27)\\
+ \quad & \frac{1}{2}\mathbb{P}(\hat{d} \leq 0|\text{ first group is true})\,.
\end{aligned}
$$

The two probabilities in (27) can easily be approximated for all Δ, p, n_1, and n_2 with high precision by applying the results of Theorem 3 via the following simulation study

(i) Fix Δ 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 $\xi_{1,b} \sim \chi_{p-1, \delta^2_{\xi_2,w_0}}$ with $\delta^2_{\xi_{2,b},w_{0,b}} = \frac{n_1 n_2}{n_i^2} \frac{\Delta^2 \xi_{2;b}}{\lambda \xi_{2;b}+(\Delta+\sqrt{\lambda}w_{0;b})^2}$.

(iv) Generate $u \sim \mathcal{F}\left(p - 1, n_1 + n_2 - p, (n_1 + n_2)^{-1}\xi_{1,b}\right)$.

(v) Calculate $\hat{d}_b^{(i)}$ following the stochastic representation (26) of Theorem 3.

(vi) Repeat steps (ii)-(v) for $b = 1, ..., B$ leading to the sample $\hat{d}_1^{(i)}, ..., \hat{d}_B^{(i)}$.

The procedure has to be perform for both values of $i = 1, 2$ where for $i = 1$ the relative numbers of events $\{\hat{d} > 0\}$ will approximate the first summand in (27) while for $i = 2$ the relative numbers of events $\{\hat{d} \leq 0\}$ will approximate the second summand in (27).

It is important to note that the difference between the error rates calculated for the two decision rules (23) ad (24) could be very large as shown in Figure 1 where $ER_p(\Delta)$ and $ER_s(\Delta)$ calculated for several $n_1 = n_2 \in \{50, 100, 150, 250\}$ 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 several values of n. However, this statement does not hold any longer when p becomes comparable to both samples sizes n_1 and n_2 as document 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, that is, when the dimension
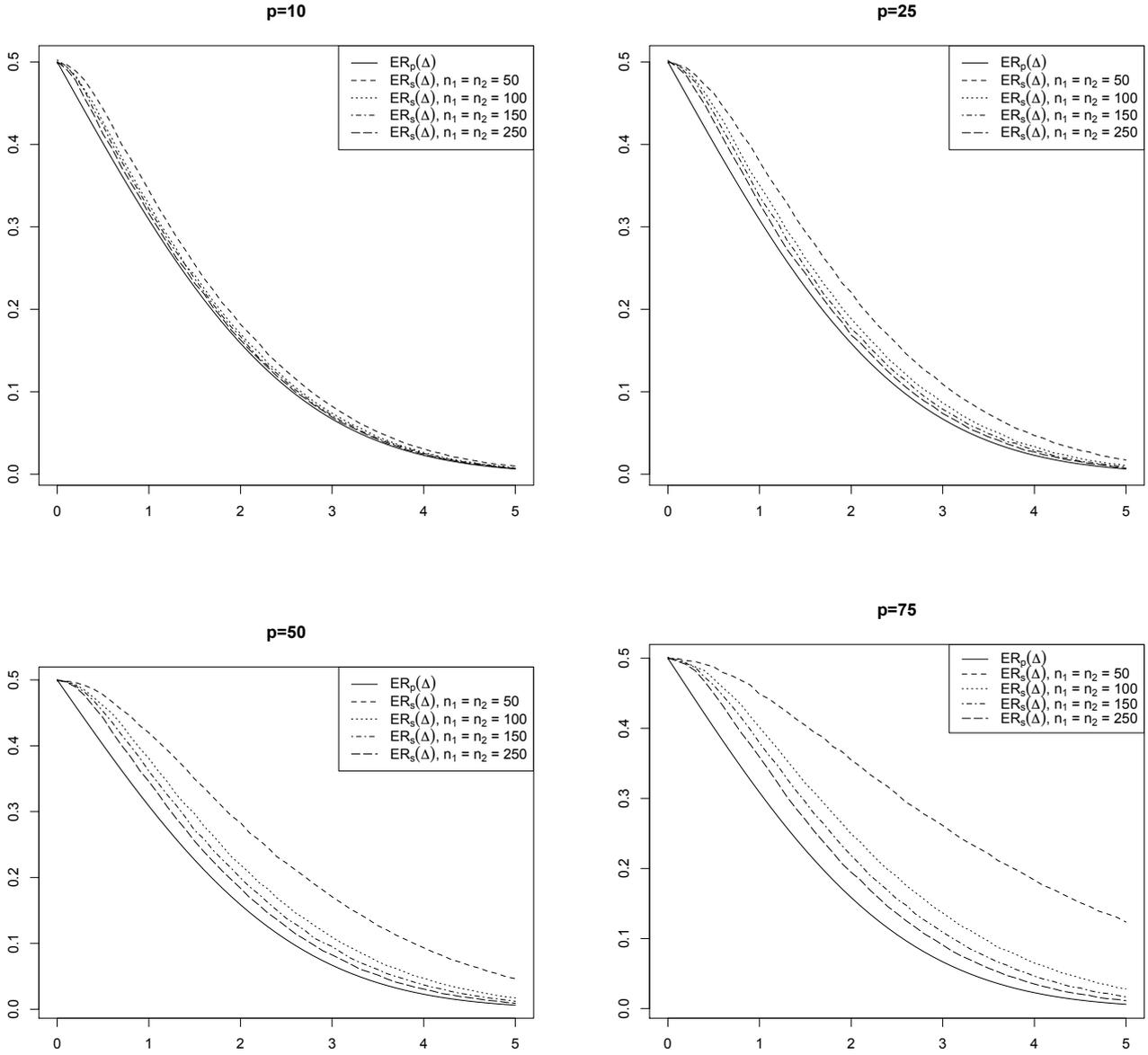
Figure 1: Error rates $ER_p(\Delta)$ and $ER_s(\Delta)$ as functions of Δ for $p \in \{10, 25, 50, 75\}$ and $ER_s(\Delta)$.

increases together with the sample size and they both tend to infinity. More precisely, we assume that $p/(n_1 + n_2) \to c \in [0, 1)$ as $n_1 + n_2 \to \infty$.

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

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

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

It is remarkable that, no assumption on the eigenvalues of the covariance matrix $\boldsymbol{\Sigma}$, like they are uniformly bounded on p, is imposed. 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 γ 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 $p^{-\gamma}\mathbf{l}^T\boldsymbol{\Sigma}^{-1}\mathbf{l} < \infty$ is uniformly on p, $\gamma \geq 0$. Then, under the conditions of Theorem 1, the asymptotic distribution of $\hat{\theta} = \mathbf{l}^T\hat{\mathbf{a}}$ is given by*

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

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

$$\sigma_\gamma^2 = \frac{1}{(1-c)^3} \bigg(\left(\mathbf{l}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right)^2 + \mathbf{l}^T\boldsymbol{\Sigma}^{-1}\mathbf{l}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \qquad (28)$$

$$+ \quad \lambda(n_1 + n_2)\mathbf{l}^T\boldsymbol{\Sigma}^{-1}\mathbf{l}\mathbb{1}_{\{0\}}(\gamma) \bigg)$$

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

Proof. Using the stochastic representation (11) of Corollary 1, we get

$$\sqrt{n_1 + n_2}\sigma^{-1} \left(\hat{\theta} - \frac{1}{1-c}\mathbf{l}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \right)$$

$$\stackrel{d}{=} \quad \sqrt{n_1 + n_2} \left((n_1 + n_2 - 2)\xi^{-1} - \frac{1}{1-c} \right) \frac{p^{-\gamma}\mathbf{l}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{p^{-\gamma}\sigma_\gamma}$$

$$+ \quad \sqrt{\lambda(n_1 + n_2)}\frac{n_1 + n_2 - 2}{\xi}\sqrt{\left(p^{-\gamma} + p^{-\gamma}\frac{p-1}{n_1 + n_2 - p}u\right)}\frac{\sqrt{p^{-\gamma}\mathbf{l}^T\boldsymbol{\Sigma}^{-1}\mathbf{l}}}{p^{-\gamma}\sigma_\gamma}z_0,$$

where $\xi \sim \chi^2_{n_1+n_2-p-1}$, $z_0 \sim \mathcal{N}(0, 1)$, $u \sim \mathcal{F}\left(p - 1, n_1 + n_2 - p, (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{R}_\mathbf{l}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)/\lambda\right)$ with $\mathbf{R}_\mathbf{l} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{l}\mathbf{l}^T\boldsymbol{\Sigma}^{-1}/\mathbf{l}^T\boldsymbol{\Sigma}^{-1}\mathbf{l}$; ξ, 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,

$$\sqrt{n_1+n_2}\left((n_1+n_2-2)\xi^{-1}-\frac{1}{1-c}\right) = \frac{\sqrt{n_1+n_2}}{\sqrt{n_1+n_2-p-1}}\frac{n_1+n_2-p-1}{\xi}\frac{1}{1-c}$$

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

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, (Bodnar and Reiß, 2016, Lemma 3))

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

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

$$\begin{aligned}
\sigma_\gamma^2 &= \frac{1}{(1-c)^3}\left(2\left(\mathbf{l}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right)^2 + \mathbf{l}^T\boldsymbol{\Sigma}^{-1}\mathbf{l}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\mathbf{R_l}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right.\\
&\quad + \left.\lambda(n_1+n_2)\mathbf{l}^T\boldsymbol{\Sigma}^{-1}\mathbf{l}\mathbb{1}_{\{0\}}(\gamma)\right)\\
&= \frac{1}{(1-c)^3}\left(\left(\mathbf{l}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right)^2 + \mathbf{l}^T\boldsymbol{\Sigma}^{-1}\mathbf{l}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\right.\\
&\quad + \left.\lambda(n_1+n_2)\mathbf{l}^T\boldsymbol{\Sigma}^{-1}\mathbf{l}\mathbb{1}_{\{0\}}(\gamma)\right)
\end{aligned}$$

$\square$

The results of Theorem 4 show that the quantity γ is present only in the asymptotic variance σ_γ^2. 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 $\gamma = 0$ we need (A2) in order to keep the variance bounded. We further investigate this point via simulations in Section 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 decision rule (23) remains the same independently of p and it is always equal to

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

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 (24) instead of (23). 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 $p/(n_1 + n_2) \to c \in [0, 1)$ as $n_1 + n_2 \to \infty$. Then, under the conditions of Theorem 1, it holds that*

$$p^{\min(\gamma,1)/2}\left(\frac{\hat{d}}{p^\gamma} - \frac{n_1 + n_2 - 2}{n_1 + n_2 - p - 1}\frac{(-1)^{i-1}}{2}p^{-\gamma}\Delta^2\right)$$

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

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

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

$$p^{\min(\gamma,1)/2}\left(\frac{\hat{d}}{p^\gamma} - \frac{n_1 + n_2 - 2}{n_1 + n_2 - p - 1}\frac{(-1)^{i-1}}{2}p^{-\gamma}\Delta^2\right)$$

$$\stackrel{d}{=} p^{\min(\gamma,1)/2-1/2}\frac{\sqrt{p}}{\sqrt{n_1 + n_2 - p - 1}}\frac{n_1 + n_2 - 2}{\xi}\sqrt{n_1 + n_2 - p - 1}\left(1 - \frac{\xi}{n_1 + n_2 - p - 1}\right)$$

$$\times \quad \frac{(-1)^{i-1}}{2}p^{-\gamma}\Delta^2 + \frac{n_1 + n_2 - 2}{\xi}\left((-1)^{i-1}\frac{\lambda n_i - 2}{2\lambda n_i}\right.$$

$$\times \quad \left(p^{\min(\gamma,1)/2-\gamma}\lambda\xi_2 + 2p^{\min(\gamma,1)/2-\gamma/2}\sqrt{p^{-\gamma}\Delta^2}\sqrt{\lambda}w_0 + p^{\min(\gamma,1)/2-\gamma}\lambda w_0^2\right)$$

$$+ \quad \frac{(-1)^{i-1}}{\lambda n_i}p^{\min(\gamma,1)/2-\gamma/2}\sqrt{p^{-\gamma}\Delta^2}\sqrt{\lambda}w_0\right)$$

$$+ \quad \frac{n_1 + n_2 - 2}{\xi}\left(\sqrt{\left(1 + \frac{1}{n_1 + n_2} + \frac{p - 1}{n_1 + n_2 - p}u\right)}\right.$$

$$\times \quad \left.\sqrt{p^{\min(\gamma,1)-2\gamma}\lambda\xi_2 + (p^{\min(\gamma,1)/2-\gamma/2}\sqrt{p^{-\gamma}\Delta^2} + p^{\min(\gamma,1)/2-\gamma}\sqrt{\lambda}w_0)^2}z_0\right)$$

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

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

where the last line follows from Lemma 3 in Bodnar and Reiß (2016) and Slutsky Theorem (see, (DasGupta, 2008, Theorem 1.5)). $\qquad\square$

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 derived the approximate error rate for the decision (24). Let $a = \frac{1}{1-c}\frac{1}{2}p^{-\gamma}\Delta$. Then, the error rate is given by

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

with

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

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

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 $h_c = \sqrt{1-c}$.

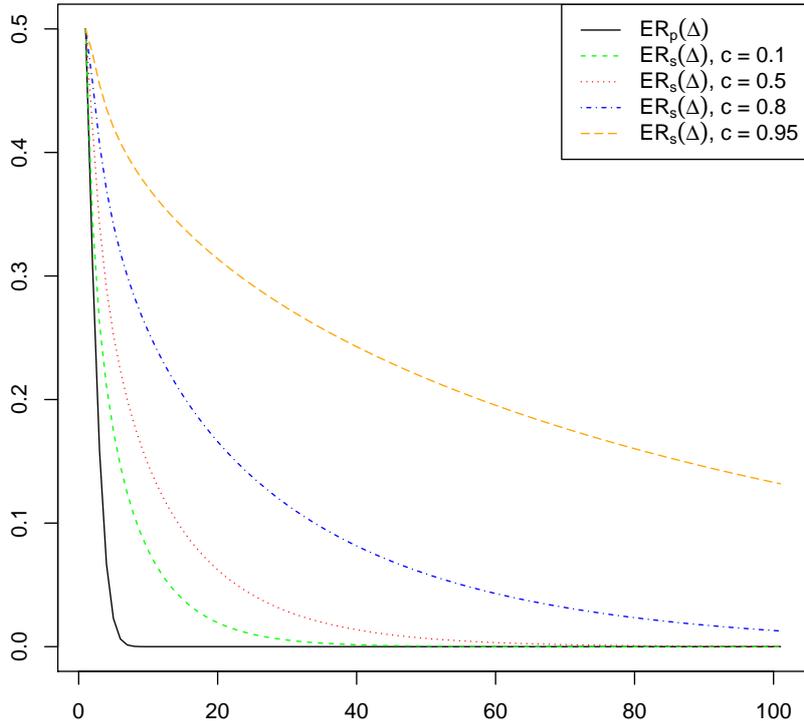In Figure 2, we plot $ER_s(\Delta)$ as a function of $\Delta \in [0,100]$ for $c \in \{0.1, 0.5, 0.8, 0.95\}$.

Figure 2: Error rates $ER_p(\Delta)$ and $ER_s(\Delta)$ as functions of Δ 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 Δ 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 Δ.

3.2 Finite-sample performance

In this section 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 $\hat{\theta} = \mathbf{l}^T \hat{\mathbf{a}}$ 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

20

univariate distributions provided in Corollary 1; (ii) second, $\hat{\theta}_b$ is computed by using (11) and standardized after that as in Theorem 4; (iii) finally, the previous two steps are repeated for $b = 1, ..., B$ times to obtained 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 γ.

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 3), while the plots in Figure 5 correspond to the case of one small sample and one large sample. We observe that the impact of the incorrect specification of γ is not large, while some deviations are observed in Figure 5 for small values of c. If c increases, the difference between 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.

References

Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. Annals of Statistics, 40(2):1171–1197.

Bai, Z., Jiang, D., Yao, J.-F., and Zheng, S. (2009). Corrections to lrt on large-dimensional covariance matrix by rmt. Annals of Statistics, 37(6B):3822–3840.

Bai, Z. and Silverstein, J. W. (2010). Spectral Analysis of Large Dimensional Random Matrices. New York, NY: Springer Science+ Business Media, LLC.

Bickel, P. J. and Levina, E. (2004). Some theory for fisher's linear discriminant function,'naive bayes', and some alternatives when there are many more variables than observations. Bernoulli, pages 989–1010.

Bodnar, T., Gupta, A., and Parolya, N. (2016). Direct shrinkage estimation of large dimensional precision matrix. Journal of Multivariate Analysis, 146:223–236.

Bodnar, T., Gupta, A. K., and Parolya, N. (2014). On the strong convergence of the optimal linear shrinkage estimator for large dimensional covariance matrix. Journal of Multivariate Analysis, 132:215–228.

Bodnar, T., Mazur, S., and Okhrin, Y. (2017a). Bayesian estimation of the global minimum variance portfolio. European Journal of Operational Research, 256:292–307.

Bodnar, T., Mazur, S., and Parolya, N. (2017b). Central limit theorems for functionals of large dimensional sample covariance matrix and mean vector in matrix-variate skewed model. Scandinavian Journal of Statistics, under revision.

Bodnar, T. and Okhrin, Y. (2008). Properties of the singular, inverse and generalized inverse partitioned Wishart distributions. Journal of Multivariate Analysis, 99:2389–2405.

Bodnar, T. and Okhrin, Y. (2011). On the product of inverse wishart and normal distributions with applications to discriminant analysis and portfolio theory. Scandinavian Journal of Statistics, 38(2):311–331.

Bodnar, T. and Reiß, M. (2016). Exact and asymptotic tests on a factor model in low and large dimensions with applications. Journal of Multivariate Analysis, 150:125–151.

Bodnar, T. and Schmid, W. (2008). A test for the weights of the global minimum variance portfolio in an elliptical model. Metrika, 67(2):127–143.

Cai, T. and Liu, W. (2011a). Adaptive thresholding for sparse covariance matrix estimation. Journal of the American Statistical Association, 106(494):672–684.

Cai, T. and Liu, W. (2011b). A direct estimation approach to sparse linear discriminant analysis. Journal of the American Statistical Association, 106(496):1566–1577.

Cai, T., Liu, W., and Luo, X. (2011). A constrained l_1 minimization approach to sparse precision matrix estimation. Journal of the American Statistical Association, 106(494):594–607.

Cai, T. T. and Jiang, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. Annals of Statistics, 39(3):1496–1525.

Chen, S. X., Zhang, L.-X., and Zhong, P.-S. (2010). Tests for high-dimensional covariance matrices. Journal of the American Statistical Association, 105(490):810–819.

DasGupta, A. (2008). Asymptotic theory of statistics and probability. Springer Science & Business Media.

Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. Journal of Econometrics, 147(1):186–197.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(4):603–680.

Fujikoshi, Y. and Seo, T. (1997). Asymptotic aproximations for epmcs of the linear and the quadratic discriminant functions when the sample sizes and the dimension are large. Technical report series, University of Toronto.

Givens, G. H. and Hoeting, J. A. (2012). Computational statistics. John Wiley & Sons.

Gupta, A. and Nagar, D. (2000). Matrix Variate Distributions. Chapman and Hall/CRC, Boca Raton.

Gupta, A., Varga, T., and Bodnar, T. (2013). Elliptically contoured models in statistics and portfolio theory. Springer, second edition.

Gupta, A. K. and Bodnar, T. (2014). An exact test about the covariance matrix. Journal of Multivariate Analysis, 125:176–189.

Jiang, T. and Yang, F. (2013). Central limit theorems for classical likelihood ratio tests for high-dimensional normal distributions. Annals of Statistics, 41(4):2029–2074.

Johnson, R. A., Wichern, D. W., et al. (2007). Applied multivariate statistical analysis. Prentice hall Upper Saddle River, NJ.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. Annals of Statistics, 29(2):295–327.

Ledoit, O. and Wolf, M. (2003). Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance, 10(5):603–621.

Mathai, A. and Provost, S. B. (1992). Quadratic Forms in Random Variables. Marcel Dekker.

Muirhead, R. J. (1982). Aspects of Multivariate Statistical Theory. Wiley, New York.

Narsky, I. and Porter, F. C. (2013). Linear and quadratic discriminant analysis, logistic regression, and partial least squares regression. Statistical Analysis Techniques in Particle Physics: Fits, Density Estimation and Supervised Learning, pages 221–249.

Rencher, A. C. (1998). Multivariate statistical inference and applications, volume 338. Wiley-Interscience.

Rencher, A. C. and Christensen, W. F. (2012). Methods of multivariate analysis. John Wiley & Sons.

Shao, J., Wang, Y., Deng, X., Wang, S., et al. (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. The Annals of statistics, 39(2):1241–1265.

Srivastava, M. S. and Kubokawa, T. (2007). Comparison of discrimination methods for high dimensional data. Journal of the Japan Statistical Society, 37:123–134.

Tamatani, M. (2015). 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, pages 15–26.

Wyman, F. J., Young, D. M., and Turner, D. W. (1990). A comparison of asymptotic error rate expansions for the sample linear discriminant function. Pattern Recognition, 23:775–783.
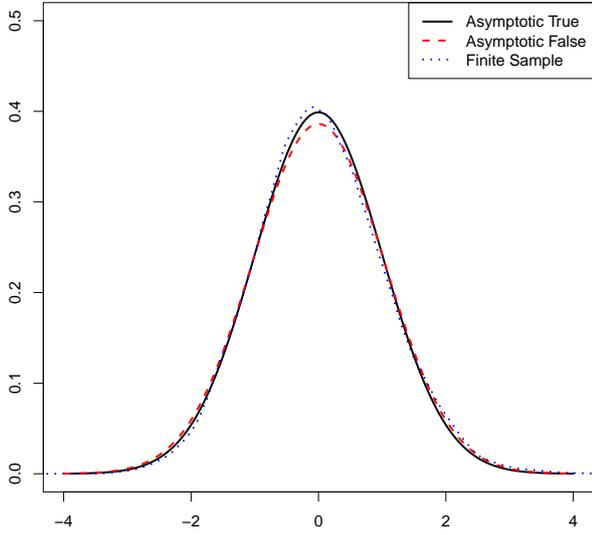
a) $p = 50$, $n_1 = n_2 = 250$
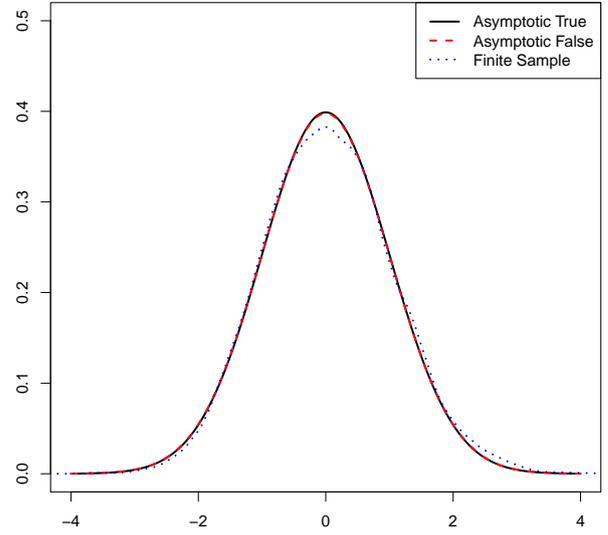
b) $p = 250$, $n_1 = n_2 = 250$
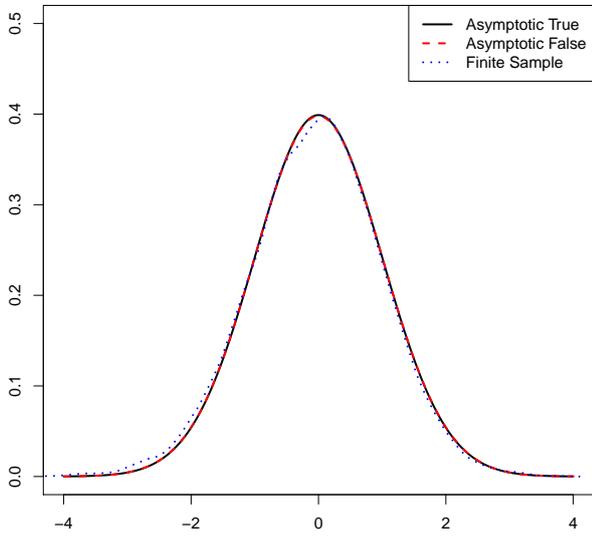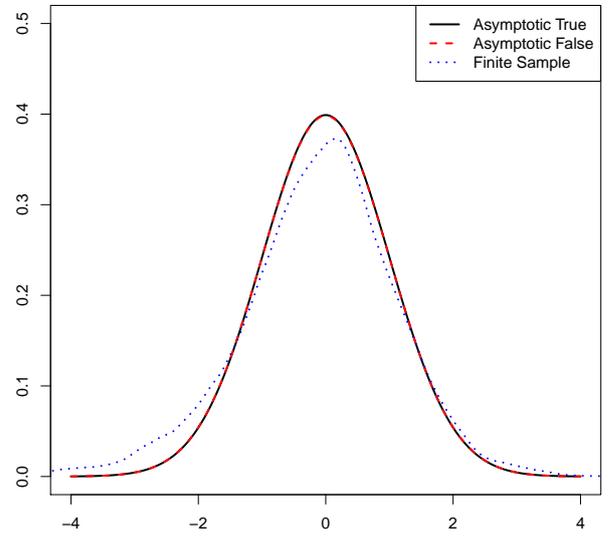
c) $p = 400$, $n_1 = n_2 = 250$
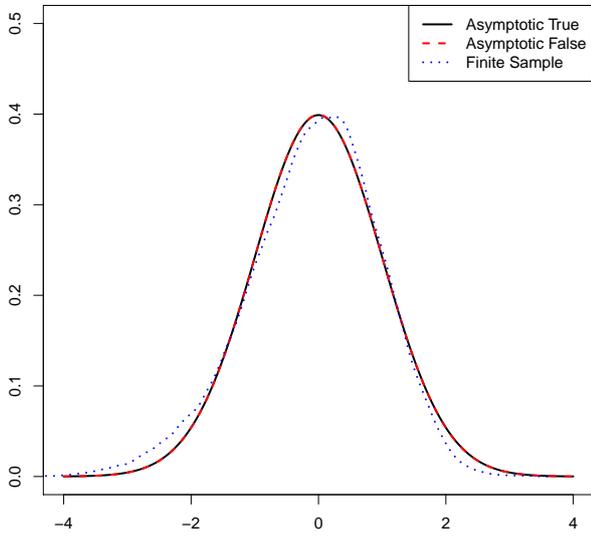
d) $p = 475$, $n_1 = n_2 = 250$

Figure 3: The kernel density estimator of the asymptotic distribution and standard normal for $\hat{\theta}$ as given in Theorem 4 for $\gamma = 0$ and $c = \{0.1, 0.5, 0.8, 0.95\}$.
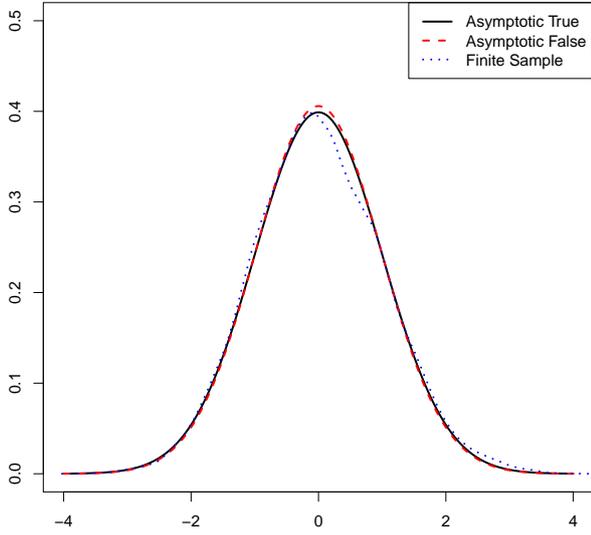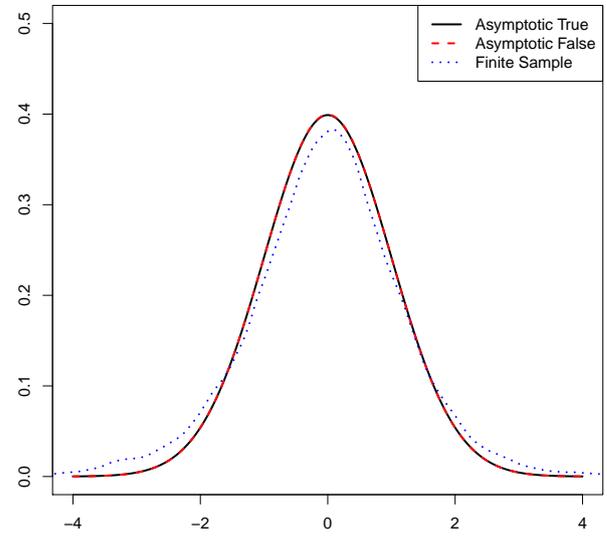
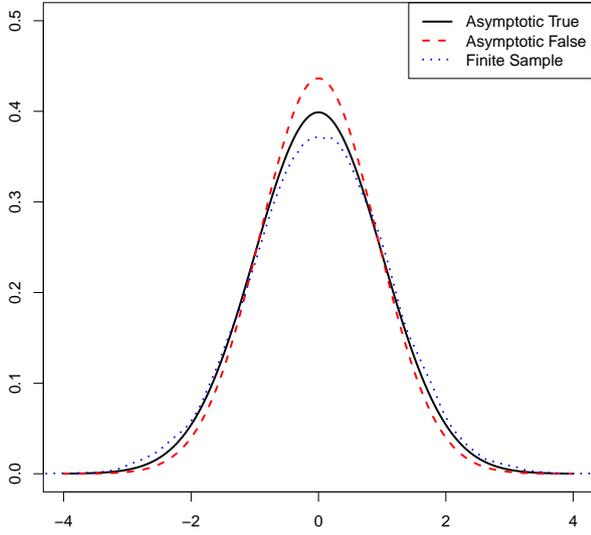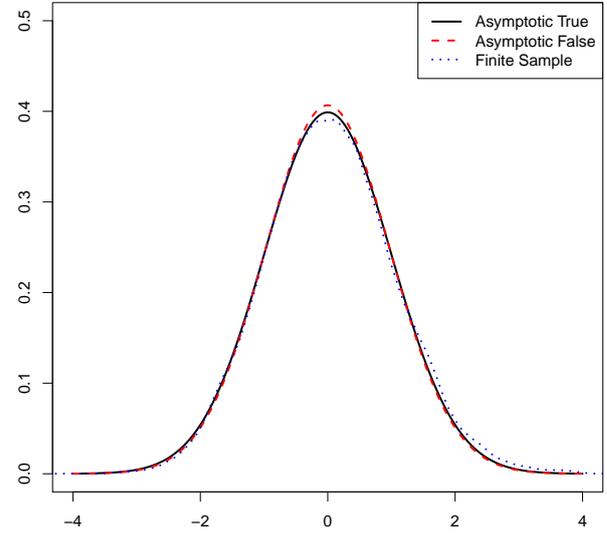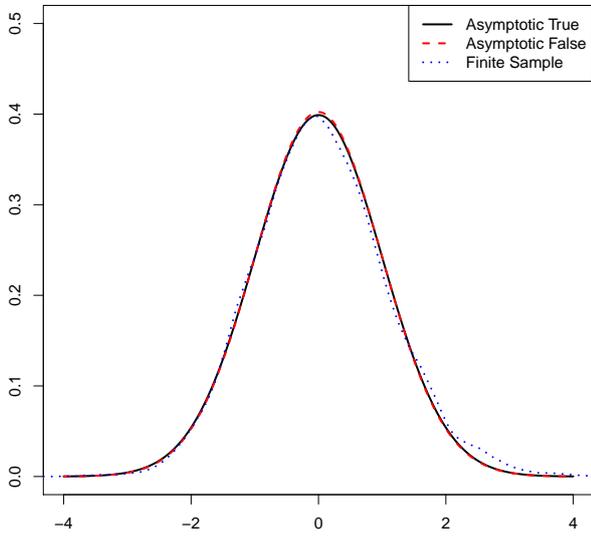a) $p = 50, \ n_1 = n_2 = 250$
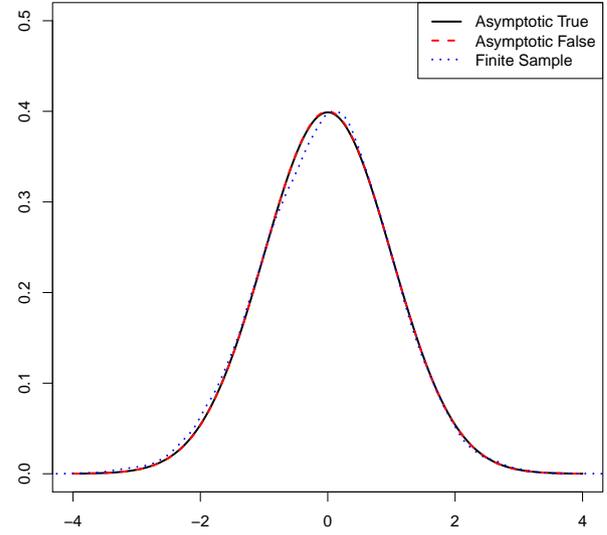
b) $p = 250, \ n_1 = n_2 = 250$

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 for $\hat{\theta}$ as given in Theorem 4 for $\gamma > 0$ and $c = \{0.1, 0.5, 0.8, 0.95\}$.
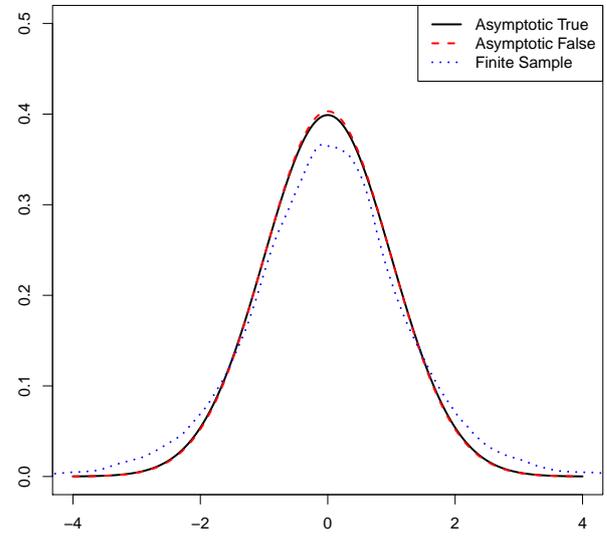
a) $p = 50, \ n_1 = 25, \ n_2 = 475$

b) $p = 250, \ n_1 = 25, \ n_2 = 475$

c) $p = 400, \ n_1 = 25, \ n_2 = 475$

d) $p = 475, \ n_1 = 25, \ n_2 = 475$

Figure 5: The kernel density estimator of the asymptotic distribution and standard normal for $\hat{\theta}$ as given in Theorem 4 for $\gamma > 0$ and $c = \{0.1, 0.5, 0.8, 0.95\}$.