Sankhyā : The Indian Journal of Statistics
2013, Volume 75-A, Part 2, pp. 300-326

On Kolmogorov asymptotics of estimators of the misclassification error rate in linear discriminant analysis

Amin Zollanvari
Texas A&M University, College Station, USA
Marc G. Genton
King Abdullah University of Science and Technology, Thuwal, Saudi Arabia

Abstract

We provide a fundamental theorem that can be used in conjunction with Kolmogorov asymptotic conditions to derive the first moments of well-known estimators of the actual error rate in linear discriminant analysis of a multivariate Gaussian model under the assumption of a common known covariance matrix. The estimators studied in this paper are plug-in and smoothed resubstitution error estimators, both of which have not been studied before under Kolmogorov asymptotic conditions. As a result of this work, we present an optimal smoothing parameter that makes the smoothed resubstitution an unbiased estimator of the true error. For the sake of completeness, we further show how to utilize the presented fundamental theorem to achieve several previously reported results, namely the first moment of the resubstitution estimator and the actual error rate. We provide numerical examples to show the accuracy of the succeeding finite sample approximations in situations where the number of dimensions is comparable or even larger than the sample size.

AMS (2000) *subject classification.* Primary 03C45: Classification theory, stability and related concepts; Secondary 62G20: Asymptotic properties.
Keywords and phrases. Double asymptotics, error estimation, Kolmogorov asymptotic analysis, plug-in error, resubstitution, smoothed resubstitution, true error.

1 Introduction

1.1. LDA and error estimation. Linear discriminant analysis (LDA), originally based on an idea proposed by R. A. Fisher using the linear regression procedure (Fisher, 1936, 1940), has a long history in statistics and pattern recognition. It was further developed by Wald (1944) in the context of decision theory and then formulated by Anderson (1951) in terms of what is known today as Anderson's statistic. From the first use on taxonomic classification by Fisher (1936), LDA-Fisher-based classification and recognition systems have been used in many disciplines, such as speech recognition (van Vuuren and Hermansky, 1997), face recognition (Swets and Weng, 1996) and, recently, cancer classification (Kim et al., 2002).

The successful applicability of a designed classifier relies on the measures of its predictive power, namely the actual or true error (Hills, 1966; Moran, 1975). However, in practice it is almost always the case that exactly evaluating the true error of a designed classifier is virtually impossible due to the lack of knowledge about the underlying distribution of the data. Therefore, methods of *error estimation* are needed to assess the performance of a classifier based on the given data.

Different error estimation techniques have been proposed through the years. For a comprehensive list of these error estimators, the reader is referred to Lachenbruch and Mickey (1968) and Schiavo and Hand (2000). Over these years, many researchers have tried to characterize the moments of error estimators and, in particular, the first moment (Fukunaga and Hayes, 1989; Kittler and DeVijver, 1982; McLachlan, 1976; Sorum, 1971, 1972). By comparing the first moment of estimator with that of true error, obtained for example in Anderson (1973), Efron (1975, 1980), McLachlan (1973, 1974), Okamoto (1963), many suggestions have been made on applicability of error estimators in practice (Foley, 1972; McLachlan, 1976; Raudys and Jain, 1991; Raudys and Pikelis, 1980). Most of the work mentioned so far has used asymptotic expansions based on the theory of infinitely large samples that do not apply to small-sample situations that we face today, especially in medical applications. We would like to highlight a quotation from Fisher (1925) that was mentioned by Martin and Hirschberg (1996): "*The traditional machinery of statistical processes is wholly unsuited to the needs of practical research ... the elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their metrics does it seem possible to apply accurate tests to practical data*". Hand (1986) also commented on this issue in the context of error estimation: "*Unfortunately, as Kittler and Devijver point out, small-sample performance of these average conditional error rate estimators often does not live up to asymptotic promise*".

Yet, it may be difficult for the readers to understand the serious implications of misusing asymptotic-performance-guaranteed tools in small-sample situations. In this regard, we cite some work reporting seriously flawed results in medical applications where large number of variables, e.g., genes, but small number of samples, e.g., patients, are available (a typical small-sample situation). For example, in Michiels and Koscielny (2005), it has been mentioned that "*Five of the seven largest published studies addressing* *cancer prognosis did not classify patients better than chance*". In another study (Dupuy and Simon, 2008), the authors mentioned 21 studies that had flawed results and were published in journals mostly with impact factor higher than 6. It is interesting to mention that the lack of reproducibility of some of these studies is partially due to misuse of error estimators in small-sample situations (Dougherty, Hua and Bittner, 2007; Dupuy and Simon, 2008; Gevaert et al., 2008).

1.2. Why Kolmogorov asymptotics? The pressing need for studying the estimators of misclassification in small-sample situations is clear. However, one may wonder if there is any method to analyze the performance of statistics and, in particular, error estimators, in small-sample situations. Fortunately, there exist some results on this issue that we discuss here.

Pioneering work in pattern recognition by Deev (1972) and Raudys (1972) in the Soviet Union characterized the error rate of LDA. They made the assumption of increasing sample size and dimension, at a fixed rate between the two, i.e., $n_0 \to \infty, n_1 \to \infty, p \to \infty, p/n_0 \to J_0 < \infty, p/n_1 \to J_1 < \infty$. This assumption generally provides accurate approximations for studying the properties of statistics in situations where the number of dimensions is comparable to that of the sample size. Serdobolskii, who made a large contribution to developing this theory in the Soviet Union, published a book (Serdobolskii, 2000) to explain the connection of this theory, which he calls the Kolmogorov asymptotic approach, to another independently developed theory initiated by work of Wigner (1958), a physicist who achieved interesting results on the spectral distribution of random matrices of increasing dimensions.

There has been a good deal of research to study the performance of linear classification rules (Deev, 1970, 1972; Fujikoshi, 2000; Fujikoshi and Seo, 1998; Meshalkin and Serdobolskii, 1978; Raudys, 1972, 1998; Raudys and Skurikhina, 1995; Serdobolskii, 1979) using Kolmogorov asymptotics. For the first time, Raudys (1972) used the assumption of increasing sample size and dimensionality to justify the use of a normal approximation to obtain the expectation of LDA true error. Later in the 1970's, under some regularity assumptions and Kolmogorov asymptotic conditions, Deev (1970, 1972), Meshalkin and Serdobolskii (1978), and Serdobolskii (2000) found the asymptotically exact expression for the LDA true error for a common unknown covariance matrix of classes. Fujikoshi (2000) essentially took the same Kolmogorov asymptotic framework, which he called a type-II approach, and using a different approach obtained asymptotically exact expressions for the first moment of the LDA true error.

1.3. Why are error estimators the focus of the current work? The *validity* of the classifier relies upon the quality of the error estimators (Dougherty et al., 2007; Dougherty, 2008). However, results on characterizing the quality of error estimators and, in particular, on their moments using Kolmogorov asymptotic conditions are much scarcer and limited to Raudys (1978). The first author of this article recently considered the approach taken by Deev (1972) and Serdobolskii (2000) to obtain asymptotically exact expressions of first and second moments of the resubstitution and leave-one-out error estimators of misclassification and their association with true error (Zollanvari, Braga-Neto and Dougherty, 2011). Here, we first present a general theory that is used later to find the first moment of smoothed resubstitution and plug-in error estimators. We show how this result can be used to determine an optimal parameter of smoothing to make a smoothed resubstitution estimator unbiased. We further show that the same general theory can be used to achieve several previously reported results on the moments of resubstitution and true error. The ultimate goal of the line of this research is to analyze the applicability of error estimators in high dimensions and further synthesize accurate error estimators in terms of *bias*, *variance* and *computational complexity*.

1.4. Why a known covariance matrix? Clearly, in practice, the assumption of a common known covariance matrix of classes imposes a limitation on use of the results. However, as has generally been the case, the results for the known covariance matrix have been obtained prior to those for the unknown covariance matrix, the latter typically being significantly more difficult. For instance, see Foley (1972), John (1961), and Sorum (1971, 1973) to cite just a few articles that have made the same assumption of a common known covariance matrix of classes in the LDA context. This assumption is not specific to the context of LDA and historically it has been a starting point for studying various statistics, e.g., see Conte, Lops and Ricci (1996), Moreira (2009), and Wacker and El-Sheikh (1984). Obtaining the corresponding results presented in this paper for an unknown covariance matrix is the next logical step in this line of research.

1.5. Organization of the paper. In Section 2, we present LDA, the definition of its true error, and several estimators of the true error that are the main focus of this work. In Section 3, we present the novel results achieved in this paper. We first present a theorem that we call the fundamental theorem of known covariance LDA, as it can be used for analyzing the performance of several estimators of LDA true error as well as the true error itself. We further apply this theorem to characterize the first moment of the plug-in error estimator and smoothed resubstitution, both of which have not been previously analyzed in the literature. We use the results of this section to find an "optimal" smoothing parameter that produces an unbiased smoothed resubstitution estimator. In Section 4, we show that the main theorem presented in Section 3 can be used in the same manner to prove several previously reported results on the first moment of the resubstitution error estimator and the true error of LDA. Section 5 provides succeeding finite sample approximations and numerical examples showing the accuracy of these approximations in situations where the number of dimensions is comparable or even larger than the sample size. The paper ends with the discussion section.

2 LDA, true error, and its estimators

In this section, we define the problem of discriminant analysis, its actual or true error, and then various estimators commonly used to estimate this true error.

2.1. LDA and its true error Consider a set of $n = n_0 + n_1$ independent and identically distributed (iid) observations of dimension p, where n_0 observations $\{X_1, X_2, \dots, X_{n_0}\}$ come from population Π_0 and n_1 observations $\{X_{n_0+1}, X_{n_0+2}, \dots, X_{n_0+n_1}\}$ come from population Π_1. Population Π_i is assumed to follow a multivariate Gaussian distribution, $N(\mu_i, \Sigma)$, for $i = 0, 1$. LDA employs Anderson's W statistic,

$$W(\bar{X}_0, \bar{X}_1, X) = \left(X - \frac{\bar{X}_0 + \bar{X}_1}{2}\right)^T \Sigma^{-1}(\bar{X}_0 - \bar{X}_1),$$

where $\bar{X}_0 = 1/n_0 \sum_{i=1}^{n_0} X_i$ and $\bar{X}_1 = 1/n_1 \sum_{i=n_0+1}^{n_0+n_1} X_i$ are the sample means for each class. The designed LDA classifier is

$$\psi(X) = \begin{cases} 1\,, & \text{if } W(\bar{X}_0, \bar{X}_1, X) \le 0\,, \\ 0\,, & \text{if } W(\bar{X}_0, \bar{X}_1, X) > 0\,; \end{cases}$$

that is, the sign of W determines the classification of the sample point, X. Here, following John (1961), Moran (1975), and Raudys (1967), we assume that the covariance matrix, Σ, is known and fixed; in particular, the W statistic is not a function of the sample covariance matrix, $\hat{\Sigma}$. In practice, however, if Σ is not known, then $\hat{\Sigma}$ may be plugged in as an estimator of Σ. Given the training data (and thus the sample means $\bar{X}_0$ and $\bar{X}_1$), the classification error is

$$\begin{aligned} \epsilon &= P\{W(\bar{X}_0, \bar{X}_1, X) \le 0, X \in \Pi_0 \mid \bar{X}_0, \bar{X}_1\} + P\{W(\bar{X}_0, \bar{X}_1, X) \\ &\qquad > 0, X \in \Pi_1 \mid \bar{X}_0, \bar{X}_1\} \qquad (2.1) \\ &= \alpha_0 \epsilon^0 + \alpha_1 \epsilon^1\,, \end{aligned}$$

where $\alpha_i = P(X \in \Pi_i)$ is the *a priori* mixing probability for population Π_i, and ϵ^i is the error rate specific to population Π_i, with

$$\begin{aligned}\epsilon^0 &= P\{W(\bar{X}_0, \bar{X}_1, X) \leq 0 | X \in \Pi_0, \bar{X}_0, \bar{X}_1\}\,,\\ \epsilon^1 &= P\{W(\bar{X}_0, \bar{X}_1, X) > 0 | X \in \Pi_1, \bar{X}_0, \bar{X}_1\},\end{aligned}$$

and therefore,

$$\begin{aligned}\epsilon = \alpha_0 \Phi &\left\{ -\frac{\left(\mu_0 - \frac{1}{2}(\bar{X}_0 + \bar{X}_1)\right)^T \Sigma^{-1}(\bar{X}_0 - \bar{X}_1)}{\sqrt{(\bar{X}_0 - \bar{X}_1)^T \Sigma^{-1} (\bar{X}_0 - \bar{X}_1)}} \right\}\\ &+ \alpha_1 \Phi \left\{ \frac{\left(\mu_1 - \frac{1}{2}(\bar{X}_0 + \bar{X}_1)\right)^T \Sigma^{-1}(\bar{X}_0 - \bar{X}_1)}{\sqrt{(\bar{X}_0 - \bar{X}_1)^T \Sigma^{-1} (\bar{X}_0 - \bar{X}_1)}} \right\},\end{aligned} \tag{2.2}$$

where Φ is the cumulative distribution function of a standard Gaussian random variable.

In order to evaluate the overall performance of the classification rule (here LDA) over all sample spaces given the parent distributions of classes, we use:

$$\begin{aligned}\mathrm{E}(\epsilon) = \alpha_0\, \mathrm{E}(\epsilon^0) + \alpha_1\, \mathrm{E}(\epsilon^1) = \alpha_0 P\{W(\bar{X}_0, \bar{X}_1, X) \leq 0 \mid X \in \Pi_0\}\\ + \alpha_1 P\{W(\bar{X}_0, \bar{X}_1, X) > 0 \mid X \in \Pi_1\}.\end{aligned}$$

2.2. The resubstitution error estimator The apparent classification error, or resubstitution error estimator (Smith, 1947), is

$$\begin{aligned}\hat{\epsilon}_r &= \frac{1}{n}\left[\sum_{i=1}^{n_0} I\{W(\bar{X}_0, \bar{X}_1, X_i) \leq 0\} + \sum_{i=n_0+1}^{n_0+n_1} I\{W(\bar{X}_0, \bar{X}_1, X_i) > 0\} \right]\\ &= \hat{\alpha}_0 \hat{\epsilon}_r^0 + \hat{\alpha}_1 \hat{\epsilon}_r^1\,,\end{aligned} \tag{2.3}$$

where $I\{A\}$ is the indicator variable for the event, A, and $\hat{\alpha}_i = n_i/n$ is the empirical mixing frequency for population Π_i. Therefore, we have:

$$\begin{aligned}\mathrm{E}(\hat{\epsilon}_r) &= \hat{\alpha}_0\, \mathrm{E}(\hat{\epsilon}_r^0) + \hat{\alpha}_1\, \mathrm{E}(\hat{\epsilon}_r^1)\\ &= \hat{\alpha}_0 P\{W(\bar{X}_0, \bar{X}_1, X_1) \leq 0\} + \hat{\alpha}_1 P\{W(\bar{X}_0, \bar{X}_1, X_{n_0+1}) > 0\}.\end{aligned}$$

2.3. The plug-in error estimator This estimator of true error, originally proposed by Fisher (1936), is obtained by replacing μ_0, α_0 and α_1 by $\bar{X}_0$, $\hat{\alpha}_0$

and $\hat{\alpha}_1$ in (2.2) and is called a *plug-in* estimator. We denote it by $\hat{\epsilon}_p$. After simplification, we have $\hat{\epsilon}_p = \Phi(-\hat{\delta}/2)$ as given by Moran (1975), where $\hat{\delta} = \sqrt{(\bar{X}_0 - \bar{X}_1)^T \Sigma^{-1} (\bar{X}_0 - \bar{X}_1)}$ is the estimated Mahalanobis distance between the classes.

Then, it is of interest to look into $\mathrm{E}(\hat{\epsilon}_p)$, which we investigate it in this paper. Using simulation studies, Lachenbruch and Mickey (1968) and Moran (1975) stated that this estimator behaves similarly to resubstitution. We confirm this claim analytically here.

2.4. The smoothed resubstitution error estimator The idea behind *smoothed estimators* (Glick, 1978) is simply to replace sharp indicator functions by "smooth" functions taking values in the interval $[0,1]$, thereby reducing the variance of the original estimator. In Glick (1978), Hirst (1996), and Snapinn and Knoke (1985, 1989), this idea is applied to the resubstitution error estimator for LDA classification, in which case one replaces the indicator function I in (2.3) by a smooth function, g:

$$\begin{aligned}\hat{\epsilon}_{sr} &= \frac{1}{n}\left(\sum_{i=1}^{n_0} g\{W(\bar{X}_0, \bar{X}_1, X_i)\} + \sum_{i=n_0+1}^{n_0+n_1} [1 - g\{W(\bar{X}_0, \bar{X}_1, X_i)\}]\right) \\ &= \hat{\alpha}_0 \hat{\epsilon}_{sr}^0 + \hat{\alpha}_1 \hat{\epsilon}_{sr}^1,\end{aligned}$$

where

$$\hat{\epsilon}_{sr}^0 = \frac{1}{n_0} \sum_{i=1}^{n_0} g\{W(\bar{X}_0, \bar{X}_1, X_i)\},$$

with a similar expression for $\hat{\epsilon}_{sr}^1$ by replacing X_1 with X_{n_0+1}. It follows that $\mathrm{E}(\hat{\epsilon}_{sr}^0) = \mathrm{E}[g\{W(\bar{X}_0, \bar{X}_1, X_1)\}]$, with a similar expression for $\mathrm{E}(\hat{\epsilon}_{sr}^1)$ by replacing X_1 with X_{n_0+1}. The choice of the smoothing function is naturally critical to the performance of the error estimator. Typical choices include linear and Gaussian smoothers. In this article, we adopt the latter choice, in the form proposed by Snapinn and Knoke (1985), i.e., $g(x) = \Phi\{-x/(b\,\hat{\delta})\}$ where $\hat{\delta} = \sqrt{(\bar{X}_1 - \bar{X}_0)^T \Sigma^{-1} (\bar{X}_1 - \bar{X}_0)}$ and b is a free parameter that must be provided, which is typical of smoothing methods. As $b \to 0$, there is no smoothing and the estimator reduces to plain resubstitution (generally negatively biased), whereas as $b \to \infty$, there is maximal smoothing and the estimator becomes a constant (generally positively biased). One approach to selecting b is thus to find the value that moves the bias to zero, yielding an unbiased estimator. We return to this point in Section 3 when we discuss optimal smoothing.

2.5. The leave-one-out error estimator The leave-one-out error estimator (Lachenbruch and Mickey, 1968) for the LDA classification rule is given by

$$\begin{aligned}\hat{\epsilon}_l &= \frac{1}{n}\left[\sum_{i=1}^{n_0} I\{W^{(i)}(\bar{X}_0,\bar{X}_1,X_i)\le 0\} \,+\, \sum_{i=n_0+1}^{n_0+n_1} I\{W^{(i)}(\bar{X}_0,\bar{X}_1,X_i)>0\}\right]\\ &= \hat{\alpha}_0\hat{\epsilon}_l^0 \,+\, \hat{\alpha}_1\hat{\epsilon}_l^1\,,\end{aligned}$$

where $W^{(i)}$ is the discriminant obtained when observation X_i is left out of the training, $\hat{\alpha}_i$ is defined as before, and $\hat{\epsilon}_l^i$ is the leave-one-out error rate specific to population Π_i, with

$$\begin{aligned}\hat{\epsilon}_l^0 &= \frac{1}{n_0}\sum_{i=1}^{n_0} I\{W^{(i)}(\bar{X}_0,\bar{X}_1,X_i)\le 0\}\,,\\ \hat{\epsilon}_l^1 &= \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} I\{W^{(i)}(\bar{X}_0,\bar{X}_1,X_i)>0\}\,.\end{aligned}$$

However, from the definition of this estimator, it is clear that we have $\mathrm{E}(\hat{\epsilon}_l) = \hat{\alpha_0}\,\mathrm{E}(\epsilon_{n_0-1}^0) + \hat{\alpha_1}\,\mathrm{E}(\epsilon_{n_1-1}^1)$ where $\epsilon_{n_0-1}^0$ and $\epsilon_{n_1-1}^1$ are conditional true errors defined in (2.1) for a problem of n_0-1 and n_1-1 observations, respectively. Therefore, studying the expectation of the true error of the misclassification suffices to determine that of the leave-one-out estimator. Consequently, we do not consider this estimator further in this study.

3 Novel results

3.1. The fundamental theorem of known covariance LDA We call the theorem derived in this section the *fundamental theorem of known covariance LDA (FTKCLDA)* since it can be used for characterizing the average performance of several error estimators of the LDA true error as well as the true error itself. As we will see later, to apply this theorem properly in studying the statistics considered here, we simply need to determine the proper parameters.

Theorem 3.1. *Let $Z_1 \sim \chi_p^2(\nu_1)$ and $Z_2 \sim \chi_p^2(\nu_2)$ denote two independent non-central chi-square random variables with degrees of freedom p and non centrality parameters ν_1 and ν_2, respectively. Assume that λ_1 and λ_2 are two parameters such that, for $\nu_1,\nu_2 \in \mathbb{R}^+\cup\{+\infty\}$ and $p\in\mathbb{N}\cup\{+\infty\}$, the following conditions hold:*

$$\lim_{\substack{p\to\infty\\ \lambda_1\to 0\\ \lambda_2\to 0}} (\lambda_1+\lambda_2)p \;=\; c_1<\infty\,, \tag{3.1}$$

$$\lim_{\substack{\nu_1\to\infty,\lambda_1\to 0\\ \nu_2\to\infty,\lambda_2\to 0}} (\lambda_1\nu_1+\lambda_2\nu_2) = c_2 < \infty\,, \tag{3.2}$$

$$\lim_{\substack{p\to\infty\\ \nu_1\to\infty,\lambda_1\to 0\\ \nu_2\to\infty,\lambda_2\to 0}} \left\{\lambda_1^2(a_1p+a_2\nu_1)+\lambda_2^2(a_1p+a_2\nu_2)\right\} = c_3 < \infty\,, \tag{3.3}$$

where c_i, $i=1,2,3$ and $a_j>0$, $j=1,2$ are some constants. Then, we have:

$$\vartheta = \lambda_1 Z_1 + \lambda_2 Z_2 \xrightarrow{D} N(\mathrm{E}_\infty(\vartheta), \mathrm{Var}_\infty(\vartheta))\,, \tag{3.4}$$

where

$$\mathrm{E}_\infty(\vartheta) = \lim_{\substack{p\to\infty\\ \nu_1\to\infty,\lambda_1\to 0\\ \nu_2\to\infty,\lambda_2\to 0}} \mathrm{E}(\vartheta)\,,$$

$$\mathrm{Var}_\infty(\vartheta) = \lim_{\substack{p\to\infty\\ \nu_1\to\infty,\lambda_1\to 0\\ \nu_2\to\infty,\lambda_2\to 0}} \mathrm{Var}(\vartheta)\,,$$

and $N(0,1)$ is a standard normal distribution.

Proof. As mentioned by Johnson, Kotz and Balakrishnan (1994), $\chi_p^2(\nu) = \chi_1^2(\nu) + \chi_{p-1}^2 = \chi_1^2(\nu) + \sum_{i=2}^p \chi_1^2$, where $\chi_1^2(\nu)$ is independent of χ_{p-1}^2. We first focus on the central chi-square part of Z_1 and Z_2. By the central limit theorem (CLT) for iid random variables, we have:

$$\vartheta_1 = \frac{\lambda_1\chi_{p-1}^2 - \lambda_1(p-1)}{\sqrt{2\lambda_1^2(p-1)}} \xrightarrow{D} N(0,1)\,. \tag{3.5}$$

Similarly, we have:

$$\vartheta_2 = \frac{\lambda_2\chi_{p-1}^2 - \lambda_2(p-1)}{\sqrt{2\lambda_2^2(p-1)}} \xrightarrow{D} N(0,1)\,. \tag{3.6}$$

Owing to the fact that p, ν_i, and a_i's are all positive, the condition stated in (3.3) implies that:

$$\lim_{\substack{p\to\infty\\ \lambda_1\to 0\\ \lambda_2\to 0}} \lambda_1^2 p = k_1,\ \lim_{\substack{p\to\infty\\ \lambda_1\to 0\\ \lambda_2\to 0}} \lambda_2^2 p = k_2,$$

where k_i, $i=1,2$ are some constants. By using Slutsky's theorem in (3.5) and (3.6) and noting the condition stated in (3.1), we have:

$$\lambda_1\chi_{p-1}^2 + \lambda_2\chi_{p-1}^2 \xrightarrow{D} N\left(\lim_{\substack{p\to\infty\\ \lambda_1\to 0\\ \lambda_2\to 0}}[(\lambda_1+\lambda_2)(p-1)],\ \lim_{\substack{p\to\infty\\ \lambda_1\to 0\\ \lambda_2\to 0}}[2\lambda_1^2(p-1)+2\lambda_2^2(p-1)]\right). \tag{3.7}$$

Now, consider the non-central part of Z_1 and Z_2. From Johnson et al. (1994), as $\nu_i \to \infty$, $i = 1, 2$:

$$\frac{\lambda_i \chi_1^2(\nu_i) - \lambda_i(1+\nu_i)}{\sqrt{2\lambda_i^2(1+2\nu_i)}} \xrightarrow{D} N(0,1)\,.$$

Owing to the fact that p, ν_i, and a_i's are all positive, the condition stated in (3.3) implies that:

$$\lim_{\substack{\nu_1\to\infty,\lambda_1\to 0\\ \nu_2\to\infty,\lambda_2\to 0}} \lambda_1^2\nu_1 = h_1, \quad \lim_{\substack{\nu_1\to\infty,\lambda_1\to 0\\ \nu_2\to\infty,\lambda_2\to 0}} \lambda_1^2\nu_2 = h_2,$$

where h_i, $i = 1, 2$ are some constants. Using Slutsky's theorem and noting the conditions stated in (3.1) and (3.2), we have:

$$\lambda_1\chi_1^2(\nu_1) + \lambda_2\chi_1^2(\nu_2) \xrightarrow{D} N\left(\lim_{\substack{\nu_1\to\infty,\lambda_1\to 0\\ \nu_2\to\infty,\lambda_2\to 0}} [\lambda_1(1+\nu_1) + \lambda_2(1+\nu_2)], \lim_{\substack{\nu_1\to\infty,\lambda_1\to 0\\ \nu_2\to\infty,\lambda_2\to 0}} [2\lambda_1^2(1+2\nu_1) + 2\lambda_2^2(1+2\nu_2)] \right). \tag{3.8}$$

Then, (3.4) follows from (3.7) and (3.8).

3.2. A double asymptotic approach in LDA For simplicity of notation, we assume that taking the limit under the Kolmogorov asymptotic conditions, i.e., $n_0 \to \infty, n_1 \to \infty, p \to \infty, p/n_0 \to J_0 < \infty, p/n_1 \to J_1 < \infty$, is simply denoted by $\lim_{p\to\infty}$. Another simplification of notation is to replace $W(\bar{X}_0, \bar{X}_1, X)$ and $W(\bar{X}_0, \bar{X}_1, X_1)$ with $W(X)$ and $W(X_1)$, respectively.

Consider a sequence of Gaussian discrimination problems defined by the sequence of parameters and sample sizes:

$$(\mu_{p,0}, \mu_{p,1}, \Sigma_p, n_{p,0}, n_{p,1})\,, \quad p = 1, 2, \ldots, \tag{3.9}$$

where the means and the covariance matrix are arbitrary except that the Mahalanobis distance, δ_p, defined as $\delta_p = \sqrt{(\mu_{p,0} - \mu_{p,1})^T \Sigma_p^{-1} (\mu_{p,0} - \mu_{p,1})} < \infty$ is such that $\lim_{p\to\infty} \delta_p = \delta$. For simplicity of notation, and at no risk of ambiguity, we subsequently omit the subscript "p" from the parameters and sample sizes in (3.9). However, we keep the subscript for δ_p whenever is needed to prevent any confusion. Furthermore, we add the $\lim_{p\to\infty} \delta_p = \delta$ condition to the aforementioned Kolmogorov asymptotic conditions (k.a.c.) and we denote taking the limit under all these conditions as $\lim_{\text{k.a.c.}}$, which means that

$$\lim_{\text{k.a.c.}} (\cdot) = \lim_{\substack{p\to\infty,\delta_p\to\delta\\ n_0\to\infty,n_1\to\infty\\ \frac{p}{n_0}\to J_0<\infty,\frac{p}{n_1}\to J_1<\infty}} (\cdot).$$

We will see in the next sections that Theorem 3.1 can be used in the same way as we used it here to find the expectation of different estimators of the true error under the Kolmogorov asymptotic conditions, which has not been considered yet.

3.3. The plug-in error estimator From Section 2.3, we have $\mathrm{E}(\hat{\epsilon}_p) = \mathrm{E}\{\Phi(-\hat{\delta}/2)\}$. Then, utilizing the FTKCLDA results in the following theorem:

THEOREM 3.2. *In the sequence of Gaussian discrimination problems defined by (3.9), under the Kolmogorov asymptotic conditions, we have:*

$$\lim_{k.a.c.} \mathrm{E}(\hat{\epsilon}_p) = \Phi\left(-\frac{1}{2}\frac{\delta^2 + J_1 + J_0}{\sqrt{\delta^2 + J_0 + J_1}}\right). \tag{3.10}$$

Therefore,

$$\lim_{k.a.c.} \mathrm{E}(\hat{\epsilon}_p) = \lim_{k.a.c.} \mathrm{E}(\hat{\epsilon}_r),$$

where $\mathrm{E}(\hat{\epsilon}_r)$ *is the first moment of the resubstitution true error, which is available from our results presented in Zollanvari et al. (2011) or a simplified version of the proof using FTKCLDA presented in Section 4.*

Proof. Moran (1975) showed that (by adapting the results stated there to the definition of discriminant used here) $\mathrm{E}(\hat{\epsilon}_p) = P(Y < 0)$, where Y is a random variable distributed as $(1+\rho)Z_1 - (1-\rho)Z_2$ in which $Z_1 \sim \chi^2_p(\nu_1)$ and $Z_2 \sim \chi^2_p(\nu_2)$ are independent and

$$\begin{aligned} \nu_i = {} & \frac{n_0 n_1}{2(1 + (-1)^{i+1}\rho)} \\ & \times \left(\frac{1}{\sqrt{n_0 + n_1}} + (-1)^{i+1}\frac{1}{\sqrt{n_0 + n_1 + 4n_0 n_1}}\right)^2 \delta_p^2, \quad i = 1, 2, \end{aligned} \tag{3.11}$$

and

$$\rho = \sqrt{\tfrac{n_0+n_1}{n_0+n_1+4n_0n_1}} \; .$$

We mention that in the corresponding formulas for ν_1 and ν_2 stated by Moran (1975), there are some typographic errors that we corrected in (3.11). To be able to use Theorem 3.1, we first define Y' as:

$$Y' = sY,$$

where $s = 1/4\sqrt{(n_0 + n_1)(n_0 + n_1 + 4n_0 n_1)}/n_0 n_1$ and note that $\mathrm{E}(\hat{\epsilon}_p) = P(Y < 0) = P(Y' < 0)$.

Under the Kolmogorov asymptotic conditions and by letting $\lambda_1 = s(1+\rho)$ and $\lambda_2 = -s(1-\rho)$, it follows that:

$$\lim_{\text{k.a.c.}} (\lambda_1 + \lambda_2)p = \lim_{\substack{p\to\infty \\ \lambda_1\to 0 \\ \lambda_2 \to 0}} (\lambda_1+\lambda_2)p = \lim_{\substack{p\to\infty \\ n_0\to\infty \\ n_1\to\infty}} \frac{p}{2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right) = \frac{1}{2}(J_0+J_1)\,, \tag{3.12}$$

$$\lim_{\text{k.a.c.}} (\lambda_1\nu_1 + \lambda_2\nu_2) = \lim_{\substack{\nu_1\to\infty,\lambda_1\to 0 \\ \nu_2\to\infty,\lambda_2\to 0}} (\lambda_1\nu_1 + \lambda_2\nu_2) = \lim_{p\to\infty} \frac{\delta_p^2}{2} = \frac{\delta^2}{2}, \tag{3.13}$$

and by taking $a_1 = 1/2$ and $a_2 = 1$, it follows that:

$$\begin{aligned}\lim_{\text{k.a.c.}} \mathrm{Var}(Y') &= \lim_{\substack{p\to\infty \\ \nu_1\to\infty,\lambda_1\to 0 \\ \nu_2\to\infty,\lambda_2\to 0}} \left(\lambda_1^2(2p+4\nu_1) + \lambda_2^2(2p+4\nu_2)\right) \\ &= \lim_{\substack{p\to\infty \\ n_0\to\infty \\ n_1\to\infty}} \left(\delta_p^2\left(1+\frac{1}{n_0}+\frac{1}{n_1}\right)\right. \\ &\qquad \left. + \, p\left(\frac{1}{n_0}+\frac{1}{n_1}+\frac{1}{2n_1^2}+\frac{1}{2n_0^2}+\frac{1}{n_0n_1}\right)\right) \\ &= \delta^2 + J_0 + J_1\,,\end{aligned} \tag{3.14}$$

and all the conditions of Theorem 3.1 hold. Noting the fact that

$$\begin{aligned}\lim_{\text{k.a.c.}} \mathrm{E}(Y') &= \lim_{\substack{p\to\infty \\ \nu_1\to\infty,\lambda_1\to 0 \\ \nu_2\to\infty,\lambda_2\to 0}} \lambda_1(p+\nu_1) + \lambda_2(p+\nu_2) \\ &= \lim_{\substack{p\to\infty \\ n_0\to\infty \\ n_1\to\infty}} \frac{p}{2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right) + \frac{\delta_p^2}{2} = \frac{1}{2}(\delta^2+J_0+J_1)\,,\end{aligned}$$

then the results follow by using (3.4) from Theorem 3.1.

3.4. The smoothed resubstitution error estimator In this section, we first derive the statistical representation of $\hat{\epsilon}_{sr}^0$ and then employ the FTK-CLDA to get $\mathrm{E}(\hat{\epsilon}_{sr})$. From Section 2, we have:

$$\mathrm{E}(\hat{\epsilon}_{sr}^0) \;=\; \mathrm{E}\left\{\Phi\left(-\frac{1}{b}\frac{W(\bar{X}_0,\bar{X}_1,X_1)}{\sqrt{(\bar{X}_1-\bar{X}_0)^T\Sigma^{-1}(\bar{X}_1-\bar{X}_0)}}\right)\right\}.$$

Then, the following theorem holds:

Theorem 3.3. *Let $X_i \sim N(\mu_0, \Sigma)$ for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \Sigma)$ for $i = n_0 + 1, \ldots, n_0 + n_1$. Then, for LDA:*

$$\mathrm{E}[\hat{\epsilon}^0_{sr}] \,=\, P\{(1+\rho)Z_1 \,-\, (1-\rho)Z_2 < 0\},$$

where

$$\rho \,=\, \sqrt{\frac{n_0 + n_1}{4n_0 n_1(1+b^2) + n_0 - n_1}}$$

and $Z_1 \sim \chi^2_p(\lambda_1)$ and $Z_2 \sim \chi^2_p(\lambda_2)$ are independent non-central chi-square random variables, with non-centrality parameters

$$\nu_i = \frac{n_0 n_1}{2(1+(-1)^{i+1}\rho)} \times \left\{ \frac{1}{\sqrt{n_0+n_1}} + (-1)^{i+1} \frac{1}{\sqrt{4n_0 n_1(1+b^2) + n_0 - n_1}} \right\}^2 \delta^2, \quad i = 1, 2$$

where $\delta \,=\, \sqrt{(\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1)}$ and $\mathrm{E}\left(\hat{\epsilon}^1_{sr}\right)$ is obtained from $\mathrm{E}\left(\hat{\epsilon}^0_{sr}\right)$ by exchanging n_0 and n_1.

Proof. The estimator $\hat{\epsilon}_{sr}$ is clearly invariant to any linear transformation such that we can use the canonically convenient form proposed by Dunn (1971), with $\Sigma = I$ and $\mu_0 = -\mu_1 = (\delta/2, 0, 0, \ldots, 0)^T$. Therefore, we have:

$$\mathrm{E}(\hat{\epsilon}^0_{sr}) \,=\, \mathrm{E}\left[\Phi \left\{ -\frac{1}{b} \frac{(X_1 - \frac{\bar{X}_0 + \bar{X}_1}{2})^T (\bar{X}_0 - \bar{X}_1)}{\sqrt{(\bar{X}_1 - \bar{X}_0)^T (\bar{X}_1 - \bar{X}_0)}} \right\} \right].$$

This can be written as:

$$\begin{aligned}
\mathrm{E}(\hat{\epsilon}^0_{sr}) \,&=\, \mathrm{E}\left[P\left(\left\{ Z + \frac{1}{b}\left(X_1 - \frac{\bar{X}_0 + \bar{X}_1}{2} \right) \right\}^T (\bar{X}_0 - \bar{X}_1) < 0 \;\middle|\; X_1, \bar{X}_0, \bar{X}_1 \right) \right] \\
&=\, P\left(\left\{ Z + \frac{1}{b}\left(X_1 - \frac{\bar{X}_0 + \bar{X}_1}{2} \right) \right\}^T (\bar{X}_0 - \bar{X}_1) < 0 \right)
\end{aligned}$$

in which $Z \sim N(0, I)$ is independent of $X_1, \ldots, X_{n_0+n_1}$. It follows that

$$\mathrm{E}(\hat{\epsilon}^0_{sr}) \,=\, P(U^T V < 0)$$

where

$$U \,=\, \frac{2\sqrt{n_0 n_1} b}{\sqrt{4n_0 n_1(1+b^2) + n_0 - n_1}} \left\{ Z + \frac{1}{b}\left(X_1 - \frac{\bar{X}_0 + \bar{X}_1}{2} \right) \right\},$$

$$V \,=\, \sqrt{\frac{n_0 n_1}{n_0 + n_1}} (\bar{X}_0 - \bar{X}_1),$$

with $\Sigma_U = \Sigma_V = I$. Therefore,

$$P(U^T V < 0) \;=\; P\left\{(U+V)^T(U+V) - (U-V)^T(U-V) < 0\right\},$$

where $U+V$ is independent from $U-V$, and $\Sigma_{U+V} = \Sigma_U + \Sigma_V + 2\Sigma_{UV} = 2(1+\rho)I$, where

$$\rho \;=\; \sqrt{\frac{n_0+n_1}{4n_0n_1(1+b^2)+n_0-n_1}}.$$

By taking $Z_1 = U + V/\sqrt{2(1+\rho)}$, we have $\Sigma_{Z_1} = I$, so that $(U+V)^T(U+V) = Z_1^T Z_1$ has the distribution of a non-central chi-square random variable, $\chi_p^2(\nu_1)$, in which $\nu_1 = \mu_{Z_1}^T \mu_{Z_1}$ with

$$\mu_{Z_1} \;=\; \sqrt{\frac{n_0n_1}{2(1+\rho)}}\left\{\frac{1}{\sqrt{n_0+n_1}} + \frac{1}{\sqrt{4n_0n_1(1+b^2)+n_0-n_1}}\right\}\delta\,.$$

Similarly, by taking $Z_2 = U - V/\sqrt{2(1-\rho)}$, we have $\Sigma_{Z_1} = I$ and $(U-V)^T(U-V)$ has the distribution of a non-central chi-square random variable, $\chi_p^2(\nu_2)$, in which $\nu_2 = \mu_{Z_2}^T \mu_{Z_2}$ with

$$\mu_{Z_2} \;=\; \sqrt{\frac{n_0n_1}{2(1-\rho)}}\left\{\frac{1}{\sqrt{n_0+n_1}} - \frac{1}{\sqrt{4n_0n_1(1+b^2)+n_0-n_1}}\right\}\delta\,,$$

and the theorem follows.

Using the FTKCLDA, we have:

Theorem 3.4. *In the sequence of Gaussian discrimination problems defined by (3.9), under the Kolmogorov asymptotic conditions, we have:*

$$\lim_{k.a.c.} \mathrm{E}(\hat{\epsilon}_{sr}) = \Phi\left(-\frac{1}{2}\frac{\delta^2+J_1+J_0}{\sqrt{(1+b^2)(\delta^2+J_0+J_1)}}\right) \tag{3.15}$$

and therefore

$$\lim_{k.a.c.} \mathrm{E}(\hat{\epsilon}_{sr}) > \lim_{k.a.c.} \mathrm{E}(\hat{\epsilon}_r),$$

where $\mathrm{E}(\hat{\epsilon}_r)$ *is the first moment of the resubstitution true error, which is available from our results presented in Zollanvari et al. (2011) or a simplified version of the proof using the FTKCLDA presented in Section 4. This means that under k.a.c., the smoothed resubstitution error estimator is less optimistically biased than is resubstitution.*

Proof. To be able to use Theorem 3.1, we first define Y' as:

$$Y' = sY,$$

where $s = 1/4\sqrt{(n_0+n_1)(4n_0n_1(1+b^2)+n_0-n_1)}/n_0n_1$ and note that $\mathrm{E}(\hat{\epsilon}_{sr}) = P(Y < 0) = P(Y' < 0)$.

Under the Kolmogorov asymptotic conditions ($n_0 \to \infty$, $n_1 \to \infty$, $p \to \infty$, $p/n_0 \to J_0$, $p/n_1 \to J_1$), we have $\nu_1 \to \infty$, $\nu_2 \to \infty$, $p \to \infty$, and by letting $\lambda_1 = s(1+\rho)$ and $\lambda_2 = -s(1-\rho)$, with ρ defined in Theorem 3.3, it follows that:

$$\lim_{\text{k.a.c.}}(\lambda_1+\lambda_2)p = \lim_{\substack{p\to\infty\\ \lambda_1\to 0\\ \lambda_2\to 0}}(\lambda_1+\lambda_2)p = \lim_{\substack{p\to\infty\\ n_0\to\infty\\ n_1\to\infty}}\frac{p}{2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right) = \frac{1}{2}(J_0+J_1)\,,$$

$$\lim_{\text{k.a.c.}}(\lambda_1\nu_1+\lambda_2\nu_2) = \lim_{\substack{\nu_1\to\infty,\lambda_1\to 0\\ \nu_2\to\infty,\lambda_2\to 0}}(\lambda_1\nu_1+\lambda_2\nu_2) = \lim_{p\to\infty}\frac{\delta_p^2}{2} = \frac{\delta^2}{2}.$$

By taking $a_1 = 1/2$ and $a_2 = 1$ and similar to (3.14), it follows that:

$$\begin{aligned}\lim_{\text{k.a.c.}}\mathrm{Var}(Y') = \lim_{\substack{p\to\infty\\ n_0\to\infty\\ n_1\to\infty}}&\left[\delta_p^2\left(1+b^2+\frac{1}{2n_0}+\frac{1}{n_1}\right)\right.\\ &\left.+\,p\left\{\frac{1}{2n_0n_1}+\frac{1}{2n_1^2}+(1+b^2)\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\right\}\right]\\ = (1+b^2)(\delta^2+J_0+J_1)\,,&\end{aligned}\tag{3.16}$$

and all the conditions of Theorem 3.1 hold. Noting the fact that

$$\begin{aligned}\lim_{\text{k.a.c.}}\mathrm{E}(Y') &= \lim_{\substack{p\to\infty\\ \nu_1\to\infty,\lambda_1\to 0\\ \nu_2\to\infty,\lambda_2\to 0}}\lambda_1(p+\nu_1)+\lambda_2(p+\nu_2)\\ &= \lim_{\substack{p\to\infty\\ n_0\to\infty\\ n_1\to\infty}}\frac{p}{2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)+\frac{\delta_p^2}{2} = \frac{1}{2}(\delta^2+J_0+J_1)\ ,\end{aligned}$$

then the results follow by using (3.4) from Theorem 3.1. Comparison of the first moment of $\hat{\epsilon}_{sr}$ and $\hat{\epsilon}_r$ can be easily done from results presented by Zollanvari et al. (2011) or a simplified version of the proof using the FTKCLDA presented in Section 4.

The simplicity and accuracy of the equations in (4.3) and (3.15) let us easily find the optimal value of b that makes the smoothed resubstitution approximately unbiased. For simplicity, we assume that $n_0 = n_1 = n$ and $\alpha_0 = \alpha_1$ in (4.3). From equating (4.3) and (3.15), the theoretical "optimal" amount of smoothing that produces an unbiased estimator is:

$$b_{\text{opt}} \;=\; \frac{1}{\delta^2}\sqrt{(J_0 + J_1)\,(J_0 + J_1 + 2\delta^2)}\,. \tag{3.17}$$

Not surprisingly, the value of b_{opt} is large for small ratios, n/p, whereas it becomes close to zero (no smoothing) for large ratios, n/p.

4 Proof of several known results using FTKCLDA

The results presented in Theorem 4.1 below are the same as those proved in Zollanvari et al. (2011) though by a different approach. Here, we show that the FTKCLDA can be applied as before to obtain these results. Also, as we explained in Section 1, the result of Theorem 4.2 is a special case of what has been considered previously in the literature for the LDA true error in a general case of the unknown covariance matrix by Deev (1970, 1972), Fujikoshi (2000), Meshalkin and Serdobolskii (1978), and Serdobolskii (2000). While the main contribution of the current paper is to introduce the FTKCLDA and its application in achieving the asymptotic expectation of *estimators* of the true error of misclassification in LDA with a known covariance matrix, the case of the true error is included to show the comprehensiveness of the FTKCLDA.

4.1. The resubstitution error estimator

Theorem 4.1. *In the sequence of Gaussian discrimination problems defined by (3.9), under the Kolmogorov asymptotic conditions, we have:*

$$W(X_1) \xrightarrow{D} N\left(\frac{1}{2}(\delta^2 + J_0 + J_1), \delta^2 + J_0 + J_1\right),$$

$$\lim_{k.a.c.} \mathrm{E}(\hat{\epsilon}_r) = \Phi\left(-\frac{1}{2}\frac{\delta^2 + J_1 + J_0}{\sqrt{\delta^2 + J_0 + J_1}}\right). \tag{4.1}$$

Proof. In Zollanvari, Braga-Neto and Dougherty (2009), we showed that $W(X_1)$ is represented statistically by

$$W(X_1) \;=\; \frac{1}{2}\,(\,\lambda_1 Z_1 \;+\; \lambda_2 Z_2\,),$$

where $Z_1 \sim \chi^2_p(\nu_1)$ and $Z_2 \sim \chi^2_p(\nu_2)$ are independent and

$$\nu_i \;=\; \frac{n_0 n_1}{2(n_0 + n_1)}\left\{1 \;+\; (-1)^{i+1}\sqrt{\frac{n_0 + n_1}{4n_0 n_1 + n_0 - 3n_1}}\right\}\delta_p^2, \quad i = 1, 2,$$

and

$$\lambda_i = \tfrac{1}{2n_0n_1}\left\{n_0+n_1+(-1)^{i+1}\sqrt{(n_0+n_1)(4n_0n_1+n_0-3n_1)}\right\}, \quad i=1,2.$$

As in (3.12) and (3.13), it can easily be seen that the limit of $(\lambda_1+\lambda_2)p$ and $\lambda_1\nu_1+\lambda_2\nu_2$ under the Kolmogorov asymptotic conditions converges to a finite constant. Then, by taking $a_1=1/2$ and $a_2=1$, it follows that:

$$\begin{aligned}\lim_{\text{k.a.c.}} \mathrm{Var}\{W(X_1)\} &= \lim_{\substack{p\to\infty\\ \nu_1\to\infty,\lambda_1\to 0\\ \nu_2\to\infty,\lambda_2\to 0}} \left\{\frac{1}{4}\lambda_1^2(2p+4\nu_1)+\frac{1}{4}\lambda_2^2(2p+4\nu_2)\right\}\\ &= \lim_{\substack{p\to\infty\\ n_0\to\infty\\ n_1\to\infty}} \left\{\delta_p^2\left(1+\frac{1}{n_1}\right)+p\left(\frac{1}{n_0}+\frac{1}{n_1}+\frac{1}{2n_1^2}-\frac{1}{2n_0^2}\right)\right\}\\ &= (\delta^2+J_0+J_1),\end{aligned} \tag{4.2}$$

and all the conditions of Theorem 3.1 hold. Noting the fact that

$$\begin{aligned}\lim_{\text{k.a.c.}} \mathrm{E}\{W(X_1)\} &= \lim_{\substack{p\to\infty\\ \nu_1\to\infty,\lambda_1\to 0\\ \nu_2\to\infty,\lambda_2\to 0}} \frac{1}{2}\lambda_1(p+\nu_1)+\frac{1}{2}\lambda_2(p+\nu_2)\\ &= \lim_{\substack{p\to\infty\\ n_0\to\infty\\ n_1\to\infty}} \frac{p}{2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)+\frac{\delta_p^2}{2} = \frac{1}{2}(\delta^2+J_0+J_1),\end{aligned}$$

the results follow by using (3.4) from Theorem 3.1.

4.2. The true error

THEOREM 4.2. *In the sequence of Gaussian discrimination problems defined by (3.9), under the Kolmogorov asymptotic conditions and for $X\in\Pi_0$, we have:*

$$W(X) \xrightarrow{D} N\left(\frac{1}{2}(\delta^2+J_1-J_0), \delta^2+J_0+J_1\right),$$

$$\begin{aligned}\lim_{k.a.c.} \mathrm{E}(\epsilon) = \alpha_0\Phi&\left(-\frac{1}{2}\frac{\delta^2+J_1-J_0}{\sqrt{\delta^2+J_0+J_1}}\right)\\ &+\alpha_1\Phi\left(-\frac{1}{2}\frac{\delta^2+J_0-J_1}{\sqrt{\delta^2+J_0+J_1}}\right).\end{aligned} \tag{4.3}$$

Proof. From Zollanvari et al. (2009), it can be easily shown that $W(X)$ is represented statistically by

$$W(X) = \frac{1}{2}(\lambda_1 Z_1 + \lambda_2 Z_2),$$

where $Z_1 \sim \chi_p^2(\nu_1)$ and $Z_1 \sim \chi_p^2(\nu_2)$ are independent and for $i = 1, 2$:

$$\begin{aligned}\nu_i &= \frac{n_0 n_1 + n_1}{2(n_0 + n_1 + 1)} \left\{ 1 + (-1)^{i+1} \sqrt{\frac{n_0 + n_1}{4n_0 n_1 + n_0 + n_1}} \right. \\ &\quad \left. + (-1)^{i+1} \frac{2n_0 n_1}{(n_0 + 1)\sqrt{(n_0 + n_1)(4n_0 n_1 + n_0 + n_1)}} \right\} \delta_p^2\end{aligned}$$

and

$$\lambda_i = \frac{1}{2n_0 n_1} \left\{ n_0 - n_1 + (-1)^{i+1} \sqrt{(n_0 + n_1)(4n_0 n_1 + n_0 + n_1)} \right\}, \quad i = 1, 2.$$

As in (3.12) and (3.13), it can easily be seen that the limit of $(\lambda_1 + \lambda_2)p$ and $\lambda_1 \nu_1 + \lambda_2 \nu_2$ under the Kolmogorov asymptotic conditions converges to a finite constant. Then, by taking $a_1 = 1/2$ and $a_2 = 1$ and similar to (4.2), we have:

$$\begin{aligned}\lim_{\text{k.a.c.}} \text{Var}\{W(X)\} &= \lim_{\substack{p \to \infty \\ n_0 \to \infty \\ n_1 \to \infty}} \left\{ \delta_p^2 \left(1 + \frac{1}{n_1} \right) + p \left(\frac{1}{n_0} + \frac{1}{n_1} + \frac{1}{2n_1^2} + \frac{1}{2n_0^2} \right) \right\} \\ &= (\delta^2 + J_0 + J_1),\end{aligned} \tag{4.4}$$

and all the conditions of Theorem 3.1 hold. Noting the fact that

$$\begin{aligned}\lim_{\text{k.a.c.}} \text{E}\{W(X)\} &= \lim_{\substack{p \to \infty \\ \nu_1 \to \infty, \lambda_1 \to 0 \\ \nu_2 \to \infty, \lambda_2 \to 0}} \frac{1}{2}\lambda_1 (p + \nu_1) + \frac{1}{2}\lambda_2 (p + \nu_2) \\ &= \lim_{\substack{p \to \infty \\ n_0 \to \infty \\ n_1 \to \infty}} \frac{p}{2} \left(\frac{1}{n_1} - \frac{1}{n_0} \right) + \frac{\delta_p^2}{2} = \frac{1}{2}(\delta^2 + J_1 - J_0),\end{aligned}$$

the results follow by using (3.4) from Theorem 3.1.

5 Numerical examples and discussion

5.1. Finite sample approximation for the expectation of ϵ, $\hat{\epsilon}_r$, $\hat{\epsilon}_p$ and $\hat{\epsilon}_{sr}$
To simplify the notation, the following function is defined:

$$\begin{aligned}f\{a, h(n_0, n_1, J_0, J_1), \alpha_0, \alpha_1\} &= \alpha_0 \Phi \left\{ -\frac{1}{2} \frac{a}{\sqrt{h(n_0, n_1, J_0, J_1)}} \right\} \\ &\quad + \alpha_1 \Phi \left\{ -\frac{1}{2} \frac{a}{\sqrt{h(n_1, n_0, J_1, J_0)}} \right\}.\end{aligned}$$

Under the Kolmogorov asymptotic conditions, Theorems 3.2, 3.3, 3.4 and 4.2 give asymptotically exact values of $\mathrm{E}(\epsilon)$, $\mathrm{E}(\hat{\epsilon}_r)$, $\mathrm{E}(\hat{\epsilon}_p)$ and $\mathrm{E}(\hat{\epsilon}_{sr})$, respectively, that could be used in their finite sample approximations by simply replacing $J_i = p/n_i$, $i = 0, 1$ and $\delta = \sqrt{(\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1)}$. On the other hand, we can use slightly different formulas by replacing the denominator in (3.10), (3.15), (4.1), and (4.3) by that in (3.14) and (3.16), (4.2), (4.4), respectively. That is, we can use the following formulas as well:

$$\begin{aligned}
\mathrm{E}(\epsilon) &\simeq f\Big(\delta^2 + J_1 - J_0, \delta^2\left(1 + \frac{1}{n_1}\right) + J_0 + J_1 + \frac{J_1}{2n_1} + \frac{J_0}{2n_0}, \alpha_0, \alpha_1\Big), \\
\mathrm{E}(\hat{\epsilon}_r) &\simeq f\Big(\delta^2 + J_1 + J_0, \delta^2\left(1 + \frac{1}{n_1}\right) + J_0 + J_1 + \frac{J_1}{2n_1} + \frac{J_0}{2n_0}, \hat{\alpha}_0, \hat{\alpha}_1\Big), \\
\mathrm{E}(\hat{\epsilon}_p) &\simeq f\Big(\delta^2 + J_1 + J_0, \delta^2\left(1 + \frac{1}{n_0} + \frac{1}{n_1}\right) + J_0 + J_1 \\
&\qquad + \frac{J_0}{2n_0} + \frac{J_1}{2n_1} + \frac{J_0}{n_1}, \hat{\alpha}_0, \hat{\alpha}_1\Big), \\
\mathrm{E}(\hat{\epsilon}_{sr}) &\simeq f\left(\delta^2 + J_1 + J_0, \delta^2\left(1 + b^2 + \frac{1}{2n_0} + \frac{1}{n_1}\right)\right. \\
&\qquad \left. + \left\{\frac{J_0}{2n_1} + \frac{J_1}{2n_1} + (1 + b^2)(J_0 + J_1)\right\}, \hat{\alpha}_0, \hat{\alpha}_1\right),
\end{aligned} \tag{5.1}$$

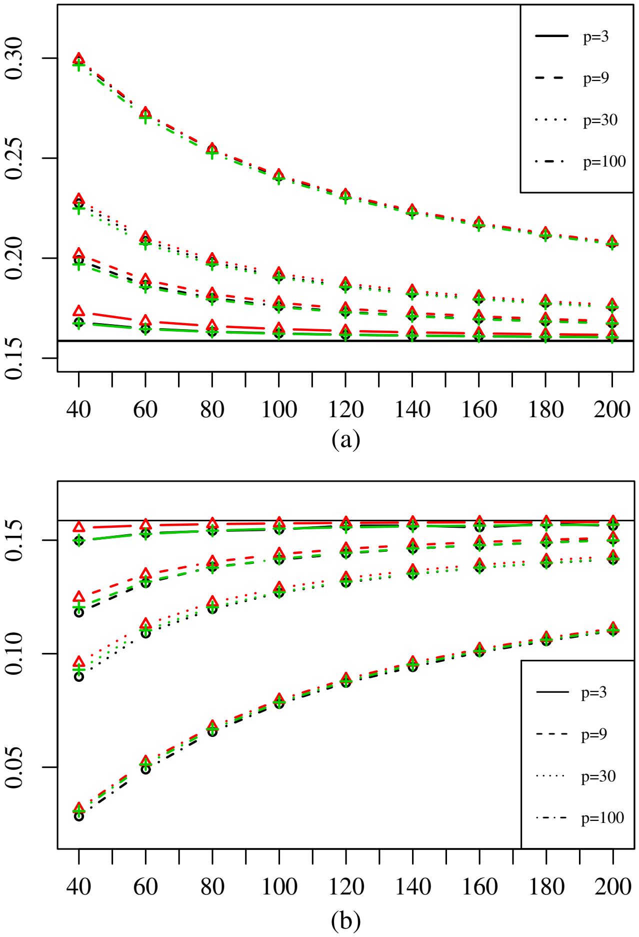
where we have $J_i = p/n_i$, $i = 0, 1$ and $\delta = \sqrt{(\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1)}$. The only difference between these equations and those finite sample approximations based on (3.10), (3.15), (4.1), and (4.3) is in the denominator. That is, $\mathrm{Var}_\infty\{\cdot\}$ is replaced simply by $\mathrm{Var}\{\cdot\}$ where $\{\cdot\}$ are the proper random variables that were used in the Theorems 2, 3, 4 and 6 to find these quantities. This is the counterpart of the Raudys approximation for different estimators of interest. In fact, since under the Kolmogorov asymptotic conditions we have $\mathrm{Var}\{\cdot\} - \mathrm{Var}_\infty\{\cdot\} \to 0$, one is free asymptotically to use the set of equations in (5.1) or those of (4.3), (4.1), (3.10) and (3.15). However, in finite sample approximations, they may have a slightly different performance; we investigate this possibility in this section. In order to distinguish between these two sets of approximations in the following figures, we call the set of equations in (5.1), *the redundant-finite sample approximation* and those of equations (3.10), (3.15), (4.1), and (4.3), *the asymptotic-finite sample approximation.* The main reason to consider the redundant-finite sample approximation is that, in fact, the Raudys formulas are all of this type approximation while the formulas developed by Deev, Serdobolskii and Meshalkin for $\mathrm{E}(\epsilon)$ are all asymptotic-finite sample approximations (Serdobolskii, 2000).

5.2. Numerical results To test the accuracy of the redundant-finite sample approximation and asymptotic-finite sample approximation, we estimate the expectation of the true error and its estimators for LDA classification with a known covariance matrix using Monte-Carlo simulations. The model that we consider consists of two Gaussian distributions with different means but equal covariance matrices having $\delta^2 = 4$, which corresponds to a Bayes error of 0.1586.

From Figures 1–3, it is easy to see that the redundant-finite sample approximation and the asymptotic-finite sample approximation agree very well with Monte Carlo estimation of the first moments even in dimensions that are much higher than the sample size (notice that since the covariance matrix is assumed to be known, these high dimensions are legitimate). The figures show that the two types of analytical results mentioned above perform similarly. Therefore, it is clear that the important terms determining the values of $\mathrm{E}(\epsilon)$, $\mathrm{E}(\hat{\epsilon}_r)$, $\mathrm{E}(\hat{\epsilon}_p)$ and $\mathrm{E}(\hat{\epsilon}_{sr})$ are singled out in the asymptotic-finite sample approximation. It is noteworthy to mention that in Figure 2b, we have used the known value of δ^2 for computing (3.17). This choice of δ^2 is used here to show the accuracy of the finite sample approximation and eliminate the randomness of $\hat{\delta}^2$. However, in practice, in order to use (3.17) for finding b_{opt}, an estimator of δ^2, namely $\hat{\delta}^2$, is needed. It is interesting to compare Figure 2b with Figure 1a to see how well the expectation of smoothed resubstitution with the choice of optimal smoothing agrees with the expectation of true error. Figure 3 is similar to Figure 2b in which the difference between the Monte Carlo estimate and the proposed finite sample approximations for $\mathrm{E}(\hat{\epsilon}_{sr})$ is shown; however, instead of using b_{opt} from (3.17), we used the choice of smoothing parameter proposed by Snapinn and Knoke (1985).

5.3. Discussion The simplicity of equations (3.10), (3.15), (4.1), and (4.3) makes them a proper tool to interpret the effect of dimensionality, sample size and the distance of classes on the overall behavior of the statistics. For simplicity, we assume that the sample size of the classes are equal and hence $J_0 = J_1$. Therefore, we have: $Bias(\hat{\epsilon}_r) = \mathrm{E}(\hat{\epsilon}_r) - \mathrm{E}(\epsilon)$ and from the asymptotic approximation stated in (4.3) and (4.1), we have:

$$Bias(\hat{\epsilon}_r) = \Phi\left(-\frac{1}{2}\frac{\delta^2 + 2J_0}{\sqrt{\delta^2 + 2J_0}}\right) - \Phi\left(-\frac{1}{2}\frac{\delta^2}{\sqrt{\delta^2 + 2J_0}}\right) < 0\,.$$

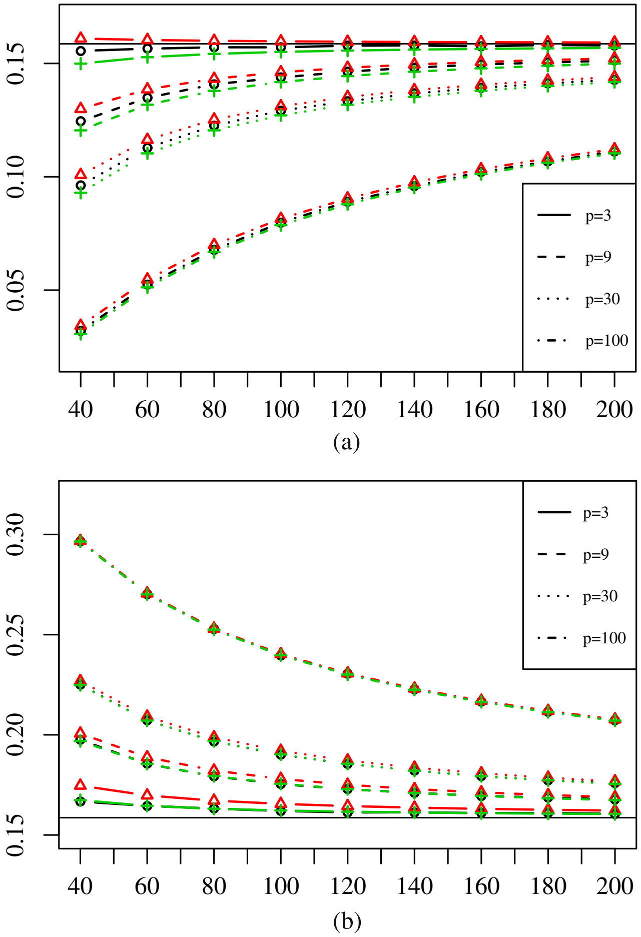


(a)

(b)

Figure 1: (a) $\mathrm{E}(\epsilon)$ (b) $\mathrm{E}(\hat{\epsilon}_r)$ as functions of the sample size for $n_0 = n_1 = n/2$ and for different dimensionality, p. ($\circ$) Monte Carlo, ($\triangle$) redundant-finite sample approximation, ($+$) asymptotic-finite sample approximation. The solid line is the $Bayes\ error = 0.1586$.

This shows that resubstitution is indeed an optimistic estimator of the true error. However, we have to mention that this well-known conclusion is asymptotically exact under the Kolmogorov asymptotic conditions. On the other hand, in Theorem 3, we showed that $\mathrm{E}(\hat{\epsilon}_r)$ and $\mathrm{E}(\hat{\epsilon}_p)$ are asymptotically the same. That is, $\lim_{\text{k.a.c.}} \mathrm{E}(\hat{\epsilon}_p) = \lim_{\text{k.a.c.}} \mathrm{E}(\hat{\epsilon}_r)$. This fact has been reported by simulations in the literature; e.g., see Lachenbruch and Mickey (1968) and Moran (1975). In Theorem 4.2, we showed that



(a)

(b)

Figure 2: (a) $\mathrm{E}(\hat{\epsilon}_p)$ (b) $\mathrm{E}(\hat{\epsilon}_{sr})$ as functions of the sample size for $n_0 = n_1 = n/2$ and for different dimensionality, p. ($\circ$) Monte Carlo, ($\triangle$) redundant-finite sample approximation, ($+$) asymptotic-finite sample approximation. The solid line is the *Bayes error* $= 0.1586$. The smoothing parameter in (b) has been chosen according to (3.17).

$\lim_{\text{k.a.c.}} \mathrm{E}(\hat{\epsilon}_{sr}) > \lim_{\text{k.a.c.}} \mathrm{E}(\hat{\epsilon}_r)$. This is perhaps one of the reasons that smoothed resubstitution was proposed by Glick (1978), i.e., to penalize the optimistic bias of resubstitution. However, as stated before, theoretically, we need to use the proposed b_{opt} to make $\hat{\epsilon}_{sr}$ an unbiased estimator of ϵ. The value of b_{opt} depends on the true parameters of the underlying distribution of classes, namely the Mahalanobis distance, which needs to be estimated in

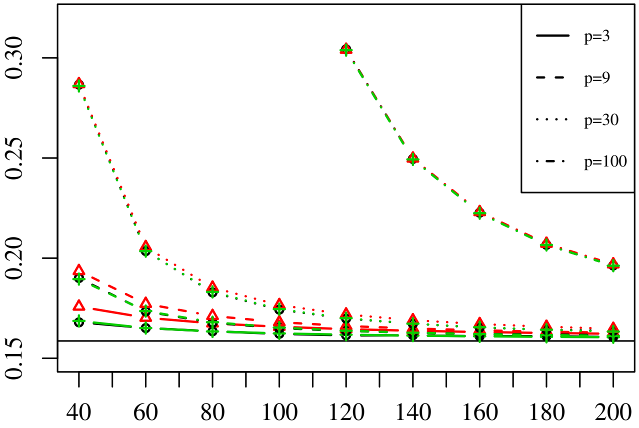


Figure 3: $\mathrm{E}(\hat{\epsilon}_{sr})$ as functions of the sample size for $n_0 = n_1 = n/2$ and for different dimensionality, p. ($\circ$) Monte Carlo, ($\triangle$) redundant-finite sample approximation, ($+$) asymptotic-finite sample approximation. The solid line is the *Bayes error* $= 0.1586$. The smoothing parameter has been chosen according to Snapinn and Knoke (1985). This choice of smoothing parameter is not computable for $n < p + 4$.

practice. In some similar situations where a statistic in the pattern recognition context depends on the Mahalanobis distance, it is suggested to use different estimators of this measure as in McLachlan (1992).

6 Conclusion

In this article, we presented the fundamental theorem of known covariance LDA (FTKCLDA) in conjunction with the so-called Kolmogorov asymptotic approach to study the behavior of different statistics of interest in linear discriminant analysis. This gives new insight into the Kolmogorov asymptotic approach by proposing a novel proof for obtaining the first moment of various estimators of the misclassification error rate as well as the distribution of the discriminant itself.

In this study, we showed analytically that in situations where the dimension of the problem is comparable to the sample size, the plug-in error estimator has an asymptotically (in terms of Kolmogorov) optimistic bias that is equivalent to that of the apparent error, something that has been confirmed by simulation studies before. In addition, we showed that in similar scenarios, the expectation of smoothed resubstitution is larger than that of resubstitution. However, to prevent optimistic or even pessimistic bias of this estimator, the amount of smoothing should be controlled as suggested, giving us the opportunity to find an optimal smoothing parameter for this estimator. The performance of the smoothed resubstitution with the choice of the optimal smoothing parameter, which itself depends on the Mahalanobis distance, remains for future studies.

In this work, we assumed that the covariance matrix used in the representation of the linear discriminant analysis is known. While this is a limitation of the results of the current work, this assumption has been historically made in studying various discriminants as a cornerstone of more complex settings. Therefore, the next logical step is to extend the results presented here to a more general setting where the covariance matrix is unknown. Another direction for future studies is to extend the results presented here to other estimators of the LDA error rate or studying the expected error rate of other classifiers and the performance of their estimators.

Acknowledgement. This research was partially supported by NSF grants DMS-1007504 and DMS-1100492, and Award No. KUS-C1-016-04 made by King Abdullah University of Science and Technology (KAUST).

References

ANDERSON, T. (1951). Classification by multivariate analysis. *Psychometrika*, **16**, 31–50.

ANDERSON, T. (1973). An asymptotic expansion of the distribution of the studentized classification statistic w. *Ann. Statist.*, **1**, 964–972.

CONTE, E., LOPS, M., and RICCI, G. (1996). Adaptive matched filter detection in spherically invariant noise. *IEEE Signal Process. Lett.*, **3**, 248–250.

DEEV, A. (1970). Representation of statistics of discriminant analysis and asymptotic expansion when space dimensions are comparable with sample size. *Dokl. Akad. Nauk SSSR*, **195**, 759–762 (in Russian).

DEEV, A. (1972). Asymptotic expansions for distributions of statistics w, m, and w* in discriminant analysis. *Statist. Methods Class.*, **31**, 6–57 (in Russian).

DOUGHERTY, E.R. (2008). On the epistemological crisis in genomics. *Curr. Genomics*, **9**, 69–79.

DOUGHERTY, E.R., HUA, J., and BITTNER, M. (2007). Validation of computational methods in genomics. *Curr. Genomics*, **8**, 1–19.

DUNN, O.J. (1971). Some expected values for probabilities of correct classification in discriminant analysis. *Technometrics*, **13**, 345–353.

DUPUY, A. and SIMON, R. (2008). Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J. Natl. Cancer Inst.*, **99**, 147–157.

EFRON, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *J. Amer. Statist. Assoc.*, **70**, 892–898.

EFRON, B. (1980). The distributions of the actual error rates in linear discriminant analysis. *J. Amer. Statist. Assoc.*, **75**, 201–205.

FISHER, R. (1936). The use of multiple measurements in taxonomic problems. *Ann. Euge.*, **7**, 79–188.

FISHER, R. (1940). The precision of discriminant function. *Ann. Euge.*, **10**, 422–429.

FISHER, R.A. (1925). *Statistical methods for research workers*, 14th edn. Oliver & Boyd, Edinburgh. The quotation is from the preface to the first (1925) edition.

FOLEY, D. (1972). Considerations of sample and feature size. *IEEE Trans. Inform. Theory*, **IT-18**, 618–626.

FUJIKOSHI, Y. (2000). Error bounds for asymptotic approximations of the linear discriminant function when the sample sizes and dimensionality are large. *J. Multivariate Anal.*, **73**, 1–17.

FUJIKOSHI, Y. and SEO, T. (1998). Asymptotic approximations for epmc's of the linear and the quadratic discriminant functions when the samples sizes and the dimension are large. *Statist. Anal. Random Arrays*, **6**, 269–280.

FUKUNAGA, K. and HAYES, R.R. (1989). Estimation of classifier performance. *IEEE Trans. Pattern Anal. Mach. Intell.*, **11**, 1087–1101.

GEVAERT, O., SMET, F.D., GORP, T.V., POCHET, N., ENGELEN, K., AMANT, F., MOOR, B.D., TIMMERMAN, D. and VERGOTE, I. (2008). Expression profiling to predict the clinical behaviour of ovarian cancer fails independent evaluation. *BMC Cancer*, **8**, 1–10.

GLICK, N. (1978). Additive estimators for probabilities of correct classification. *Pattern Recognit.*, **10**, 211–222.

HAND, D. (1986). Recent advances in error rate estimation. *Pattern Recognit. Lett.*, **4**, 335–346.

HILLS, M. (1966). Allocation rules and their error rates. *J. R. Stat. Soc. Ser. B (Methodological)*, **28**, 1–31.

HIRST, D. (1996). Error-rate estimation in multiple-group linear discriminant analysis. *Technometrics*, **38**, 389–399.

JOHN, S. (1961). Errors in discrimination. *Ann. Math. Stat.*, **32**, 1125–1144.

JOHNSON, N., KOTZ, S. and BALAKRISHNAN, N. (1994.). *Continuous univariate distributions*. John Wiley, New York.

KIM, S., DOUGHERTY, E.R., SHMULEVICH, I., HESS, K.R., HAMILTON, S.R., TRENT, J.M., FULLER, G.N. and ZHANG, W. (2002). Identification of combination gene sets for glioma classification. *Mol. Cancer Ther.*, **1**, 1229–1236.

KITTLER, J. and DEVIJVER, P. (1982). Statistical properties of error estimators in performance assessment of recognition systems. *IEEE Trans. Pattern Anal. Mach. Intell.*, **4**, 215–220.

LACHENBRUCH, P. and MICKEY, M. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, **10**, 1–11.

MARTIN, J.K. and HIRSCHBERG, D.S. (1996). Small Sample Statistics for Classification Error Rates II: Confidence Intervals and Significance Tests. Tech. Rep. 96-22, University of California, Irvine, CA.

MCLACHLAN, G.J. (1973). An asymptotic expansion of the expectation of the estimated error rate in discriminant analysis. *Aust. J. Statistics*, **15**, 210–214.

MCLACHLAN, G.J. (1974). Estimation of the errors of misclassification on the criterion of asymptotic mean square error. *Technometrics*, **16**, 255–260.

MCLACHLAN, G.J. (1976). The bias of the apparent error in discriminant analysis. *Biometrika*, **63**, 239–244.

MCLACHLAN, G.J. (1992). *Discriminant analysis and statistical pattern recognition.* John Wiley, New York.

MESHALKIN, L.D. and SERDOBOLSKII, V.I. (1978). Errors in the classification of multivariate observations. *Theory Probab. Appl.*, **23**, 741–750.

MICHIELS, C.H.S. and KOSCIELNY, S. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*, **365**, 488–492.

MORAN, M. (1975). On the expectation of errors of allocation associated with a linear discriminant function. *Biometrika*, **62**, 141–148.

MOREIRA, M. (2009). Tests with correct size when instruments can be arbitrarily weak. *J. Econometrics*, **152**, 131–140.

OKAMOTO, M. (1963). An asymptotic expansion for the distribution of the linear discriminant function. *Ann. Math. Stat.*, **34**, 1286–1301 (Correction: *Ann. Math. Stat.*, **39**, 1358–1359, 1968).

RAUDYS, S. (1967). On determining training sample size of a linear classifier. *Comput. Syst.*, **28**, 79–87 (in Russian).

RAUDYS, S. (1972). On the amount of a priori information in designing the classification algorithm. *Tech. Cybern.*, **4**, 168–174 (in Russian).

RAUDYS, S. (1978). Comparison of the Estimates of the Probability of Misclassification. In *Proc. International Joint Conference on Pattern Recognition*, pp 280–282.

RAUDYS, S. (1998). Expected classification error of the fisher linear classifier with pseudoinverse covariance matrix. *Pattern Recognit. Lett.*, **19**, 385–392.

RAUDYS, S. and JAIN, A.K. (1991). Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans. Pattern Anal. Mach. Intell.*, **13**, 252–264.

RAUDYS, S. and PIKELIS, V. (1980). On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, **2**, 242–252.

RAUDYS, S. and SKURIKHINA, M. (1995). Small-sample properties of ridge-estimate of the covariance matrix in statistical and neural net classification. *Multivariate Statist.*, 237–245.

SCHIAVO, R.A. and HAND, D.J. (2000). Ten more years of error rate research. *Internat. Statist. Rev.*, **68**, 295–310.

SERDOBOLSKII, V. (2000). *Multivariate statistical analysis: a high-dimensional approach.* Kluwer Academic Publishers, Netherlands.

SERDOBOLSKII, V.I. (1979). The Moments of Discriminant Function and Classification for a Large Number of Variables. In *Statistical Problems of Control* (S. Raudys, ed.). Inst. of Math. and Cyb. Press, Vilnius, pp 27–51, in Russian.

SMITH, C. (1947). Some examples of discrimination. *Ann. Euge.*, **18**, 272–282.

SNAPINN, S. and KNOKE, J. (1985). An evaluation of smoothed classification error-rate estimators. *Technometrics*, **27**, 199–206.

SNAPINN, S. and KNOKE, J. (1989). Estimation of error rates in discriminant analysis with selection of variables. *Biometrics*, **45**, 289–299.

SORUM, M.J. (1971). Estimating the conditional probability of misclassification. *Technometrics*, **13**, 333–343.

SORUM, M.J. (1972). Estimating the expected and the optimal probabilities of misclassification. *Technometrics*, **14**, 935–943.

SORUM, M.J. (1973). Estimating the expected probability of misclassification for a rule based on the linear discriminant function: Univariate normal case. *Technometrics*, **15**, 329–339.

SWETS, D. and WENG, J. (1996). Using discriminant eigenfeatures for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, **18**, 891–896.

VAN VUUREN, S. and HERMANSKY, H. (1997). Data-driven design of rasta-like filters. In: *Proc. Eurospeech*, 1607–1610.

WACKER, A. and EL-SHEIKH, T. (1984). Average classification accuracy over collections of gaussian problems—common covariance matrix case. *Pattern Recognit.*, **17**, 259–274.

WALD, A. (1944). On a statistical problem arising in the classification of an individual into one of two groups. *Ann. Math. Stat.*, **15**, 145–162.

WIGNER, E.P. (1958). On the distribution of the roots of certain symmetric matrices. *Ann. Math.*, **67**, 325–327.

ZOLLANVARI, A., BRAGA-NETO, U. and DOUGHERTY, E. (2009). On the sampling distribution of resubstitution and leave-one-out error estimators for linear classifiers. *Pattern Recognit.*, **42**, 2705–2723.

ZOLLANVARI, A., BRAGA-NETO, U. and DOUGHERTY, E. (2011). Analytic study of performance of error estimators for linear discriminant analysis. *IEEE Trans. Signal Process.*, **59**, 4238–4255.

Amin Zollanvari
Department of Statistics
and Department of Electrical
and Computer Engineering
Texas A&M University
College Station, TX 77843, USA
E-mail: amin_zoll@neo.tamu.edu

Marc G. Genton
CEMSE Division, King Abdullah
University of Science and Technology
Thuwal 23955-6900, Saudi Arabia
E-mail: marc.genton@kaust.edu.sa

Paper received: 16 August 2012.