ANALYTIC STUDY OF PERFORMANCE OF ERROR ESTIMATORS FOR LINEAR DISCRIMINANT ANALYSIS WITH APPLICATIONS IN GENOMICS

A Dissertation

by

AMIN ZOLLANVARI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2010

Major Subject: Electrical Engineering

ANALYTIC STUDY OF PERFORMANCE OF ERROR ESTIMATORS FOR LINEAR DISCRIMINANT ANALYSIS WITH APPLICATIONS IN GENOMICS

A Dissertation

by

AMIN ZOLLANVARI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Ulisses M. Braga-Neto
	Edward R. Dougherty
Committee Members,	Aniruddha Datta
	Guy L. Curry
Head of Department,	Costas N. Georghiades

December 2010

Major Subject: Electrical Engineering

ABSTRACT

Analytic Study of Performance of Error Estimators for Linear Discriminant Analysis with Applications in Genomics. (December 2010)

Amin Zollanvari, B.S., Shiraz University, Iran;
M.S., Shiraz University, Iran

Co–Chairs of Advisory Committee: Ulisses M. Braga-Neto
Edward R. Dougherty

Error estimation must be used to find the accuracy of a designed classifier, an issue that is critical in biomarker discovery for disease diagnosis and prognosis in genomics and proteomics. This dissertation is concerned with the analytical formulation of the joint distribution of the true error of misclassification and two of its commonly used estimators, resubstitution and leave-one-out, as well as their marginal and mixed moments, in the context of the Linear Discriminant Analysis (LDA) classification rule.

In the first part of this dissertation, we obtain the joint sampling distribution of the actual and estimated errors under a general parametric Gaussian assumption. Exact results are provided in the univariate case and an accurate approximation is obtained in the multivariate case. We show how these results can be applied in the computation of conditional bounds and the regression of the actual error, given the observed error estimate. In practice the unknown parameters of the Gaussian distributions, which figure in the expressions, are not known and need to be estimated. Using the usual maximum-likelihood estimates for such parameters and plugging them into the theoretical exact expressions provides a sample-based approximation to the joint distribution, and also sample-based methods to estimate upper conditional bounds.

In the second part of this dissertation, exact analytical expressions for the bias, variance, and Root Mean Square (RMS) for the resubstitution and leave-one-out error estimators in the univariate Gaussian model are derived. All probabilistic characteristics of an error estimator are given by the knowledge of its joint distribution with the true error. Partial information is contained in their mixed moments, in particular, their second mixed moment. Marginal information regarding an error estimator is contained in its marginal moments, in particular, its mean and variance. Since we are interested in estimator accuracy and wish to use the RMS to measure that accuracy, we desire knowledge of the second-order moments, marginal and mixed, with the true error. In the multivariate case, using the double asymptotic approach with the assumption of knowing the common covariance matrix of the Gaussian model, analytical expressions for the first moments, second moments, and mixed moment with the actual error for the resubstitution and leave-one-out error estimators are derived. The results provide accurate small sample approximations and this is demonstrated in the present situation via numerical comparisons. Application of the results is discussed in the context of genomics.

To Ghazal

ACKNOWLEDGMENTS

I would like to sincerely and gratefully thank my advisors Dr. Ulisses M. Braga-Neto and Dr. Edward R. Dougherty. Many thanks to Dr. Braga-Neto for his great guidance in research, encouragement, and support he provided. He helped me not only to pursue my interests in research, but also helped me to believe in myself. He generously listened to my ideas and taught me many new concepts for which I am grateful. My special thanks to Dr. Dougherty for his understanding, guidance throughout the research, sharing new ideas, and all the support he provided. I feel extremely privileged to have had the opportunity to work with him; a great scientist, a wonderful teacher, and amazing mentor. His patience let me pass the hardest time I had during these years. I am also grateful to my committee members, Dr. Aniruddha Datta for giving me useful advice; and Dr. Guy L. Curry for serving on my committee.

I dedicate this dissertation to Ghazal, my lovely wife and my best friend, who constantly supported me in all ups and downs. She is my inspiration in every aspect of life. Without her dedication, this work would have never been completed.

I am gratefully indebted to my parents, Shahnaz and Jafar, for their indispensable love, dedication, and encouragement throughout my life. They are a pillar of strength to me. Many thanks to my brother, Ali, and my sister, Noushin. Being with them is a breath of fresh air. Last but not the least, I would like to express my gratitude to Dr. Nader Ghahramani who has been a paragon to me.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

CHAPTER I

ESTIMATION OF THE MISCLASSIFICATION ERROR RATE

Supervised learning is about predicting an output variable using some input variables. A continuous output variable results in a *regression* problem while a categorial output variable constructs a *classification* problem. A *Regression estimator* and a *classifier* are the two predictors used to accomplish these tasks, respectively. The performance of the designed predictor is assessed by how accurate it can predict the future samples. However, the accuracy of the predictor mostly depends on the underlying distribution of samples, which is usually unknown. This is where *error estimation* plays a significant role.

A. Classification Problem

Let $x \in \Re^p$ be a sample of p dimensions coming from one of the the $t \geq 2$ subgroups or classes Π_0, Π_1, $\ldots \Pi_t$ of population Π. Further, assume that we have a set of *training* samples; that is a set of samples that their classes are known. The problem of *classification* is to design a classifier, $\psi(x) : \Re^p \rightarrow \{0,1\}$ based on the training sample set to classify x into one of these subgroups. This problem is known as classification, discrimination or allocation [1]. In the case where there are two subgroups Π_0, Π_1, the problem is known as *binary classification*; we will refer to that case simply as classification throughout this dissertation. In this scenario, we will assume that $\{X_1, X_2, \ldots, X_{n_0}\}$ and $\{X_{n_0+1}, X_{n_0+2}, \ldots, X_{n_0+n_1}\}$ are training random samples from Π_0 and Π_1, respectively. We will assume that Π_0, Π_1 are described by probability density functions, namely the *class-conditional densities* $f(X = x|X \in \Pi_i), i = 0, 1$.

The journal model is *IEEE Transactions on Automatic Control.*

Depending on prior knowledge of the class-conditional densities, we may consider different problems of classification, as described next.

1. Complete Knowledge of Underlying Distributions

In this case, it is assumed that complete knowledge about $f(X|X \in \Pi_i), i = 0,1$ and the *prior* probabilities, $P(X \in \Pi_i) = \alpha_i, i = 0,1$, is available. The prior probabilities give the probability that a sample X taken from population Π belongs to Π_i (before seeing the specific value of X). For simplicity, we assume that prior probabilities are known. However, one may want to estimate it from the data in hand in which case they can be estimated by the frequency of the data in each class. Using the strong law of large numbers this estimator of prior probability converge to its true value with probability 1. Using Bayes theorem we can assign x to the class with higher *posterior* probability. Letting $\frac{\alpha_1}{\alpha_0} = c$, this is equivalent to the following *likelihood-ratio rule*:

$$\psi(x) = \begin{cases} 1\,, & \text{if } \dfrac{f(\Pi_0|X = x)}{f(\Pi_1|X = x)} < c \\ 0\,, & \text{if } \dfrac{f(\Pi_0|X = x)}{f(\Pi_1|X = x)} > c \end{cases} \tag{1.1}$$

In the boundary region where $\frac{f(\Pi_0|X = x)}{f(\Pi_1|X = x)} = c$, we can either do randomization or break the tie in favor of one region. The above solution constructs a complete class of admissible rules [2–4]. If c in the above formulation is known, then $\psi(x)$ minimizes the expected risk defined as the expectation of the probability of misallocating a member of Π_i. For the cases where c is not known, then other criteria should be chosen instead of the likelihood-ratio rule given in (1.1); for example, choosing c such that the probability of misallocating a member of Π_0 equals the probability of misallocating a member of Π_1. For other examples of criteria see [5].

2. Parametric Models

Parametric models are the result of partial knowledge about the class-conditional densities; they exist in different forms. One source of knowledge can be the general form of the distributions governing the problem, for example, Gaussian distributions [6,7], t-distributions [8], inverse normal distributions [9,10], elliptically contoured distributions [11], and skew normal distributions [12].

The knowledge can also be about the ratio of class-conditional densities. Depending on different assumptions made on the ratio of class-conditional densities, different family of discriminants have been proposed; for example, the linear logistic model [13], the quadratic logistic model [13], and the probit model [14].

Another source of knowledge commonly used appears in discrete data classification, in which the constraint on probabilities that states that their sum must be one results in the samples coming from a multinomial distribution; thus, we have made no assumption on the form of the distribution, and hence we cannot call multinomial model a parametric one. However, there have been various attempts to smooth the non-parametric estimates of the multinomial distribution; for example, by making the assumption of having independent features [15] or assuming a log-linear model [16]. These attempts of smoothing are categorized as parametric models for discrete data classification.

3. Non-parametric Models

There are many popular non-parametric methods of classification. These methods can be commonly categorized into three main types. One is based on density estimation of the class-conditional densities; another is based on optimization schemes, and the last one is based on tree classification approaches.

The first and the simplest rule based on density estimation is the multinomial discrimination rule [1, 17]. In this case, the continuous data is handled by discretizing it; however, at the expense of loss of discriminatory power [18]. Another widely used classification rule of this type is based on kernel density estimation of class-conditional densities. Depending on the nature of the data, different kernel based rules have been proposed. For the continuous data classification problem, normal and Cauchy kernels have been proposed in [19] and [20], respectively. For dealing with discrete data and for a mixture of discrete and continuous data see [21], and for handling missing data, see [22]. Two of the most popular classification rules, support vector machines and neural networks, are based on optimization schemes. There are many variants of these methods, which can be found in [23] and [24]. Two commonly used types of tree based rules are CART and binary space partition trees. The reader is referred to [23] for more information on these rules.

B. Linear Discriminant Analysis

Among all the classifiers mentioned in previous section, we are particularly concerned with *Linear Discriminant Analysis* (LDA), which was originally based on an idea from R. A. Fisher using the linear regression procedure [6, 7], and has a long history in statistics and pattern recognition. LDA was further developed by Wald [25] in the context of decision theory and then formulated by Anderson [26] in terms of what is known today as Anderson's statistic.

From the first use on taxonomic classification by R. A. Fisher [6], LDA-Fisher based classification and recognition systems have been applied in many disciplines such as speech recognition [27, 28], face recognition [29, 30] and recently in cancer classification [31, 32].

Here, 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,

$$W(\hat{\mu}_0, \hat{\mu}_1, X) = \left(X - \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}\right)^T \Sigma^{-1} (\hat{\mu}_0 - \hat{\mu}_1) \tag{1.2}$$

where $\hat{\mu}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} X_i$ and $\hat{\mu}_1 = \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} X_i$ are the sample means for each class.

The designed LDA classifier is given by

$$\psi(X) = \begin{cases} 1\,, & \text{if } W(\hat{\mu}_0, \hat{\mu}_1, X) \leq 0 \\ 0\,, & \text{if } W(\hat{\mu}_0, \hat{\mu}_1, X) > 0 \end{cases} \tag{1.3}$$

that is, the sign of W determines the classification of the sample point X. Throughout this dissertation, following for example [33–35], we are assuming 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 used as an estimator of Σ. Given the training data (and thus the sample means $\hat{\mu}_0$ and $\hat{\mu}_1$), the classification error is given by

$$\begin{aligned} \varepsilon &= P(W(\hat{\mu}_0, \hat{\mu}_1, X) \leq 0, X \in \Pi_0 \mid \hat{\mu}_0, \hat{\mu}_1) \;+\; P(W(\hat{\mu}_0, \hat{\mu}_1, X) > 0, X \in \Pi_1 \mid \hat{\mu}_0, \hat{\mu}_1) \\ &= \alpha_0 \varepsilon^0 \;+\; \alpha_1 \varepsilon^1 \end{aligned} \tag{1.4}$$

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

$$\varepsilon^0 = P(W(\hat{\mu}_0, \hat{\mu}_1, X) \leq 0 | X \in \Pi_0, \hat{\mu}_0, \hat{\mu}_1)\,, \quad \varepsilon^1 = P(W(\hat{\mu}_0, \hat{\mu}_1, X) > 0 | X \in \Pi_1, \hat{\mu}_0, \hat{\mu}_1) \tag{1.5}$$

and therefore,

$$\begin{aligned}\varepsilon \;=\; & \alpha_0\Phi\left(-\frac{\left(\mu_0 - \frac{1}{2}(\hat{\mu}_0 + \hat{\mu}_1)\right)^T \Sigma^{-1}(\hat{\mu}_0 - \hat{\mu}_1)}{\sqrt{(\hat{\mu}_0 - \hat{\mu}_1)^T\Sigma^{-1}(\hat{\mu}_0 - \hat{\mu}_1)}}\right) \\ & + \alpha_1\Phi\left(\frac{\left(\mu_1 - \frac{1}{2}(\hat{\mu}_0 + \hat{\mu}_1)\right)^T \Sigma^{-1}(\hat{\mu}_0 - \hat{\mu}_1)}{\sqrt{(\hat{\mu}_0 - \hat{\mu}_1)^T\Sigma^{-1}(\hat{\mu}_0 - \hat{\mu}_1)}}\right)\end{aligned} \tag{1.6}$$

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

$$\begin{aligned}\mathrm{E}[\varepsilon] &= \alpha_0\mathrm{E}[\varepsilon^0]+\alpha_1\mathrm{E}[\varepsilon^1] \\ &= \alpha_0 P(W(\hat{\mu}_0, \hat{\mu}_1, X) \leq 0|X \in \Pi_0)+\alpha_1 P(W(\hat{\mu}_0, \hat{\mu}_1, X) > 0|X \in \Pi_1)\end{aligned} \tag{1.7}$$

C. Error Estimation in Biomarker Discovery

Classifiers have the role of diagnostic and prognostic tools for cancer stratification; hence, it is of main concern to assess their predictive power. The successful applicability of a designed classifier relies on its predictive power, namely, the actual or true error [1, 35]. However, in practice it is almost always the case that one cannot evaluate exactly the true error of a designed classifier due to the lack of knowledge about the underlying distribution of the data. Therefore, one needs methods of *error estimation* to assess the performance of a classifier based on the given data. However, with the emergence of high-throughput measurement technologies, these biological data are now often characterized by an extremely large number of measurements made on a small number of samples, which creates significant challenges in the statistical analysis and interpretation of such data, in particular, difficult challenges in the application of error estimation methods.

Different error estimation techniques have been proposed through the years. For a comprehensive list of these error estimators the reader is referred to [36, 37]. Researchers have tried to characterize the performance of different error estimators in terms of their moments [38–42]. By comparing these results with those for the true error, obtained for example in [43–48], many suggestions have been made on applicability of these error estimators in practice [38, 49–51]. Most of this work has used asymptotic expansions based on the theory of infinitely large samples that do not apply to small-sample situations that are prevalent in medical applications. We would like to highlight a quote from R. A. Fisher [52], which appears in [53]:

> 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.

As another comment on this subject that has been made particularly in the context of error estimation, consider the comment by D. Hand [54] on asymptotic results by Kittler and Devijver [41] on the variance of so-called average conditional error rate estimators:

> 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, one may not see the serious implications of misusing asymptotic performance-guarantee tools in small-sample situations. There have been already 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, according to [55], "Five of the seven largest published studies addressing cancer prognosis did not classify patients better than chance." In another study [56], the authors have mentioned 21 studies that have flawed results mostly published in journals with impact factor larger 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 [56–58].

The resubstitution error estimator [59] and the leave-one-out cross-validation error estimator (variously credited to [60–63]) are the main focus of the present dissertation and have been used extensively in the literature dealing with small-sample biological high-throughput data – for instance, see [64–69], to cite just a few. It is noteworthy that some of these cited works have been subsequently criticized for lack of reproducible results due to the improper use of resubstitution and leave-one-out error estimation [56, 57], which only highlights further the critical need to study the performance of these error estimators in small-sample settings.

1. Resubstitution Error Estimator

The apparent classification error, or resubstitution error estimator [59], is given by

$$\hat{\varepsilon}_r = \frac{1}{n}\left[\sum_{i=1}^{n_0} I_{\{W(\hat{\mu}_0,\hat{\mu}_1,X_i)\leq 0\}} + \sum_{i=n_0+1}^{n_0+n_1} I_{\{W(\hat{\mu}_0,\hat{\mu}_1,X_i)>0\}}\right] = \hat{\alpha}_0\hat{\varepsilon}_r^0 + \hat{\alpha}_1\hat{\varepsilon}_r^1 \tag{1.8}$$

where I_A is the indicator variable for event A, $\hat{\alpha}_i = n_i/n$ is the empirical mixing frequency for population Π_i, and $\hat{\epsilon}_i^r$ is the apparent error rate specific to population Π_i, with

$$\begin{aligned}\hat{\epsilon}_0^r &= \frac{1}{n_0}\sum_{i=1}^{n_0} I_{\{W(\bar{X}_0,\bar{X}_1,X_i)\leq 0\}}\\ \hat{\epsilon}_1^r &= \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} I_{\{W(\bar{X}_0,\bar{X}_1,X_i)>0\}}\end{aligned} \tag{1.9}$$

2. Leave-one-out Estimator

The leave-one-out error estimator [60] for the LDA classification rule is given by

$$\hat{\epsilon}_l = \frac{1}{n}\left[\sum_{i=1}^{n_0} I_{\{W^{(i)}(\hat{\mu}_0,\hat{\mu}_1,X_i)\leq 0\}} + \sum_{i=n_0+1}^{n_0+n_1} I_{\{W^{(i)}(\hat{\mu}_0,\hat{\mu}_1,X_i)>0\}}\right] = \hat{\alpha}_0\hat{\varepsilon}_l^0 + \hat{\alpha}_1\hat{\varepsilon}_l^1 \tag{1.10}$$

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

$$\begin{aligned} \hat{\varepsilon}_l^0 &= \frac{1}{n_0}\sum_{i=1}^{n_0} I_{\{W^{(i)}(\hat{\mu}_0,\hat{\mu}_1,X_i)\leq 0\}} \\ \hat{\varepsilon}_l^1 &= \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} I_{\{W^{(i)}(\hat{\mu}_0,\hat{\mu}_1,X_i)>0\}} \end{aligned} \tag{1.11}$$

However, from the definition of this estimator, it is clear that we have $\mathrm{E}(\hat{\varepsilon}_l) = \hat{\alpha}_0\mathrm{E}(\varepsilon_{n_0-1}^0) + \hat{\alpha}_1\mathrm{E}(\varepsilon_{n_1-1}^1)$ where $\varepsilon_{n_0-1}^0$ and $\varepsilon_{n_1-1}^1$ are the true errors defined in (1.4) for a problem of $n_0 - 1$ and $n_1 - 1$ observations, respectively. Therefore, studying the expectation of true error of misclassification suffices to determine that of leave-one-out. However, the variance of leave-one-out and its cross-moment with true error still need to be investigated separately.

3. Plug-in Error Estimator

The *plug-in error estimator*, originally proposed in [6], is obtained by replacing μ_0, α_0 and α_1 by $\hat{\mu}_0$, $\hat{\alpha}_0$ and $\hat{\alpha}_1$ in (1.6). If we denote this estimator by $\hat{\epsilon}_p$, then after simplification we have $\hat{\epsilon}_p = \Phi(-\hat{\delta}/2)$ as given in [35], where $\hat{\delta} = \sqrt{(\hat{\mu}_0 - \hat{\mu}_1)^T\Sigma^{-1}(\hat{\mu}_0 - \hat{\mu}_1)}$. Based on simulation experiments, it has been stated in [35] and [60] that this estimator has a similar behavior as resubstitution.

D. Bibliography on the True Error and Its Estimators for Linear Discriminant

In this section, we survey the main results on the development of LDA and its variants and the rigorous analytical results regarding the distributional knowledge of estimators of true error for LDA; however, it is out of scope of this dissertation to survey results on error estimators for all classification rules. Interested readers are encouraged to combine the papers mentioned here with those of [36] for results until 1974, [54] for results from 1974 to 1986, and [37] for results until 2000, to get a complete bibliography of the papers on error estimation.

1. From 1936 to 1966

There was a large body of work in these years on development of linear discriminant function and its variants. Interested readers are encouraged to see [7, 70–74].

Fisher in his seminal paper in 1936 [6], proposed a linear function that maximizes the ratio of between to within scatter of classes. It is noteworthy to mention that to find this linear function, Fisher did not assume any parametric assumption on the class-conditional densities; in fact he used a linear regression procedure. In addition, the Fisher linear function is not a discriminant itself; however, we can build a discriminant using it.

The ratio he considered for maximization purpose was:

$$\hat{F}(a) = \frac{(a^T\hat{\mu}_0 - a^T\hat{\mu}_1)^2}{aSa} \tag{1.12}$$

where a is the weight vector and is a column vector of dimension p, $\hat{\mu}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} X_i$ and $\hat{\mu}_1 = \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} X_i$ and S is the pooled sample covariance matrix:

$$S = \frac{(n_0-1)S_0 + (n_1-1)S_1}{n_0+n_1-2}\,, \qquad S_j = \frac{1}{n_j - 1}\sum_{i:Y_i=j}(X_i - \hat{\mu}_j)(X_i - \hat{\mu}_j)^T \tag{1.13}$$

A reasonable question to ask is why maximization of $\hat{F}(a)$ is a proper criterion for classification purposes. A simple justification is given by Theorem 4.4 in [24], which states that for any linear discrimination rule with weight vector a, the probability of error ϵ is upper bounded as follows:

$$\epsilon \leq \inf_{a \in \Re^p} \frac{1}{1 + F(a)} \tag{1.14}$$

where

$$F(a) = \frac{(a^T \mu_0 - a^T \mu_1)^2}{a^T \Sigma_0 a + a^T \Sigma_1 a} \tag{1.15}$$

and $\mu_i = \mathrm{E}[X_i]$ and $\Sigma_0 = E[(X_i - \mu_i)(X_i - \mu_i)^T]$. Here no parametric assumptions about the distributions are made. If these parameters were known, then maximizing $F(a)$ leads to tightest upper bound on true error; however, in practice we can replace these unknown parameters by their estimates that leads to $\hat{F}(a)$.

In 1944, Wald constructed the most powerful test for testing the class of a sample, using the Neyman-Pearson lemma [25]. Further, he suggested replacing the true distributional parameters appearing in the critical region by their sample estimates, thereby providing the first instance of linear discriminant analysis. It is very closely related to Fisher discriminant function, being in fact a linear function of measurements that best discriminate the populations (i.e., maximizes the Fisher ratio). He also suggested the problem of finding the distribution of the discriminant itself, which is required to determine the probability of misclassification. In this regard, he represented the distribution of his statistic (called Wald's statistic later) in terms of three statistics that he called them m_1, m_2, and m_3 and were used later in literature [26, 75].

In 1947, Smith [59] proposed for the first time the use of the apparent error, also called resubstitution, in connection with the sample quadratic discriminant. In 1951, Harter obtained the exact distribution of Wald's statistic in the univariate case and approximated the multivariate case with the assumption that at least one of the populations has zero mean [76].

The LDA discriminant, also known as Anderson's statistic, was proposed by Anderson in 1951 [26]. He also proposed multi-class classification using various discriminants. Additional work on the distribution of Anderson's, Wald's, $W^\star$, Rao's, and Z statistics, which are all variants of linear discriminants, can be found in [34, 43, 46–48, 75, 77–95]. These results can be used to find the expectation of the true error. One can replace the true parameters of the class-conditional densities that appear in these expressions by their ML estimators to build plug-in types of estimators of the expected true error; however, it should be noted that when a classifier is designed, it is of more interest to estimate its true error, not the expected true error.

In 1964, the first use of the leave-one-out error estimator, known in the Soviet Union as *sliding egzam*, was proposed in [96]. This Russian paper precedes that of [60], to which this estimator is usually credited.

2. From 1966 to 2000

Hereafter, unless otherwise stated explicitly, the statistic under study is Anderson's statistic. Our attempt here is to briefly mention rigorous analytical results on the distributional knowledge of error estimators in the literature.

In 1966, Hills attempted to unify the notation commonly used in the literature. In addition, he considered different scenarios, such as multinomial, multivariate Bernoulli distribution, and normal distribution in univariate and multivariate cases. For the univariate normal model, he derived the exact expectation of the resubstitution and plug-in error estimators, with the assumption of knowing the common variance of classes [1]. In this scenario, he also established some inequalities between the expectation of resubstitution, expectation of true error, and bayes error when the class sample sizes are equal.

In 1972, Foley represented the expectation of resubstitution by an infinite series of certain gamma functions. He made the assumption that the common covariance matrix of classes is known [51].

In 1973, Sorum [97] obtained the exact expressions for the expectation and variance of of resubstitution, leave-one-out, and different parametric estimators in the univariate case. She also expanded these exact results in an asymptotic sense to simplify the comparison of these estimators in the univariate case. She made the assumption that the common variance of the classes is known.

In the same year, McLachlan gave an asymptotic expression for expectation of the plug-in error estimator [44]. McLachlan derived his results under the multivariate normal model with unknown common covariance matrix.

In 1974, McLachlan obtained asymptotic expressions for the expectation and variance of of several parametric error estimators, such as the usual plug-in error estimator [45]. Here again he considered a multivariate normal model with unknown common covariance matrix.

In 1975, Moran gave exact expressions for the expectation of the resubstitution and plug-in error estimators under a multivariate normal model with known common covariance matrix [35].

In 1992, Davison and Hall demonstrated the smaller variance but larger bias of bootstrap compared to leave-one-out. They showed this fact analytically in the univariate case with unknown and possibly different class variances [98].

There have been numerous Monte Carlo studies [38, 60, 99–113], unquantified approximations [38, 42, 114], and unproven statements [115] on error estimation for LDA.

In addition to the results we have mentioned above, there has been a tremendous effort in the eastern hemisphere mostly associated with properties of true error for discriminant analysis. Interested readers can consult [116] for more information.

3. A History Chart of the Distributional Knowledge of Error Estimation for LDA

Below we provide a chart summarizing the rigorous analytical work on the distributional knowledge of error estimators in the context of LDA.

- 1966 Hills: Univariate with Known Common Variance; Exact Expectation; Resubstitution; Plug-in Estimator; Optimistic Bias of Resubstitution for Equal Class Sample Sizes [1].
- 1972 Foley: Multivariate with Known Covariance; Exact Expectation; Resubstitution [51].
- 1973 Sorum: Univariate with Known Common Variance; Exact and Asymptotic Expectation and Variance; Resubstitution; Leave-one-out; Several Parametric Estimators [97].
- 1973 McLachlan: Multivariate with Unknown Covariance; Asymptotic Expectation; Plug-in Estimator [44].
- 1974 McLachlan: Multivariate with Unknown Covariance; Asymptotic Expectation and Variance; Several Parametric Estimators [45].
- 1975 Moran: Multivariate with Known Covariance; Exact Expectation; Resubstitution; Plug-in Estimator [35].
- 1992 Davison and Hall: Univariate with Unknown Possibly Different Variances of Classes; Asymptotic Expectation and Variance; Bootstrap; Leave-one-out [98].

CHAPTER II

JOINT SAMPLING DISTRIBUTION BETWEEN ACTUAL AND ESTIMATED CLASSIFICATION ERRORS FOR LINEAR DISCRIMINANT ANALYSIS*

The present chapter furthers the analytical study of error estimation by deriving, for what is believed to be the first time, the analytical formulation for the joint sampling distribution of the actual and estimated errors for a classification rule. We consider here the LDA classification rule and the resubstitution and leave-one-out error estimators, under a general parametric Gaussian assumption.

We will give in this chapter exact and approximate expressions that allow the computation of the joint probability:

$$P\left(\hat{\varepsilon} = \frac{k}{n_0 + n_1},\, \varepsilon < z\right), \quad k = 0, 1, \ldots, n_0 + n_1\,,\; 0 \le z \le 1 \tag{2.1}$$

where ε is the actual classification error rate, and $\hat{\varepsilon}$ is either the resubstitution estimator $\hat{\varepsilon}_r$ or the leave-one-out estimator $\hat{\varepsilon}_l$, in the case where the classes are Gaussian distributed. By simple summation along the discrete variable, this allows one to easily compute the associated joint (cumulative) distribution functions, if so desired. More importantly, from the expressions for the joint probability in (2.1), one can compute the exact bias, deviation variance, and RMS of estimation (in terms of the mean, variance and second moment of $\hat{\varepsilon} - \varepsilon$), as well as exact conditional probability $P(\varepsilon < z \,|\, \hat{\varepsilon})$, which leads to the computation of exact conditional bounds on the actual error, as well as the exact regression $E[\varepsilon \,|\, \hat{\varepsilon}]$ of the actual on the estimated error, as will be detailed in Section C.

*Reprinted with permission from “Joint Sampling Distribution Between Actual and Estimated Classification Errors for Linear Discriminant Analysis” by A. Zollanvari, U.M. Braga-Neto and E.R. Dougherty, *IEEE Transactions on Information Theory*, vol. 56, no. 2, pp. 784-804, 2010. Copyright 2010 by IEEE.

Likewise, we will give expressions, in the univariate case, that allow computation of the joint probability density

$$p\left(\hat{\varepsilon} = \frac{k}{n_0 + n_1}, \, \varepsilon = z\right), \quad k = 0, 1, \ldots, n_0 + n_1 \, , \, 0 \le z \le 1 \tag{2.2}$$

where $\hat{\varepsilon}$ is again either the resubstitution estimator $\hat{\varepsilon}_r$ or the leave-one-out estimator $\hat{\varepsilon}_l$, in the case where the classes are Gaussian distributed. Note that, even though we are using the terminology "density," the quantity in (2.2) is in fact a combination of density in ε and probability mass function in $\hat{\varepsilon}$.

A. Univariate Case

Consider a set of $n = n_0 + n_1$ i.i.d. univariate samples, where n_0 samples, represented by $\{X_1, X_2, \ldots X_{n_0}\}$, come from population Π_0 distributed as $N(\mu_0, \sigma_0^2)$, and n_1 samples $\{X_{n_0+1}, X_{n_0+2}, \ldots X_{n_0+n_1}\}$ come from population Π_1 distributed as $N(\mu_1, \sigma_1^2)$. The problem is to assign a new sample $X = x$ from the mixture population $p\,\Pi_0 + (1-p)\Pi_1$, $0 < p < 1$, to one of the classes. Without loss of generality, we will assume throughout this Section that $\mu_0 > \mu_1$. We will assume, for simplicity, that $p = \frac{1}{2}$, but the approach is easily generalizable to the case $p \neq \frac{1}{2}$.

In the univariate case, the LDA classifier and discriminant reduces to the

$$\psi(x) = \begin{cases} 0, & \text{if } W(x) = (x - \hat{\mu})(\hat{\mu}_0 - \hat{\mu}_1) > 0 \\ 1, & \text{otherwise} \end{cases} \tag{2.3}$$

where $\hat{\mu}_0$ and $\hat{\mu}_1$ are the sample means for each class and $\hat{\mu} = \frac{1}{2}(\hat{\mu}_0 + \hat{\mu}_1)$.

1. Resubstitution

From (2.3), we see that ψ can be written simply as $\psi(x) = I_{\{x<\hat{\mu}\}}$, if $\hat{\mu}_0 > \hat{\mu}_1$, i.e., sample means are on the same side of the cutpoint $\hat{\mu}$ as the corresponding actual means, and $\psi(x) = I_{\{x>\hat{\mu}\}}$, if $\hat{\mu}_0 < \hat{\mu}_1$, i.e., sample means are on the wrong side of the cutpoint (the case $\hat{\mu}_0 = \hat{\mu}_1$ having probability 0). The first case may be called "direct" classification, while the second case characterizes "reverse" classification.

Let us introduce the functions $\varepsilon^{\uparrow} : R \to [0,1]$ and $\varepsilon^{\downarrow} : R \to [0,1]$ as follows.

$$\varepsilon^{\uparrow}(w) = \frac{1}{2}\left[\Phi\left(\frac{w-\mu_0}{\sigma_0}\right) + \Phi\left(\frac{\mu_1 - w}{\sigma_1}\right)\right] \tag{2.4}$$

and

$$\varepsilon^{\downarrow}(w) = 1 - \varepsilon^{\uparrow}(w) = \frac{1}{2}\left[\Phi\left(\frac{\mu_0 - w}{\sigma_0}\right) + \Phi\left(\frac{w-\mu_1}{\sigma_1}\right)\right] \tag{2.5}$$

where $\Phi(x)$ is the Gaussian cumulative distribution function evaluated at x.

The actual error for the classifier ψ in (2.3) is a function of $\hat{\mu}$ and of the "direction" of classification:

$$\varepsilon = \begin{cases} \varepsilon^{\uparrow}(\hat{\mu}), & \hat{\mu}_0 > \hat{\mu}_1 \text{ (direct classification)} \\[2ex] \varepsilon^{\downarrow}(\hat{\mu}), & \hat{\mu}_0 < \hat{\mu}_1 \text{ (reverse classification)} \end{cases} \tag{2.6}$$

a. Equal-variance Case

In this section, it is assumes that $\sigma_0 = \sigma_1 = \sigma$ (this assumption will be dropped in the next Section). The restriction $\varepsilon < z$ in (2.1) puts a corresponding restriction on where $\hat{\mu}$ may lie on the real line, which in turn affects the derivation of the joint probability in (2.1). For direct classification, ε is always under 0.5, while for reverse classification, ε is always above 0.5. In addition, if ε^* denotes the optimal (Bayes) classification error, then

- Direct classification $\Rightarrow \varepsilon^* = \varepsilon^{\uparrow}(w_1) \leq \varepsilon < 0.5$
- Reverse classification $\Rightarrow 0.5 < \varepsilon \leq 1 - \varepsilon^* = \varepsilon^{\downarrow}(w_1)$,

where $w_1 = \frac{1}{2}(\mu_0 + \mu_1)$ is the single point where the two densities $N(\mu_0, \sigma^2)$ and $N(\mu_1, \sigma^2)$ are equal. See the example in Figure 1, where the actual error rate ε is plotted as a function of $\hat{\mu}$, for the case $\mu_0 = 1$, $\mu_1 = 0$, and $\sigma_0 = \sigma_1 = 1$.

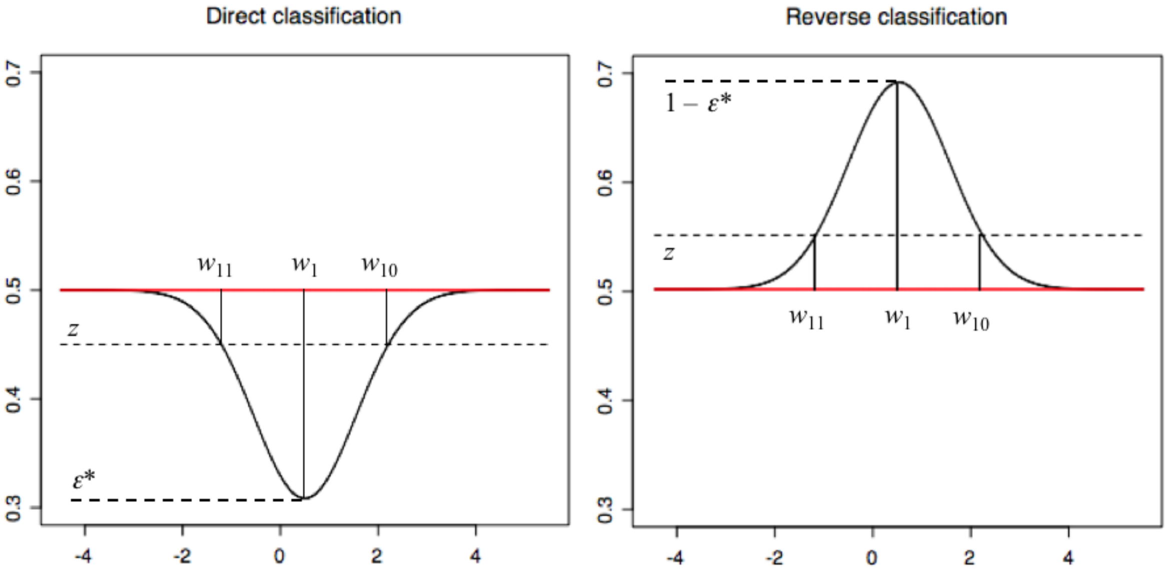

Fig. 1. Plots of actual error as a function of $\hat{\mu}$, for $\mu_0 = 1$, $\mu_1 = 0$, and $\sigma_0 = \sigma_1 = 1$. Left: plot of $\varepsilon^{\uparrow}(w)$, direct classification ($\hat{\mu}_0 > \hat{\mu}_1$). Right: plot of $\varepsilon^{\downarrow}(w)$, reverse classification ($\hat{\mu}_0 < \hat{\mu}_1$).

The event $[\varepsilon < z]$ is characterized as follows (see Figure 1):

$$[\varepsilon < z] = \begin{cases} \varnothing\,, & \text{for } z < \varepsilon^* \\ [\,\hat{\mu} \in (w_{11}, w_{10}), \hat{\mu}_0 > \hat{\mu}_1\,]\,, & \text{for } \varepsilon^* \leq z \leq 0.5 \\ [\,\hat{\mu}_0 > \hat{\mu}_1] \cup [\,\hat{\mu} \notin (w_{11}, w_{10}), \hat{\mu}_0 < \hat{\mu}_1\,]\,, & \text{for } 0.5 < z \leq 1 - \varepsilon^* \\ \Omega\,, & \text{for } z > 1 - \varepsilon^* \end{cases} \tag{2.7}$$

where Ω denotes the entire sample space, and the cutpoints $w_{11} < w_{10}$ can be found easily in each case by numerical inversion of the respective function $\varepsilon^{\uparrow}$ or $\varepsilon^{\downarrow}$. We have thus established the following Lemma.

Lemma 1. *For $\sigma_0 = \sigma_1$,*

$$P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \varepsilon < z\right) = \begin{cases} 0\,, & \text{for } z < \varepsilon^* \\[2ex] P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} \in (w_{11}, w_{10}), \hat{\mu}_0 > \hat{\mu}_1\right)\,, & \text{for } \varepsilon^* \le z \le 0.5 \\[2ex] P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu}_0 > \hat{\mu}_1\right) + & \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} \notin (w_{11}, w_{10}), \hat{\mu}_0 < \hat{\mu}_1\right)\,, & \text{for } 0.5 < z \le 1 - \varepsilon^* \\[2ex] P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}\right)\,, & \text{for } z > 1 - \varepsilon^* \end{cases} \tag{2.8}$$

The following theorem specifies how to compute these probabilities in the case $k = 0$ (no apparent error). This result is next extended to $k > 0$.

Theorem 1. *Let $X_i \sim N(\mu_0, \sigma^2)$ be i.i.d. observations for $i = 1, \dots, n_0$, and $X_i \sim N(\mu_1, \sigma^2)$ be i.i.d. observations for $i = n_0 + 1, \dots, n_0 + n_1$ used to derive the classifier in (2.3). Then*

$$\begin{aligned} P\left(\hat{\varepsilon}_r = 0, \hat{\mu} \in (a,b), \hat{\mu}_0 > \hat{\mu}_1\right) &= P(Z_1 > \mathbf{0}) \\ P\left(\hat{\varepsilon}_r = 0, \hat{\mu} \notin (a,b), \hat{\mu}_0 < \hat{\mu}_1\right) &= P(Z_2 < \mathbf{0}) + P(Z_3 < \mathbf{0}) \\ P\left(\hat{\varepsilon}_r = 0, \hat{\mu}_0 > \hat{\mu}_1\right) &= P(Z_4 > \mathbf{0}) \\ P\left(\hat{\varepsilon}_r = 0\right) &= P(Z_4 > \mathbf{0}) + P(Z_4 < \mathbf{0}) \end{aligned} \tag{2.9}$$

where Z_1 is a Gaussian random vector of size n_0+n_1+3, with mean μ_{Z_1} given by:

$$\begin{bmatrix} (\mu_0 - \mu_1)\mathbf{1}_{n_0+n_1+1} \\ \\ (\mu_0 + \mu_1) - 2a \\ \\ -(\mu_0 + \mu_1) + 2b \end{bmatrix} \tag{2.10}$$

and covariance matrix $\Sigma_{Z_1} = \sigma^2 H$, where:

$$H_{ij} = \begin{cases} \frac{4n_0-3}{n_0} + \frac{1}{n_1}, & i,j = 1,\dots,n_0, i = j \\ -\frac{3}{n_0} + \frac{1}{n_1}, & i,j = 1,\dots,n_0, i \neq j \\ \frac{1}{n_0} + \frac{4n_1-3}{n_1}, & i,j = n_0+1,\dots,n_0+n_1, i = j \\ \frac{1}{n_0} - \frac{3}{n_1}, & i,j = n_0+1,\dots,n_0+n_1, i \neq j \\ \frac{1}{n_0} - \frac{1}{n_1}, & \begin{cases} i = n_0+n_1+2, j = 1,\dots,n_0+n_1+1 \\ j = n_0+n_1+2, i = 1,\dots,n_0+n_1+1 \end{cases}, \\ \frac{1}{n_1} - \frac{1}{n_0}, & \begin{cases} i = n_0+n_1+3, j = 1,\dots,n_0+n_1+1 \\ j = n_0+n_1+3, i = 1,\dots,n_0+n_1+1 \end{cases}, \\ -\frac{1}{n_0} - \frac{1}{n_1}, & \begin{cases} i = n_0+n_1+2, j = n_0+n_1+3 \\ i = n_0+n_1+3, j = n_0+n_1+2 \end{cases}, \\ \frac{1}{n_0} + \frac{1}{n_1}, & otherwise \end{cases} \tag{2.11}$$

Furthermore, Z_2 (resp. Z_3) is a Gaussian random vector of size n_0+n_1+2, obtained from Z_1 by eliminating component n_0+n_1+3 (resp. n_0+n_1+2), while Z_4 is Gaussian random vector of size n_0+n_1+1, obtained from Z_1 by eliminating both components n_0+n_1+2 and n_0+n_1+3.

Proof. See Appendix.

Now observe that the probability of committing $k > 0$ errors on the training data can be written as

$$\begin{aligned} P([k \text{ errors}]) &= \sum_{l=0}^{k} P([l \text{ errors in class 0 and } k-l \text{ errors in class 1}]) \\ &= \sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} P([X_1, \ldots, X_l \text{ in error and } X_{n_0+1}, \ldots, X_{n_0+l-k} \text{ in error}]) \end{aligned} \tag{2.12}$$

Furthermore, the random vectors Z_i in Theorem 1 assume that no training point in $X_1, \ldots, X_{n_0+n_1}$ is misclassified; misclassification of X_j implies flipping the sign of the j-th component of Z_i, as can be easily checked in the proof of Theorem 1. This establishes the following theorem.

Theorem 2. *Under the same conditions as in Theorem 1,*

$$\begin{aligned} P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} \in (a,b), \hat{\mu}_0 > \hat{\mu}_1\right) &= \sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} P(E^2_{l,k-l} Z_1 > \mathbf{0}) \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} \notin (a,b), \hat{\mu}_0 < \hat{\mu}_1\right) &= \sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} \left[P(E^1_{l,k-l} Z_2 < \mathbf{0}) + P(E^1_{l,k-l} Z_3 < \mathbf{0})\right] \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu}_0 > \hat{\mu}_1\right) &= \sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} P(E^0_{l,k-l} Z_4 > \mathbf{0}) \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}\right) &= \sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} \left[P(E^0_{l,k-l} Z_4 > \mathbf{0}) + P(E^0_{l,k-l} Z_4 < \mathbf{0})\right] \end{aligned} \tag{2.13}$$

where the vectors Z_i, $i = 1, \ldots, 4$, are defined in Theorem 1, and $E^r_{l,k-l}$ is a diagonal matrix of size $n_0 + n_1 + 1 + r$, for $r = 0, 1, 2$, with diagonal elements defined to be $(-\mathbf{1}_l, \mathbf{1}_{n_0-l}, -\mathbf{1}_{k-l}, \mathbf{1}_{n_1-(k-l)}, 1, \mathbf{1}_r)$.

Theorem 2, in conjunction with Lemma 1, allows the exact computation of the joint probability in (2.1) for the resubstitution error estimator. The probabilities of the kind $P(Z > 0)$, where Z is a Gaussian vector, which are needed in the computations above, can be readily computed using an algorithm for integration of multivariate Gaussian densities over rectangular regions, due to Genz and Bretz [117]. This provides an efficient and very accurate method for the exact computation of the joint probability in (2.1).

b. Unequal-variance Case

In this section, we consider the case where $\sigma_0 \neq \sigma_1$. As was seen in the previous section, when the variances are equal, the class densities are equal at a single point $w_1 = \frac{1}{2}(\mu_0 + \mu_1)$, which also is an extremum point of the classification error functions $\varepsilon^{\uparrow}$ and $\varepsilon^{\downarrow}$. In the present unequal-variance case, the class densities are equal at two points w_1 and w_2,

$$\begin{aligned} w_1 &= \frac{\mu_1\sigma_0^2 - \mu_0\sigma_1^2 + \sigma_0\sigma_1\sqrt{(\mu_1-\mu_0)^2 + 2(\sigma_1^2-\sigma_0^2)\ln\frac{\sigma_1}{\sigma_0}}}{\sigma_0^2-\sigma_1^2} \\ w_2 &= \frac{\mu_1\sigma_0^2 - \mu_0\sigma_1^2 - \sigma_0\sigma_1\sqrt{(\mu_1-\mu_0)^2 + 2(\sigma_1^2-\sigma_0^2)\ln\frac{\sigma_1}{\sigma_0}}}{\sigma_0^2-\sigma_1^2} \end{aligned} \tag{2.14}$$

where $w_1 > w_2$ for $\sigma_0 > \sigma_1$ and $w_1 < w_2$ for $\sigma_0 < \sigma_1$. These points are extrema of the classification error, in the sense that

- Direct classification $\Rightarrow \varepsilon^* = \varepsilon^{\uparrow}(w_1) \leq \varepsilon \leq \varepsilon^{\uparrow}(w_2)$, with $\varepsilon^{\uparrow}(w_1) < 0.5 < \varepsilon^{\uparrow}(w_2)$.
- Reverse classification $\Rightarrow \varepsilon^{\downarrow}(w_2) \leq \varepsilon \leq \varepsilon^{\downarrow}(w_1) = 1-\varepsilon^*$, with $\varepsilon^{\downarrow}(w_2) < 0.5 < \varepsilon^{\downarrow}(w_1)$.

This is illustrated in Figure 2, where the actual error rate ε is plotted as a function of $\hat{\mu}$, for the case $\mu_0 = 1$, $\mu_1 = 0$, $\sigma_0 = 3$, and $\sigma_1 = 1$.

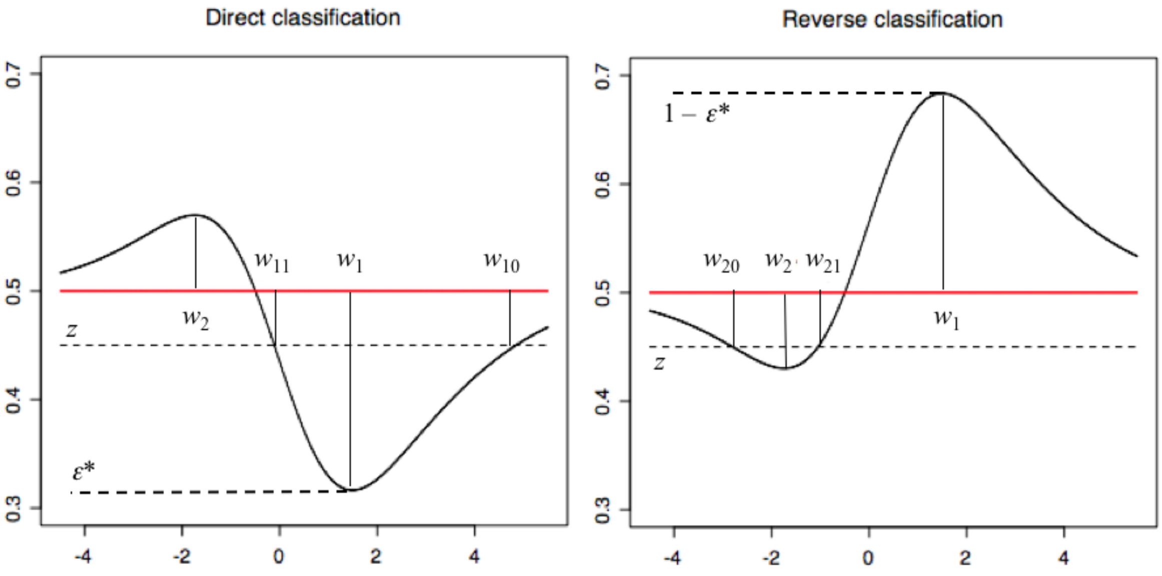

Fig. 2. Plots of actual error as a function of $\hat{\mu}$, for $\mu_0 = 1$, $\mu_1 = 0$, $\sigma_0 = 3$, $\sigma_1 = 1$ and $\varepsilon^{\downarrow}(w_2) < z \leq 0.5$. Left: plot of $\varepsilon^{\uparrow}(w)$, direct classification ($\hat{\mu}_0 > \hat{\mu}_1$). Right: plot of $\varepsilon^{\downarrow}(w)$, reverse classification ($\hat{\mu}_0 < \hat{\mu}_1$).

The event $[\varepsilon < z]$ is characterized as follows (see Figure 2):

$$[\varepsilon < z] = \begin{cases} \varnothing\,, & \text{for } z < \varepsilon^* \\ [\,\hat{\mu} \in (w_{11}, w_{10}), \hat{\mu}_0 > \hat{\mu}_1]\,, & \text{for } \varepsilon^* \leq z \leq \varepsilon^{\downarrow}(w_2) \\ [\,\hat{\mu} \in (w_{11}, w_{10}), \hat{\mu}_0 > \hat{\mu}_1] \cup [\,\hat{\mu} \in (w_{21}, w_{20}), \hat{\mu}_0 < \hat{\mu}_1]\,, & \text{for } \varepsilon^{\downarrow}(w_2) < z \leq 0.5 \\ [\,\hat{\mu} \notin (w_{11}, w_{10}), \hat{\mu}_0 < \hat{\mu}_1] \cup [\,\hat{\mu} \notin (w_{21}, w_{20}), \hat{\mu}_0 > \hat{\mu}_1]\,, & \text{for } 0.5 < z \leq \varepsilon^{\uparrow}(w_2) \\ [\,\hat{\mu}_0 > \hat{\mu}_1] \cup [\,\hat{\mu} \notin (w_{11}, w_{10}), \hat{\mu}_0 < \hat{\mu}_1]\,, & \text{for } \varepsilon^{\uparrow}(w_2) < z \leq 1 - \varepsilon^* \\ \Omega\,, & \text{for } z > 1 - \varepsilon^* \end{cases} \tag{2.15}$$

where the cutpoints $w_{11} < w_{10}$ and $w_{21} < w_{20}$ can be found easily in each case by numerical inversion of the respective function $\varepsilon^{\uparrow}$ or $\varepsilon^{\downarrow}$, such that $w_1 \in (w_{11}, w_{10})$ and

$w_2 \in (w_{21}, w_{20})$. We have thus established the following Lemma.

Lemma 2. *For arbitrary $\sigma_0 \neq \sigma_1$,*

$$P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \varepsilon < z\right) = \begin{cases} 0\,, & \textit{for } z < \varepsilon^* \\[2ex] P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} \in (w_{11}, w_{10}), \hat{\mu}_0 > \hat{\mu}_1\right)\,, & \textit{for } \varepsilon^* \le z \le \varepsilon^{\downarrow}(w_2) \\[2ex] P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} \in (w_{11}, w_{10}), \hat{\mu}_0 > \hat{\mu}_1\right) + & \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} \in (w_{21}, w_{20}), \hat{\mu}_0 < \hat{\mu}_1\right)\,, & \textit{for } \varepsilon^{\downarrow}(w_2) < z \le 0.5 \\[2ex] P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} \notin (w_{11}, w_{10}), \hat{\mu}_0 < \hat{\mu}_1\right) + & \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} \notin (w_{21}, w_{20}), \hat{\mu}_0 > \hat{\mu}_1\right)\,, & \textit{for } 0.5 < z \le \varepsilon^{\uparrow}(w_2) \\[2ex] P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu}_0 > \hat{\mu}_1\right) + & \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} \notin (w_{11}, w_{10}), \hat{\mu}_0 < \hat{\mu}_1\right)\,, & \textit{for } \varepsilon^{\uparrow}(w_2) < z \le 1 - \varepsilon^* \\[2ex] P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}\right)\,, & \textit{for } z > 1 - \varepsilon^* \end{cases} \tag{2.16}$$

The following theorem specifies how to compute these probabilities in the case $k = 0$ (no apparent error). The proof of this theorem is similar to the proof of Theorem 1 and is thus omitted.

Theorem 3. *Let $X_i \sim N(\mu_0, \sigma^2)$ be i.i.d. observations for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \sigma^2)$ be i.i.d. observations for $i = n_0+1, \ldots, n_0+n_1$ used to derive the classifier* *in (2.3). Then*

$$\begin{aligned}
P\left(\hat{\varepsilon}_r = 0, \hat{\mu} \in (a,b), \hat{\mu}_0 > \hat{\mu}_1\right) &= P(Z_1 > \mathbf{0}) \\
P\left(\hat{\varepsilon}_r = 0, \hat{\mu} \in (a,b), \hat{\mu}_0 < \hat{\mu}_1\right) &= P(Z_1' < \mathbf{0}) \\
P\left(\hat{\varepsilon}_r = 0, \hat{\mu} \notin (a,b), \hat{\mu}_0 < \hat{\mu}_1\right) &= P(Z_2 < \mathbf{0}) + P(Z_3 < \mathbf{0}) \\
P\left(\hat{\varepsilon}_r = 0, \hat{\mu} \notin (a,b), \hat{\mu}_0 > \hat{\mu}_1\right) &= P(Z_2' > \mathbf{0}) + P(Z_3' > \mathbf{0}) \\
P\left(\hat{\varepsilon}_r = 0, \hat{\mu}_0 > \hat{\mu}_1\right) &= P(Z_4 > \mathbf{0}) \\
P\left(\hat{\varepsilon}_r = 0\right) &= P(Z_4 > \mathbf{0}) + P(Z_4 < \mathbf{0})
\end{aligned} \tag{2.17}$$

where Z_1 is a Gaussian random vector of size n_0+n_1+3, with mean μ_{Z_1} given by:

$$\mu_{Z_1} = \left[(\mu_0 - \mu_1)\mathbf{1}^T_{n_0+n_1+1},\ (\mu_0 + \mu_1) - 2a,\ -(\mu_0 + \mu_1) + 2b\right]^T \tag{2.18}$$

and covariance matrix Σ_{Z_1} given by

$$(\Sigma_{Z_1})_{ij} = \begin{cases}
(4n_0 - 3)\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & i,j = 1,\ldots,n_0, i = j \\
-3\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & i,j = 1,\ldots,n_0, i \neq j \\
\frac{\sigma_0^2}{n_0} + (4n_1 - 3)\frac{\sigma_1^2}{n_1}, & i,j = n_0+1,\ldots,n_0+n_1, i = j \\
\frac{\sigma_0^2}{n_0} - 3\frac{\sigma_1^2}{n_1}, & i,j = n_0+1,\ldots,n_0+n_1, i \neq j \\
\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1}, & \begin{cases} i = n_0+n_1+2, j = 1,\ldots,n_0+n_1+1 \\ j = n_0+n_1+2, i = 1,\ldots,n_0+n_1+1 \end{cases}, \\
\frac{\sigma_1^2}{n_1} - \frac{\sigma_0^2}{n_0}, & \begin{cases} i = n_0+n_1+3, j = 1,\ldots,n_0+n_1+1 \\ j = n_0+n_1+3, i = 1,\ldots,n_0+n_1+1 \end{cases}, \\
-\left(\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\right), & \begin{cases} i = n_0+n_1+2, j = n_0+n_1+3 \\ i = n_0+n_1+3, j = n_0+n_1+2 \end{cases}, \\
\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & \textit{otherwise}
\end{cases} \tag{2.19}$$

Here Z_1' is a Gaussian random vector of size n_0+n_1+3, obtained from Z_1 by multiplying by -1 the last two components of Z_1. Furthermore, Z_2 (resp. Z_3) is a Gaussian random vector of size n_0+n_1+2, obtained from Z_1 by eliminating component n_0+n_1+3 (resp. n_0+n_1+2), while Z_2' (resp. Z_3') is a Gaussian random vector of size n_0+n_1+2, obtained from Z_1' by eliminating component n_0+n_1+3 (resp. n_0+n_1+2) . Finally, Z_4 is Gaussian random vector of size n_0+n_1+1, obtained from Z_1 by eliminating both components n_0+n_1+2 and n_0+n_1+3.

The previous result can be extended to the case $k > 0$ by using the same reasoning employed before in connection with Theorem 2, which establishes the following result.

Theorem 4. *Under the same conditions as in Theorem 3,*

$$\begin{aligned}
P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} \in (a,b), \hat{\mu}_0 > \hat{\mu}_1\right) &= \sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} P(E_{l,k-l}^{2} Z_1 > \mathbf{0}) \\
P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} \in (a,b), \hat{\mu}_0 < \hat{\mu}_1\right) &= \sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} P(E_{l,k-l}^{2} Z_1' < \mathbf{0}) \\
P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} \notin (a,b), \hat{\mu}_0 < \hat{\mu}_1\right) &= \sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} \left[P(E_{l,k-l}^{1} Z_2 < \mathbf{0}) + P(E_{l,k-l}^{1} Z_3 < \mathbf{0})\right] \\
P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} \notin (a,b), \hat{\mu}_0 > \hat{\mu}_1\right) &= \sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} \left[P(E_{l,k-l}^{1} Z_2' > \mathbf{0}) + P(E_{l,k-l}^{1} Z_3' > \mathbf{0})\right] \\
P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu}_0 > \hat{\mu}_1\right) &= \sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} P(E_{l,k-l}^{0} Z_4 > \mathbf{0}) \\
P\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}\right) &= \sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} \left[P(E_{l,k-l}^{0} Z_4 > \mathbf{0}) + P(E_{l,k-l}^{0} Z_4 < \mathbf{0})\right]
\end{aligned} \tag{2.20}$$

where the vectors Z_i, $i = 1, \ldots, 4$, Z_i', $i = 1, \ldots, 3$, are defined in Theorem 3, and $E_{l,k-l}^{r}$ is a diagonal matrix of size $n_0 + n_1 + 1 + r$, for $r = 0, 1, 2$, with diagonal elements $(-\mathbf{1}_l, \mathbf{1}_{n_0-l}, -\mathbf{1}_{k-l}, \mathbf{1}_{n_1-(k-l)}, 1, \mathbf{1}_r)$.

Theorem 4, in conjunction with Lemma 2, allows the exact computation of the joint probability in (2.1) for the resubstitution error estimator. The probabilities of the kind $P(Z > 0)$, where Z is a Gaussian vector, which are needed in the computations above, can be readily computed using the algorithm for integration of multivariate Gaussian densities over rectangular regions due to Genz and Bretz [117]. This provides an efficient and very accurate method for the exact computation of the joint probability in (2.1) in the resubstitution case.

c. Joint Density

It is relatively easy to apply a methodology similar to the one in the previous sections to obtain the joint density in (2.2) for the resubstitution error estimator. Let the value of the Gaussian density with mean μ and variance σ^2 at x be denoted by $\varphi(x, \mu, \sigma^2)$, and let $\psi(w) = |\varphi(x, \mu_0, \sigma_0^2) - \varphi(x, \mu_1, \sigma_1^2)|$. Lemma 3 can be easily shown. In addition, Lemma 3 holds for the case of equal variances $\sigma_0 = \sigma_1$, by considering only two regions with $z < 0.5$ and $z > 0.5$ and eliminating all terms that include w_{20} and w_{21}.

Lemma 3. *For arbitrary $\sigma_0 \neq \sigma_1$,*

$$p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \varepsilon = z\right) = \begin{cases} 0\,, & \textit{for } z < \varepsilon^* \\ \frac{1}{\psi(w_{11})}\, p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = w_{11}, \hat{\mu}_0 > \hat{\mu}_1\right) + & \\ \frac{1}{\psi(w_{10})}\, p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = w_{10}, \hat{\mu}_0 > \hat{\mu}_1\right)\,, & \textit{for } \varepsilon^* \leq z \leq \varepsilon^{\downarrow}(w_2) \\ \frac{1}{\psi(w_{11})}\, p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = w_{11}, \hat{\mu}_0 > \hat{\mu}_1\right) + & \\ \frac{1}{\psi(w_{10})}\, p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = w_{10}, \hat{\mu}_0 > \hat{\mu}_1\right) + & \\ \frac{1}{\psi(w_{21})}\, p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = w_{21}, \hat{\mu}_0 < \hat{\mu}_1\right) + & \\ \frac{1}{\psi(w_{20})}\, p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = w_{20}, \hat{\mu}_0 < \hat{\mu}_1\right)\,, & \textit{for } \varepsilon^{\downarrow}(w_2) < z \leq 0.5 \\ \frac{1}{\psi(w_{21})}\, p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = w_{21}, \hat{\mu}_0 > \hat{\mu}_1\right) + & \\ \frac{1}{\psi(w_{20})}\, p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = w_{20}, \hat{\mu}_0 > \hat{\mu}_1\right) + & \\ \frac{1}{\psi(w_{11})}\, p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = w_{11}, \hat{\mu}_0 < \hat{\mu}_1\right) + & \\ \frac{1}{\psi(w_{10})}\, p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = w_{10}, \hat{\mu}_0 < \hat{\mu}_1\right)\,, & \textit{for } 0.5 < z \leq \varepsilon^{\uparrow}(w_2) \\ \frac{1}{\psi(w_{11})}\, p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = w_{11}, \hat{\mu}_0 < \hat{\mu}_1\right) + & \\ \frac{1}{\psi(w_{10})} p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = w_{10}, \hat{\mu}_0 < \hat{\mu}_1\right)\,, & \textit{for } \varepsilon^{\uparrow}(w_2) < z \leq 1 - \varepsilon^* \\ 0 & \textit{for } z > 1 - \varepsilon^* \end{cases} \tag{2.21}$$

The following theorem specifies how to compute the terms on the right hand side of (2.21).

Theorem 5. *Under the same conditions as in Theorem 3,*

$$p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = a, \hat{\mu}_0 > \hat{\mu}_1\right) = \sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} P(E^0_{l,k-l}Y > \mathbf{0})\varphi\left(0, \mu_0+\mu_1-2a, \frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1}\right)$$
$$p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = a, \hat{\mu}_0 < \hat{\mu}_1\right) = \sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} P(E^0_{l,k-l}Y < \mathbf{0})\varphi\left(0, \mu_0+\mu_1-2a, \frac{\sigma_0^2}{n_0}+\frac{\sigma_1^2}{n_1}\right) \tag{2.22}$$

Here Y is a Gaussian random vector of size $n_0 + n_1 + 1$ with mean μ_Y given by:

$$\mu_Y = 2\tfrac{n_1\sigma_0^2(a-\mu_1)-n_0\sigma_1^2(a-\mu_0)}{n_1\sigma_0^2+n_0\sigma_1^2}\mathbf{1}_{n_0+n_1+1} \tag{2.23}$$

and covariance matrix Σ_Y given by

$$\Sigma_Y = \Sigma_{Y_{11}} - \frac{1}{n_0 n_1}\frac{(n_1\sigma_0^2 - n_0\sigma_1^2)^2}{n_1\sigma_0^2 + n_0\sigma_1^2}\mathbf{1}_{(n_0+n_1+1)\times(n_0+n_1+1)} \tag{2.24}$$

where:

$$\Sigma_{Y_{11}} = \begin{cases} (4n_0-3)\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & i,j = 1,\ldots,n_0, i = j \\ -3\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & i,j = 1,\ldots,n_0, i \neq j \\ \frac{\sigma_0^2}{n_0} + (4n_1-3)\frac{\sigma_1^2}{n_1}, & i,j = n_0+1,\ldots,n_0+n_1, i = j \\ \frac{\sigma_0^2}{n_0} - 3\frac{\sigma_1^2}{n_1}, & i,j = n_0+1,\ldots,n_0+n_1, i \neq j \\ \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & otherwise \end{cases} \tag{2.25}$$

and $E^0_{l,k-l}$ is the diagonal matrix used in theorem 4.

Proof. See Appendix.

Theorem 5, in conjunction with Lemma 3, allows the exact computation of the joint density in (2.2) for the resubstitution error estimator.

d. Numerical Examples

Figures 3 and 4 display examples of the joint probability in (2.1) and the corresponding joint density in (2.2), respectively, for the resubstitution error estimator, computed using the expressions given previously.

2. Leave-one-out

We consider only the general unequal-variance case. The development here is considerably more complex than in the case of resubstitution. However, Lemma 2 still holds for the case of leave-one-out, by replacing $\hat{\varepsilon}_r$ with $\hat{\varepsilon}_l$. The probabilities required in the Lemma are given in the next Theorem, which is the counterpart of Theorem 3.

Theorem 6. *Let $X_i \sim N(\mu_0, \sigma^2)$ be i.i.d. observations for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \sigma^2)$ be i.i.d. observations for $i = n_0 + 1, \ldots, n_0 + n_1$ used to derive the classifier in (2.3). Then*

$$\begin{aligned}
P\left(\hat{\varepsilon}_l = 0, \hat{\mu} \in (a,b), \hat{\mu}_0 > \hat{\mu}_1\right) &= \sum_{m=0}^{n_0}\sum_{n=0}^{n_1}\binom{n_0}{m}\binom{n_1}{n} P(E^2_{m,n} Z_1 > \mathbf{0}) \\
P\left(\hat{\varepsilon}_l = 0, \hat{\mu} \in (a,b), \hat{\mu}_0 < \hat{\mu}_1\right) &= \sum_{m=0}^{n_0}\sum_{n=0}^{n_1}\binom{n_0}{m}\binom{n_1}{n} P(E^2_{m,n} Z_1' < \mathbf{0}) \\
P\left(\hat{\varepsilon}_l = 0, \hat{\mu} \notin (a,b), \hat{\mu}_0 < \hat{\mu}_1\right) &= \sum_{m=0}^{n_0}\sum_{n=0}^{n_1}\binom{n_0}{m}\binom{n_1}{n} \Big(P(E^1_{m,n} Z_2 < \mathbf{0}) + P(E^1_{m,n} Z_3 < \mathbf{0})\Big) \\
P\left(\hat{\varepsilon}_l = 0, \hat{\mu} \notin (a,b), \hat{\mu}_0 > \hat{\mu}_1\right) &= \sum_{m=0}^{n_0}\sum_{n=0}^{n_1}\binom{n_0}{m}\binom{n_1}{n} \Big(P(E^1_{m,n} Z_2' > \mathbf{0}) + P(E^1_{m,n} Z_3' > \mathbf{0})\Big) \\
P\left(\hat{\varepsilon}_l = 0, \hat{\mu}_0 > \hat{\mu}_1\right) &= \sum_{m=0}^{n_0}\sum_{n=0}^{n_1}\binom{n_0}{m}\binom{n_1}{n} P(E^0_{m,n} Z_4 > \mathbf{0}) \\
P\left(\hat{\varepsilon}_l = 0\right) &= \sum_{m=0}^{n_0}\sum_{n=0}^{n_1}\binom{n_0}{m}\binom{n_1}{n} \Big(P(E^0_{m,n} Z_4 > \mathbf{0}) + P(E^0_{m,n} Z_4 < \mathbf{0})\Big)
\end{aligned} \tag{2.26}$$

where $E^r_{m,n}$ is a diagonal matrix of size $2(n_0{+}n_1){+}r{+}1$, for $r = 0, 1, 2$, with diagonal elements

$(-\mathbf{1}_m, \mathbf{1}_{n_0-m}, -\mathbf{1}_m, \mathbf{1}_{n_0-m}, -\mathbf{1}_n, \mathbf{1}_{n_1-n}, -\mathbf{1}_n, \mathbf{1}_{n_1-n}, \mathbf{1}_{r+1})$. *Here Z_1 is a Gaussian random*

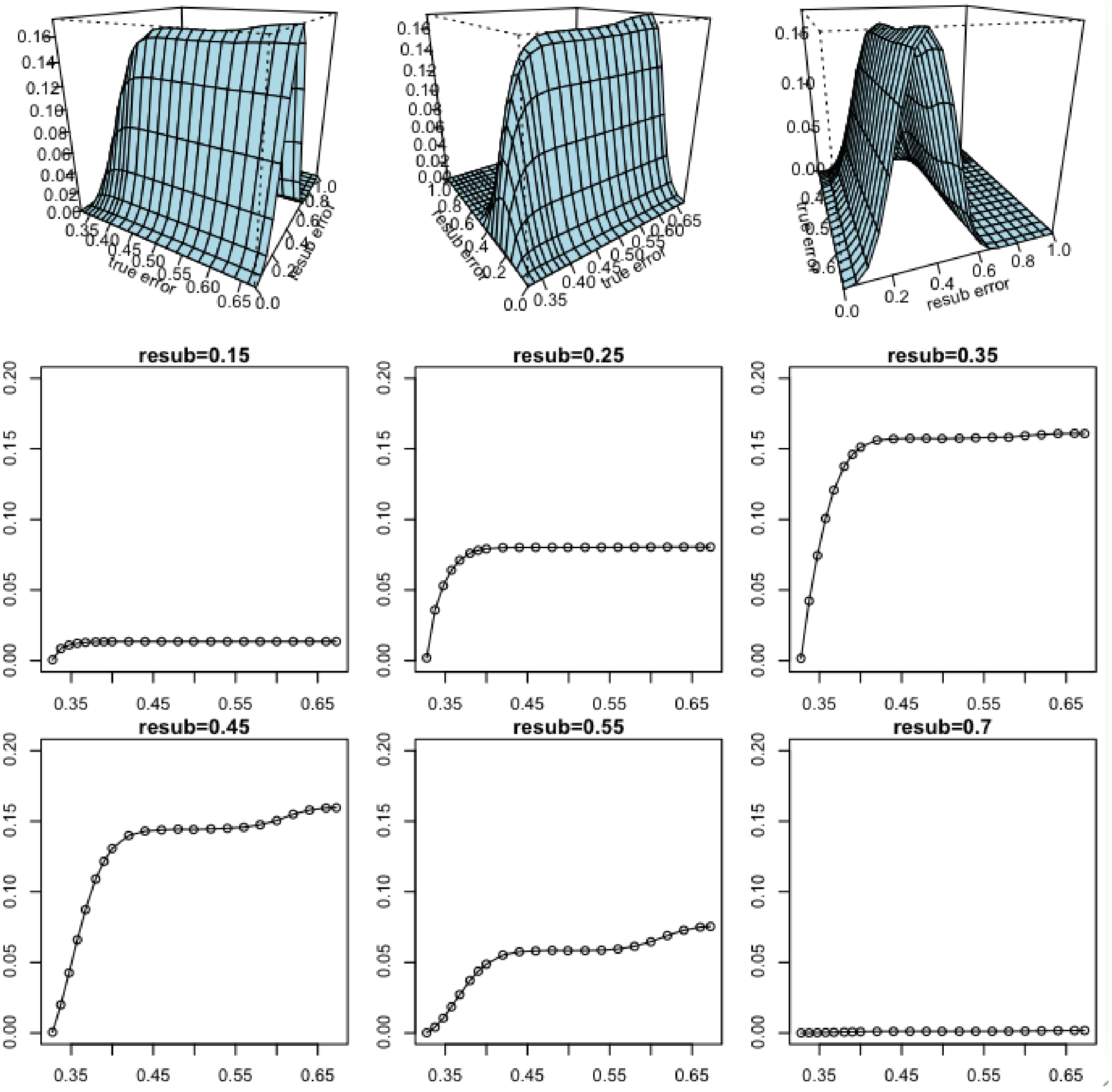

Fig. 3. Joint probability in (2.1) for the resubstitution error estimator: $n_0 = n_1 = 10$, $m_0 = 1, m_1 = 0$, $\sigma_0 = 2, \sigma_1 = 1$. Bayes error = 0.32742.

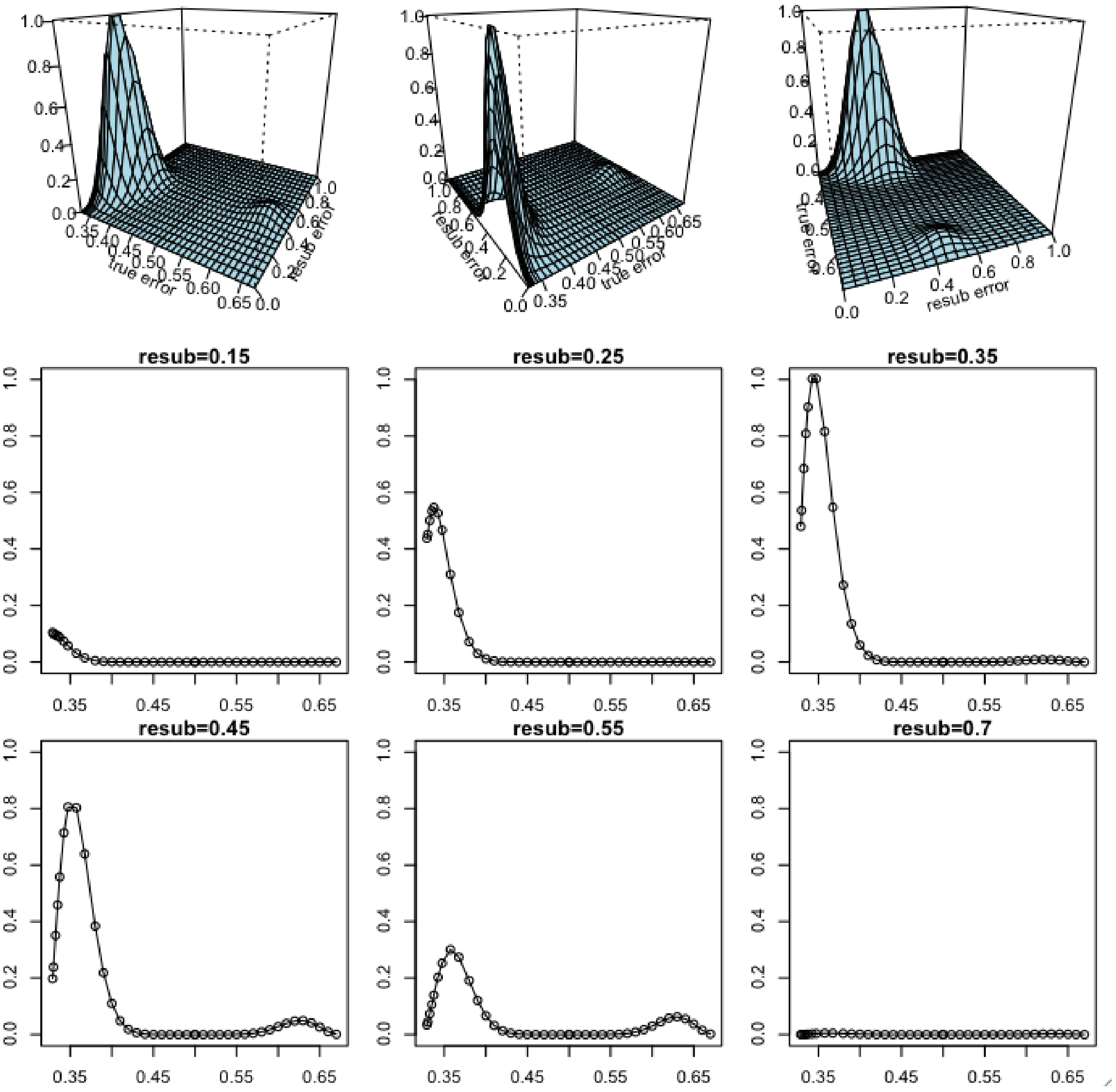

Fig. 4. Joint density in (2.2) for the resubstitution error estimator: $n_0 = n_1 = 10$, $m_0 = 1, m_1 = 0$, $\sigma_0 = 2, \sigma_1 = 1$. Bayes error = 0.32742.

vector of size $2(n_0+n_1)+3$*, with mean* μ_{Z_1} *given by:*

$$\mu_{Z_1} = \begin{bmatrix} \frac{n_0-1}{n_0}(\mu_0 - \mu_1)\mathbf{1}_{2n_0} \\ \\ \frac{n_1-1}{n_1}(\mu_0 - \mu_1)\mathbf{1}_{2n_1} \\ \\ \mu_0 - \mu_1 \\ \\ (\mu_0 + \mu_1) - 2a \\ \\ -(\mu_0 + \mu_1) + 2b \end{bmatrix}$$

and covariance matrix Σ_{Z_1} *given by*

$$\Sigma_{Z_1} = \left[\begin{array}{c|c|c} C^1 & C^2 & C^4 \\ \hline C^{2^T} & C^3 & C^5 \\ \hline C^{4^T} & C^{5^T} & C^6 \end{array}\right] \tag{2.27}$$

where

$$(C^1)_{ij} = \begin{cases} \left(4 - \frac{7}{n_0} + \frac{3}{n_0^2}\right)\sigma_0^2 + \frac{(n_0-1)^2\sigma_1^2}{n_0^2 n_1}, & i,j = 1,\ldots,n_0, i = j \\ \left(\frac{-3}{n_0} + \frac{2}{n_0^2}\right)\sigma_0^2 + \frac{(n_0-1)^2\sigma_1^2}{n_0^2 n_1}, & i,j = 1,\ldots,n_0, i \neq j \\ \frac{(n_0-1)\sigma_0^2}{n_0^2} + \frac{(n_0-1)^2\sigma_1^2}{n_0^2 n_1}, & i,j = n_0+1,\ldots,2n_0, i = j \\ \frac{(n_0-2)\sigma_0^2}{n_0^2} + \frac{(n_0-1)^2\sigma_1^2}{n_0^2 n_1}, & i,j = n_0+1,\ldots,2n_0, i \neq j \\ \frac{-(n_0-1)\sigma_0^2}{n_0^2} + \frac{(n_0-1)^2\sigma_1^2}{n_0^2 n_1}, & \begin{cases} i = n_0+1,\ldots,2n_0, j = i - n_0 \\ j = n_0+1,\ldots,2n_0, i = j - n_0 \end{cases} \\ \frac{\sigma_0^2}{n_0} + \frac{(n_0-1)^2\sigma_1^2}{n_0^2 n_1}, & otherwise \end{cases} \tag{2.28}$$

$$C^2 = \left[\frac{(n_1-1)(n_0-1)\sigma_0^2}{n_0^2 n_1} + \frac{(n_1-1)(n_0-1)\sigma_1^2}{n_1^2 n_0} \right] \mathbf{1}_{2n_0 \times 2n_1} \tag{2.29}$$

$$(C^3)_{ij} = \begin{cases} \left(4 - \frac{7}{n_1} + \frac{3}{n_1^2}\right)\sigma_1^2 + \frac{(n_1-1)^2\sigma_0^2}{n_0 n_1^2}\,, & i,j = 1,\ldots,n_1, i = j \\ \left(\frac{-3}{n_1} + \frac{2}{n_1^2}\right)\sigma_1^2 + \frac{(n_1-1)^2\sigma_0^2}{n_0 n_1^2}\,, & i,j = 1,\ldots,n_1, i \neq j \\ \frac{(n_1-1)\sigma_1^2}{n_1^2} + \frac{(n_1-1)^2\sigma_0^2}{n_0 n_1^2}\,, & i,j = n_1+1,\ldots,2n_1, i = j \\ \frac{(n_1-2)\sigma_1^2}{n_1^2} + \frac{(n_1-1)^2\sigma_0^2}{n_0 n_1^2}\,, & i,j = n_1+1,\ldots,2n_1, i \neq j \\ \frac{-(n_1-1)\sigma_1^2}{n_1^2} + \frac{(n_1-1)^2\sigma_0^2}{n_0 n_1^2}\,, & \begin{cases} i = n_1+1,\ldots,2n_1, j = i - n_1 \\ j = n_1+1,\ldots,2n_1, i = j - n_1 \end{cases} \\ \frac{\sigma_1^2}{n_1} + \frac{(n_1-1)^2\sigma_0^2}{n_0 n_1^2}\,, & otherwise \end{cases} \tag{2.30}$$

$$C^4 = \frac{(n_0-1)}{n_0} \left[(\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1})_{2n_0\times 1} \;\; (\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1})_{2n_0\times 1} \;\; (\frac{\sigma_1^2}{n_1} - \frac{\sigma_0^2}{n_0})_{2n_0\times 1} \right]_{2n_0\times 3} \tag{2.31}$$

$$C^5 = \frac{(n_1-1)}{n_1} \left[(\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1})_{2n_1\times 1} \;\; (\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1})_{2n_1\times 1} \;\; (\frac{\sigma_1^2}{n_1} - \frac{\sigma_0^2}{n_0})_{2n_1\times 1} \right]_{2n_1\times 3} \tag{2.32}$$

$$C^6 = \begin{pmatrix} (\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}) & (\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1}) & (\frac{\sigma_1^2}{n_1} - \frac{\sigma_0^2}{n_0}) \\ (\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1}) & (\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}) & -(\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}) \\ (\frac{\sigma_1^2}{n_1} - \frac{\sigma_0^2}{n_0}) & -(\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}) & (\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}) \end{pmatrix}, \tag{2.33}$$

whereas Z_1' *is a Gaussian random vector of size* $2(n_0+n_1)+3$*, obtained from* Z_1 *by multiplying by* -1 *the last two components of* Z_1*. Furthermore,* Z_2 *(resp.* Z_3*) is a Gaussian random vector of size* $2(n_0+n_1)+2$*, obtained from* Z_1 *by eliminating component* $2(n_0+n_1)+3$ *(resp.* $2(n_0+n_1)+2$*), while* Z_2' *(resp.* Z_3'*) is a Gaussian random vector of size* $2(n_0+n_1)+2$ *, obtained from* Z_1' *by eliminating component* $2(n_0+n_1)+3$ *(resp.* $2(n_0+n_1)+2$*) . Finally,* Z_4 *is Gaussian random vector of size*

$2(n_0+n_1)+1$, obtained from Z_1 by eliminating both components $2(n_0+n_1)+2$ and $2(n_0+n_1)+3$.

Proof. See Appendix.

The previous result can be extended to the case $k > 0$ by using the same reasoning employed before in connection with Theorem 2 and 4, which establishes the following result.

Theorem 7. *Under the same conditions as in Theorem 6,*

$$\begin{aligned}
&P\left(\hat{\varepsilon}_l = \tfrac{k}{n_0+n_1}, \hat{\mu} \in (a,b), \hat{\mu}_0 > \hat{\mu}_1\right) = \\
&\textstyle\sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} \sum_{p=0}^{l} \sum_{q=0}^{k-l} \binom{l}{p}\binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m}\binom{n_1-(k-l)}{n} P(E_{m,n}^{2,p,q,k,l} Z_1 > \mathbf{0}) \\[1em]
&P\left(\hat{\varepsilon}_l = \tfrac{k}{n_0+n_1}, \hat{\mu} \in (a,b), \hat{\mu}_0 < \hat{\mu}_1\right) = \\
&\textstyle\sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} \sum_{p=0}^{l} \sum_{q=0}^{k-l} \binom{l}{p}\binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m}\binom{n_1-(k-l)}{n} P(E_{m,n}^{2,p,q,k,l} Z_1' < \mathbf{0}) \\[1em]
&P\left(\hat{\varepsilon}_l = \tfrac{k}{n_0+n_1}, \hat{\mu} \notin (a,b), \hat{\mu}_0 < \hat{\mu}_1\right) = \\
&\textstyle\sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} \sum_{p=0}^{l} \sum_{q=0}^{k-l} \binom{l}{p}\binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m}\binom{n_1-(k-l)}{n} \big[P(E_{m,n}^{1,p,q,k,l} Z_2 < \mathbf{0}) \\
&\hspace{25em} + P(E_{m,n}^{1,p,q,k,l} Z_3 < \mathbf{0})\big] \\
&P\left(\hat{\varepsilon}_l = \tfrac{k}{n_0+n_1}, \hat{\mu} \notin (a,b), \hat{\mu}_0 > \hat{\mu}_1\right) = \\
&\textstyle\sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} \sum_{p=0}^{l} \sum_{q=0}^{k-l} \binom{l}{p}\binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m}\binom{n_1-(k-l)}{n} \big[P(E_{m,n}^{1,p,q,k,l} Z_2' > \mathbf{0}) \\
&\hspace{25em} + P(E_{m,n}^{1,p,q,k,l} Z_3' > \mathbf{0})\big] \\
&P\left(\hat{\varepsilon}_l = \tfrac{k}{n_0+n_1}, \hat{\mu}_0 > \hat{\mu}_1\right) = \\
&\textstyle\sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} \sum_{p=0}^{l} \sum_{q=0}^{k-l} \binom{l}{p}\binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m}\binom{n_1-(k-l)}{n} P(E_{m,n}^{0,p,q,k,l} Z_4 > \mathbf{0}) \\[1em]
&P\left(\hat{\varepsilon}_l = \tfrac{k}{n_0+n_1}\right) = \\
&\textstyle\sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} \sum_{p=0}^{l} \sum_{q=0}^{k-l} \binom{l}{p}\binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m}\binom{n_1-(k-l)}{n} \big[P(E_{m,n}^{0,p,q,k,l} Z_4 > \mathbf{0}) \\
&\hspace{25em} + P(E_{m,n}^{0,p,q,k,l} Z_4 < \mathbf{0})\big]
\end{aligned} \tag{2.34}$$

where the vectors Z_i, $i = 1, \ldots, 4$, and Z'_i, $i = 1, \ldots, 3$, are defined in Theorem 6, and $E^{r,p,q,k,l}_{m,n}$ is a diagonal matrix of size $2(n_0 + n_1) + r + 1$ with diagonal elements given by the component-wise product of the vectors E_1 and E_2 where:

$$\begin{aligned} E_1 &= (-\mathbf{1}_p, \mathbf{1}_{n_0}, -\mathbf{1}_{l-p}, \mathbf{1}_{n_0-l}, -\mathbf{1}_q, \mathbf{1}_{n_1}, -\mathbf{1}_{k-l-q}, -\mathbf{1}_{n_1-k+l}, \mathbf{1}_{r+1}) \\ E_2 &= (-\mathbf{1}_l, \mathbf{1}_m, -\mathbf{1}_{n_0-m}, \mathbf{1}_m, -\mathbf{1}_{n_0-l-m}, -\mathbf{1}_{k-l}, \mathbf{1}_n, -\mathbf{1}_{n_1-n}, \mathbf{1}_n, -\mathbf{1}_{n_1-k+l-n}, \mathbf{1}_{r+1}) \end{aligned} \tag{2.35}$$

Theorem 7, in conjunction with Lemma 2, with $\hat{\varepsilon}_r$ replaced by with $\hat{\varepsilon}_l$, allows the exact computation of the joint probability in (2.1) for the leave-one-out error estimator.

a. Joint Density

As in the resubstitution case, it is possible to apply a methodology similar to the one in the previous sections to obtain the joint density in (2.2) for the leave-one-out error estimator. As mentioned previously, Lemma 2 still holds for the case of leave-one-out, by replacing $\hat{\varepsilon}_r$ with $\hat{\varepsilon}_l$, whereas the following result is the counterpart of Theorem 5. The proof of this theorem is similar to the proof of Theorem 5 and is thus omitted.

Theorem 8. *Under the same conditions as in Theorem 6,*

$$\begin{aligned} & p\left(\hat{\varepsilon}_l = \tfrac{k}{n_0+n_1}, \hat{\mu} = a, \hat{\mu}_0 > \hat{\mu}_1\right) = \\ & \textstyle\sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} \sum_{p=0}^{l} \sum_{q=0}^{k-l} \binom{l}{p}\binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m}\binom{n_1-(k-l)}{n} P(E^{0,p,q,k,l}_{m,n} Y > \mathbf{0}) \\ & \qquad\qquad\qquad\qquad \times \varphi\left(0, \mu_0 + \mu_1 - 2a, \tfrac{\sigma_0^2}{n_0} + \tfrac{\sigma_1^2}{n_1}\right) \\ & p\left(\hat{\varepsilon}_l = \tfrac{k}{n_0+n_1}, \hat{\mu} = a, \hat{\mu}_0 < \hat{\mu}_1\right) = \\ & \textstyle\sum_{l=0}^{k} \binom{n_0}{l}\binom{n_1}{k-l} \sum_{p=0}^{l} \sum_{q=0}^{k-l} \binom{l}{p}\binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m}\binom{n_1-(k-l)}{n} P(E^{0,p,q,k,l}_{m,n} Y < \mathbf{0}) \\ & \qquad\qquad\qquad\qquad \times \varphi\left(0, \mu_0 + \mu_1 - 2a, \tfrac{\sigma_0^2}{n_0} + \tfrac{\sigma_1^2}{n_1}\right) \end{aligned} \tag{2.36}$$

in which $E^{0,p,q,k,l}_{m,n}$ is the diagonal matrix used in Theorem 7, and Y is a Gaussian *random vector of size* $2(n_0+n_1)+1$ *with mean* μ_Y *given by:*

$$\mu_Y = 2\tfrac{n_1\sigma_0^2(a-\mu_1)-n_0\sigma_1^2(a-\mu_0)}{n_1\sigma_0^2+n_0\sigma_1^2}\left[\tfrac{(n_0-1)}{n_0}\mathbf{1}_{2n_0}^T \; \tfrac{(n_1-1)}{n_1}\mathbf{1}_{2n_1}^T \; 1\right]^T_{(2n_0+2n_1+1)\times 1} \tag{2.37}$$

and covariance matrix Σ_Y *given by*

$$\Sigma_Y = \Sigma_{Y_{11}} - \frac{1}{n_0 n_1}\frac{(n_1\sigma_0^2 - n_0\sigma_1^2)^2}{n_1\sigma_0^2 + n_0\sigma_1^2} H_{(2n_0+2n_1+1)\times(2n_0+2n_1+1)} \tag{2.38}$$

where

$$\Sigma_{Y_{11}} = \left[\begin{array}{c|c|c} C^1 & C^2 & \mathrm{ad}\mathbf{1}_{2n_0} \\ \hline {C^2}^T & C^3 & \mathrm{cd}\mathbf{1}_{2n_1} \\ \hline \mathrm{ad}\mathbf{1}_{2n_0}^T & \mathrm{cd}\mathbf{1}_{2n_1}^T & \mathrm{d} \end{array}\right] \tag{2.39}$$

and

$$H = \left[\begin{array}{c|c|c} a^2\mathbf{1}_{2n_0\times 2n_0} & b\mathbf{1}_{2n_0\times 2n_1} & a\mathbf{1}_{2n_0} \\ \hline b\mathbf{1}_{2n_1\times 2n_0} & c^2\mathbf{1}_{2n_1\times 2n_1} & c\mathbf{1}_{2n_1} \\ \hline a\mathbf{1}_{2n_0}^T & c\mathbf{1}_{2n_1}^T & 1 \end{array}\right] \tag{2.40}$$

with $C^i, i = 1,2,3$ *as defined in theorem 6, and* $\mathrm{a} = \frac{(n_0-1)}{n_0}$, $\mathrm{b} = \frac{(n_0-1)(n_1-1)}{n_0 n_1}$, $\mathrm{c} = \frac{(n_1-1)}{n_1}$, *and* $\mathrm{d} = (\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1})$.

Theorem 8, in conjunction with Lemma 3, with $\hat{\varepsilon}_r$ replaced by with $\hat{\varepsilon}_l$, allows the exact computation of joint density in (2.2) for the leave-one-out error estimator.

b. Numerical Examples

Figures 5 and 6 display examples of the joint probability in (2.1) and the corresponding joint density in (2.2), respectively, for the leave-one-out error estimator, computed using the expressions given previously. Comparing these figures to Figures 3 and 4, one observes, among other interesting facts, that there is in the present case more probability mass at large values of the error estimator, as expected due to the generally larger variance of leave-one-out with respect to resubstitution.

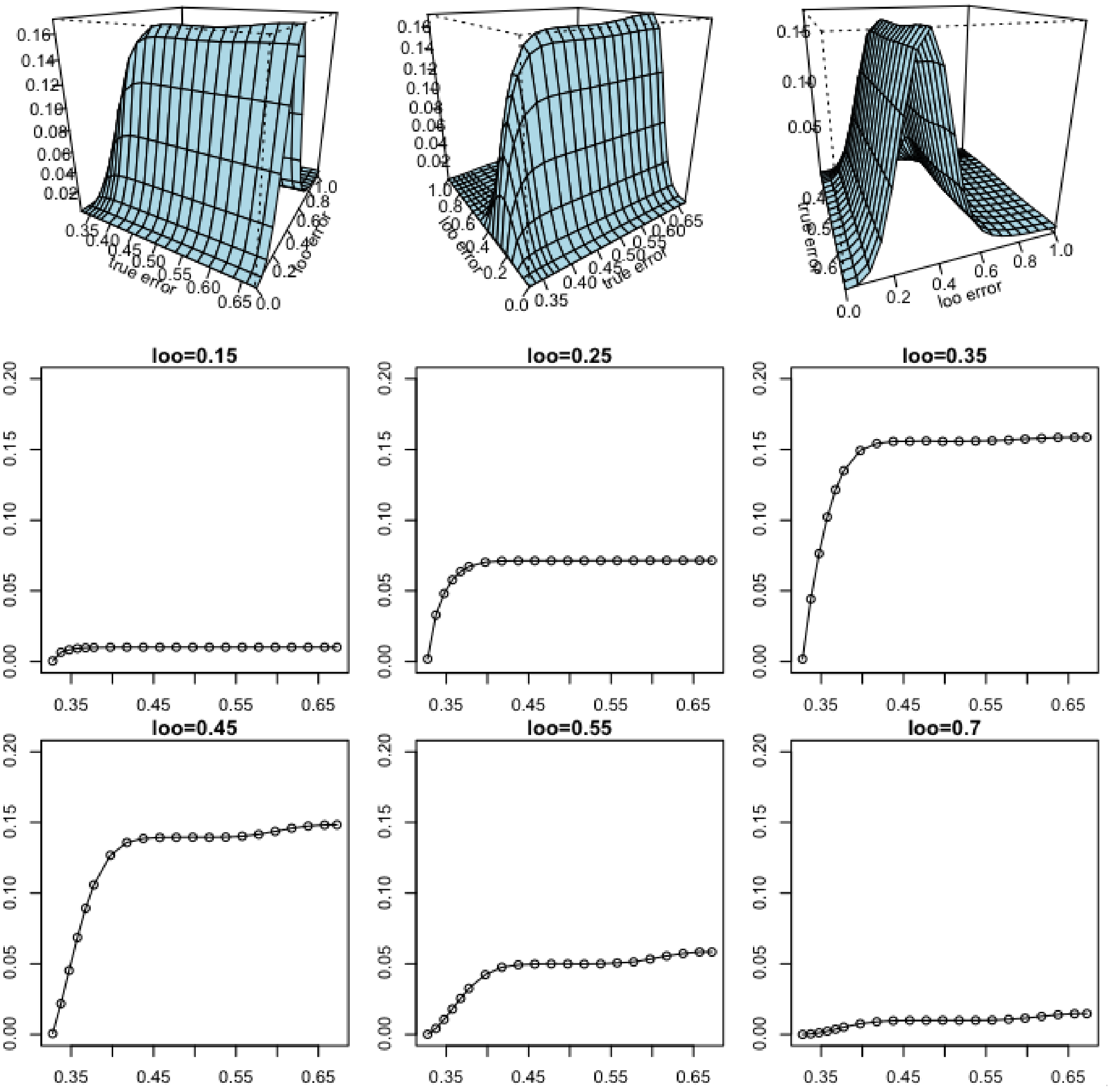

Fig. 5. Joint probability in (2.1) for the leave-one-out error estimator: $n_0 = n_1 = 10$, $m_0 = 1, m_1 = 0$, $\sigma_0 = 2, \sigma_1 = 1$. Bayes error = 0.32742.

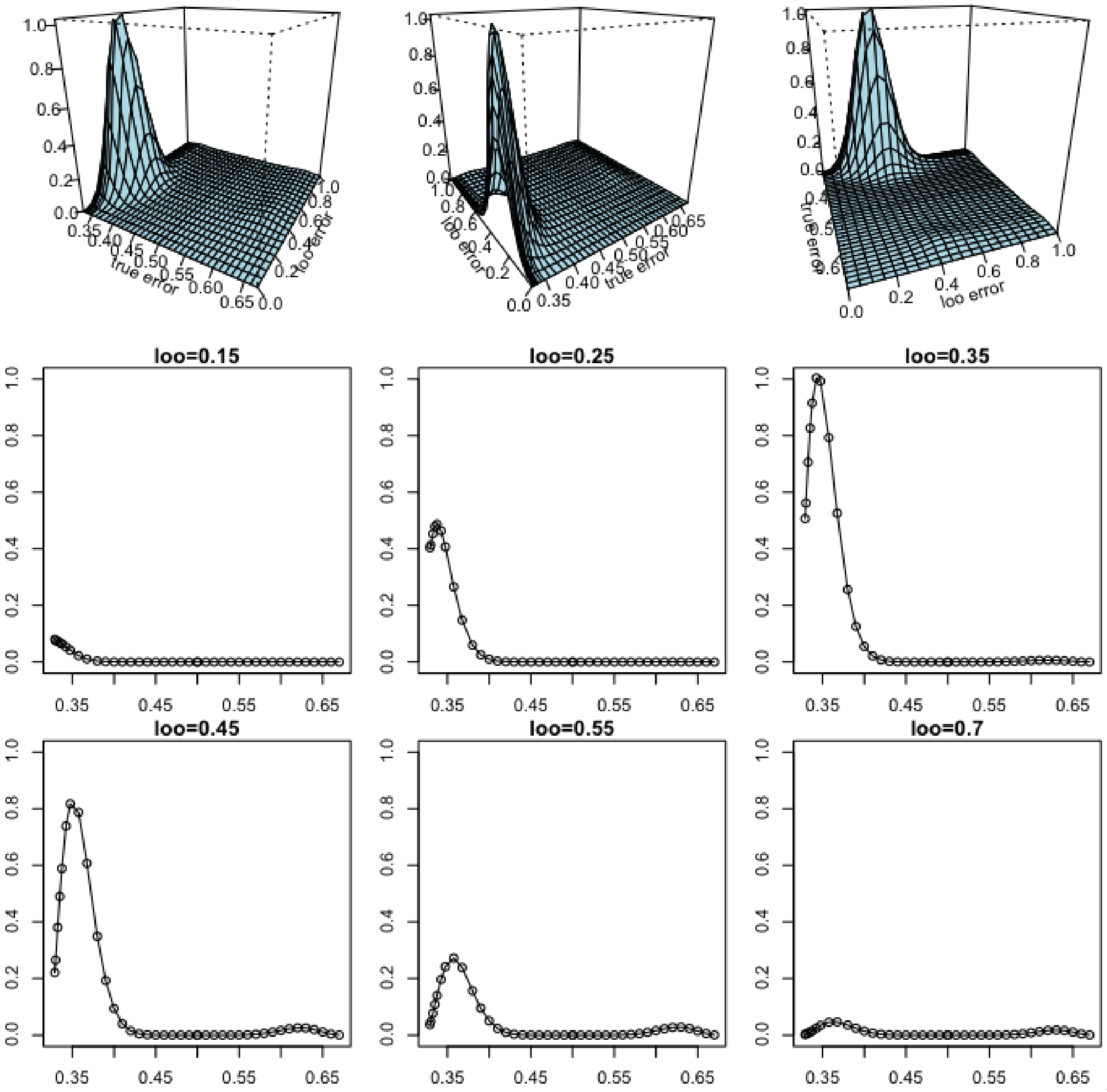

Fig. 6. Joint density in (2.2) for the leave-one-out error estimator: $n_0 = n_1 = 10$, $m_0 = 1, m_1 = 0$, $\sigma_0 = 2, \sigma_1 = 1$. Bayes error = 0.32742.

B. Multivariate Case

Consider now a set of $n = n_0 + n_1$ independent distributed samples, where n_0 samples $\{X_1, X_2, \ldots X_{n_0}\}$ come from the multivariate Gaussian distribution $N(\mu_0, \Sigma_{p\times p})$, and n_1 samples denoted by $\{X_{n_0+1}, X_{n_0+2}, \ldots X_{n_0+n_1}\}$ come from the multivariate Gaussian distribution $N(\mu_1, \Sigma_{p\times p})$, where μ_0 and μ_1 are arbitrary $p\times 1$ mean vectors and $\Sigma_{p\times p}$ is a covariance matrix common to both classes. The approach used in deriving the joint distribution of actual and estimated errors in the univariate case is not applicable here; however, we will employ an approximation method, which is based on the previously derived exact expressions for the univariate case.

This is done by using the Fisher discriminant $w = \Sigma^{-1}(\mu_0 - \mu_1)$ to project the data to the real line, which gives the maximum separation possible between the classes, and then we use the exact results stated in previous section on the resultant distributions, namely, the univariate Gaussian distributions $N(\eta_0, \Delta^2)$ and $N(\eta_1, \Delta^2)$, where

$$\eta_i = (\mu_0 - \mu_1)^T \Sigma^{-1} \mu_i \,, \qquad \Delta^2 = (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1)$$

for $i = 0, 1$.

1. Numerical Examples

In Figure 7, we have assumed mean vectors of opposite signs $\mu_0 = m_0 = d\mathbf{1}_{p\times 1}$ and $\mu_1 = m_1 = -d\mathbf{1}_{p\times 1}$, and covariance Σ matrix with variance 1 on diagonal and correlation r for the off-diagonal elements, where $|r| \leq 1$. The MC approximation uses 3×10^6 random samples.

Differences between the proposed approximation and the MC approximation arise in two cases. In the first case, they are different for values of actual error very close to Bayes error. This could happen because the MC approximation is poor very close to Bayes error, since there are not enough MC samples that can be used in that case. However, this case is not so important anyway, given that the actual classification error usually is not this close to the Bayes error. In the second case, they differ as the value of n/p becomes smaller. We have observed that the proposed approximation is less accurate in such small-sample settings. For fixed n/p, the proposed approximation is better for smaller Bayes error.

C. Conditional Bounds and Regression for the Actual Error Given the Estimated Error

A problem of great importance in practice is to bound the actual classification error given the observed value of the error estimator, which is akin to finding confidence intervals in classical parameter estimation. In addition, great insight can be obtained by finding the expected classification error conditioned on the observed value of the error estimator, which contains "local" information on the accuracy of the classification rule, as opposed to the "global" information contained in the unconditional expected classification error. These are called, respectively, *conditional bounds* and *regression* of the actual error given the observed error estimated error, and they can be readily computed given the knowledge of the joint distribution of actual and estimated error, as detailed in the sequel.

Given the knowledge of the joint probability in (2.1), one can write the conditional distribution of the actual error given the estimated error as

$$P\left(\varepsilon < z \,\middle|\, \hat{\varepsilon} = \frac{k}{n_0+n_1}\right) = P\left(\hat{\varepsilon} = \frac{k}{n_0+n_1}, \varepsilon < z\right) \Big/ P\left(\hat{\varepsilon} = \frac{k}{n_0+n_1}\right), \quad k = 0, 1, \ldots, n_0+n_1 \tag{2.41}$$

provided that the denominator $P\left(\hat{\varepsilon} = \frac{k}{n_0+n_1}\right)$ is nonzero (this probability is determined by Theorems 2, 4, or 6).

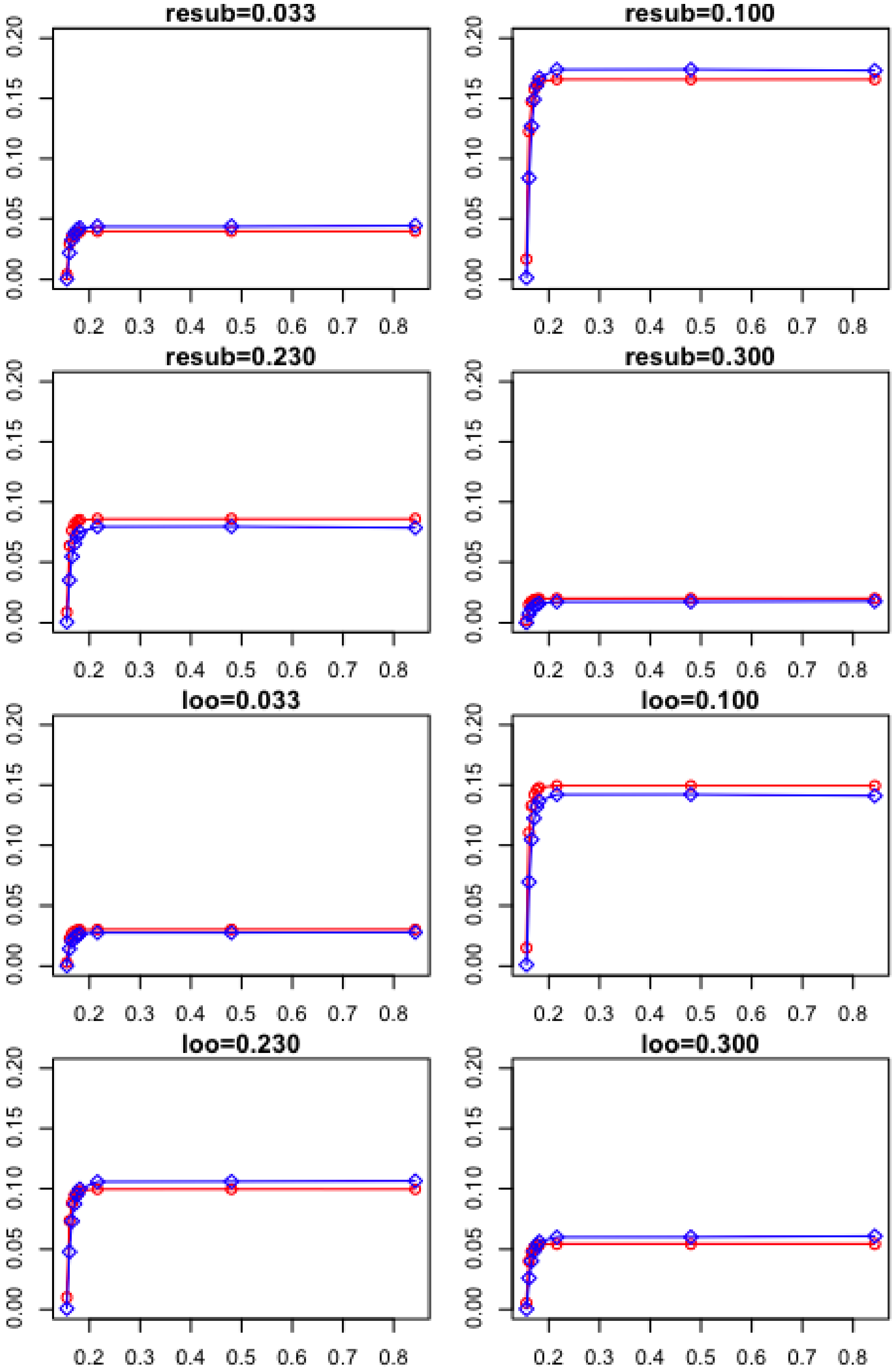

Fig. 7. Joint probability in (2.1) for the resubstitution (top panels) and leave-one-out (bottom panels) in the multivariate case: $n_0 = n_1 = 15$, $m_0 = m_1 = -\mathrm{d}\mathbf{1}_{p\times 1}$, d = 0.75, $r = 0.1$, $p = 2$. Bayes error = 0.1559. Legend key: proposed approximation (∘), MC approximation (⋄).

To find an exact $100(1-\alpha)\%$ upper bound on the actual error given the resubstitution estimate, we would like to find z_α such that

$$P\left(\varepsilon < z_\alpha \,\middle|\, \hat{\varepsilon} = \frac{k}{n_0 + n_1}\right) = 1 - \alpha \tag{2.42}$$

The value z_α can be found by means of a simple one-dimensional search.

As for the regression, from the conditional distribution in (2.41), one can obtain the conditional expectation of the actual error given the error estimator, via

$$E\left(\varepsilon \,\middle|\, \hat{\varepsilon}_r = \frac{k}{n_0 + n_1}\right) = \int_0^1 \left(1 - P\left(\varepsilon < z \,\middle|\, \hat{\varepsilon}_r = \frac{k}{n_0 + n_1}\right)\right) dz \tag{2.43}$$

by using the fact that $E[X] = \int P(X > z)dz$ for any nonnegative random variable X.

Figure 8 illustrates the exact 95% upper conditional bound and regression in the univariate case, using the expressions for the joint probability in (2.1) obtained previously, whereas Figure 9 provides similar examples in the bivariate case ($p = 2$), using the proposed approximation for the joint probability in (2.1) developed previously. The total number of sample points is kept to 20 to facilitate computation. In the multivariate case, we have assumed mean vectors of opposite signs $\mu_0 = m_0 = d\mathbf{1}_{p\times 1}$ and $\mu_1 = m_1 = -d\mathbf{1}_{p\times 1}$, and covariance matrix Σ with variance 1 on the diagonal and correlation r for the off-diagonal elements, where $|r| \leq 1$. In all examples, the conditional bounds and regression are calculated for only those values of the error estimate such that $P(\hat{\varepsilon} = \frac{k}{n_0+n_1}) > 0.001$. In particular, note that the latter probability is displayed in the plots to show the concentration of mass of the observed error estimates. Values of the error estimate of very small probability (< 0.001) are difficult to handle owing to poor accuracy of the required Gaussian probability computations and are therefore avoided here (these cases are very rare and thus of little practical importance in any case); nonetheless, such cases could be obtained at the expense of more computational work.

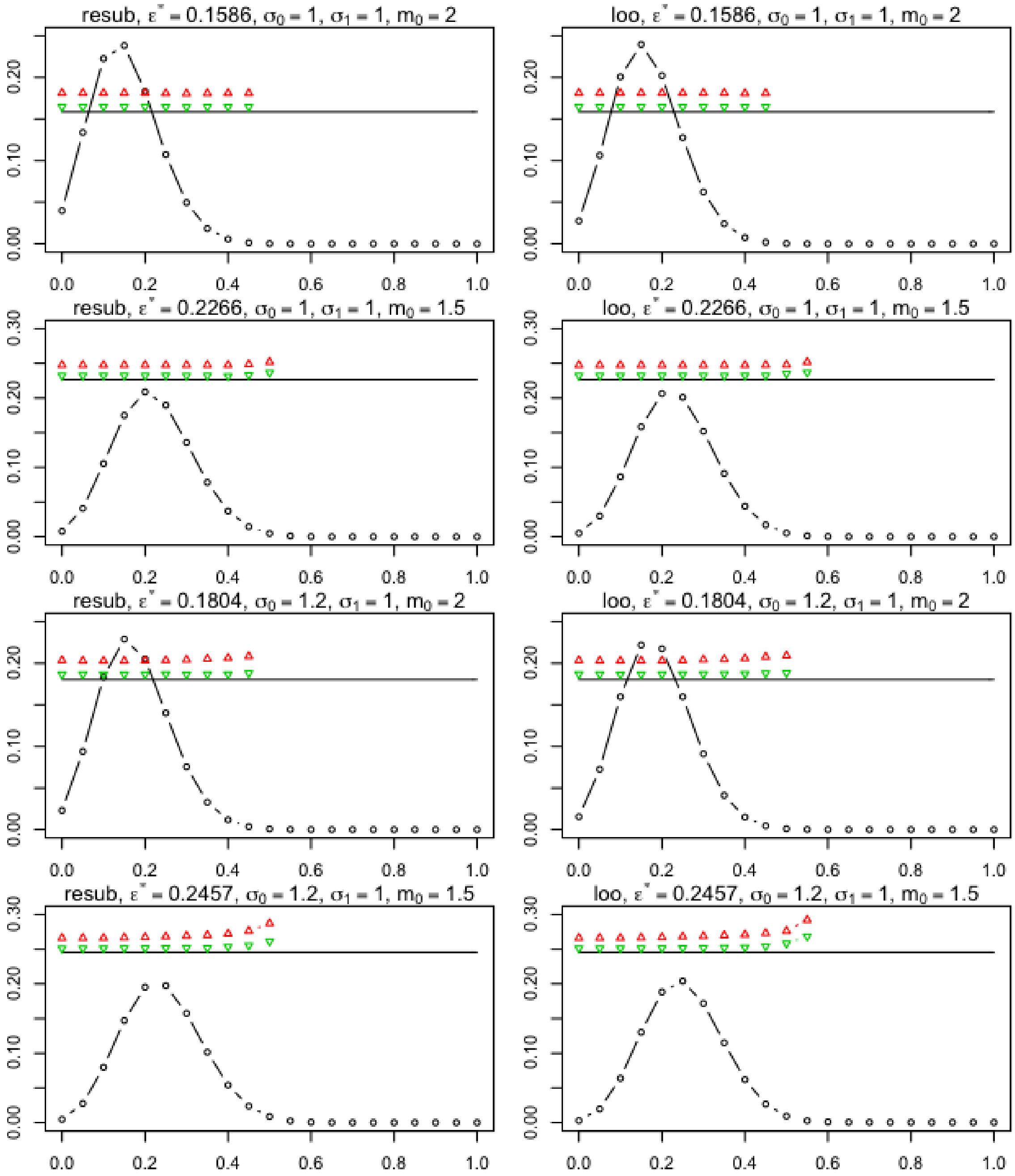

Fig. 8. The 95% upper conditional bounds and regression of actual error given the resubstitution and leave-one-out error estimates in the univariate case. In all cases, $m_1 = 0$. The horizontal solid line displays the Bayes error. The marginal probability mass function for the error estimators in each case is also plotted for reference. Legend key: 95% upper conditional bound ($\triangle$), regression (∇), probability mass function ($\circ$).

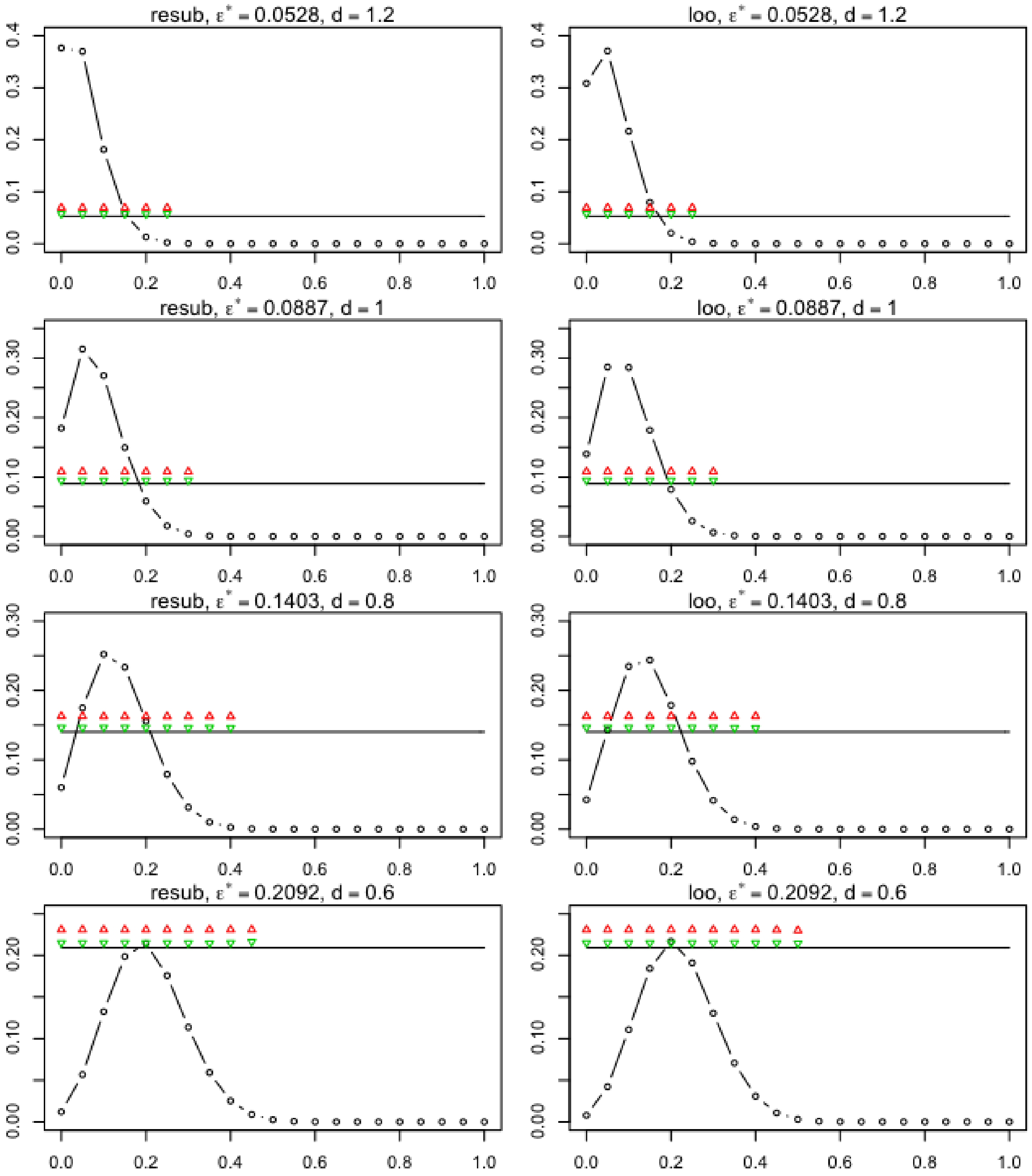

Fig. 9. The 95% upper conditional bounds and regression of actual error given the resubstitution and leave-one-out error estimates in the bivariate case: $m_0 = -m_1 = \mathrm{d}\mathbf{1}_p$, $r = 0.1$, $p = 2$. The marginal probability mass function for the error estimators in each case is also plotted for reference. The horizontal solid line displays the Bayes error. Legend key: 95% upper conditional bound ($\triangle$), regression (∇), probability mass function ($\circ$).

Figure 10 presents univariate and bivariate examples derived from gene-expression data from a recently-published breast cancer study [118]. Discrimination is between good (class 0) vs. bad (class 1) prognosis. A subset of 30 samples was randomly selected among the total of 295 included in the aforementioned study, with $n_0 = 12$ and $n_1 = 18$ to reflect the proportion between classes observed in the full data set, and corresponding normalized gene expression measurements were extracted for the genes "LOC51203" and "FGF18." Those are the top genes according to both the t-test and fold change. Univariate and bivariate Shapiro-Wilk tests (using the R statistical software) applied on the full data set, for more sensitivity, did not reject Gaussianity of these genes, either individually or as a pair, over either of the classes at a 95% significance level. Sample means and variances (the pooled covariance matrix was used in the bivariate case) were used as estimates of the unknown true means and variances.

These results confirm the lack of regression for small-sample error estimation observed in the simulation study in [105], as one can see in the figures that both the confidence bounds and the nonlinear regressions are virtually horizontal, except for a slight bit of upward movement at the extreme right, where there is very little error-estimator mass and therefore negligible practical significance. This means that the error estimate provides essentially no information regarding the error as in practically useless, both for predicting the actual error or bounding it with confidence in the small-sample setting for this Gaussian model. As might be expected, the situation is worse with two features as opposed to one, but there is virtually no regression in either case. This is a very small sample, a total of 20 sample points, but the number of features is also very small. Consider the much larger numbers of features often used in practice and consider the much more complex classification rules being employed. These results provide analytic support for the synthetic results obtained in [105].

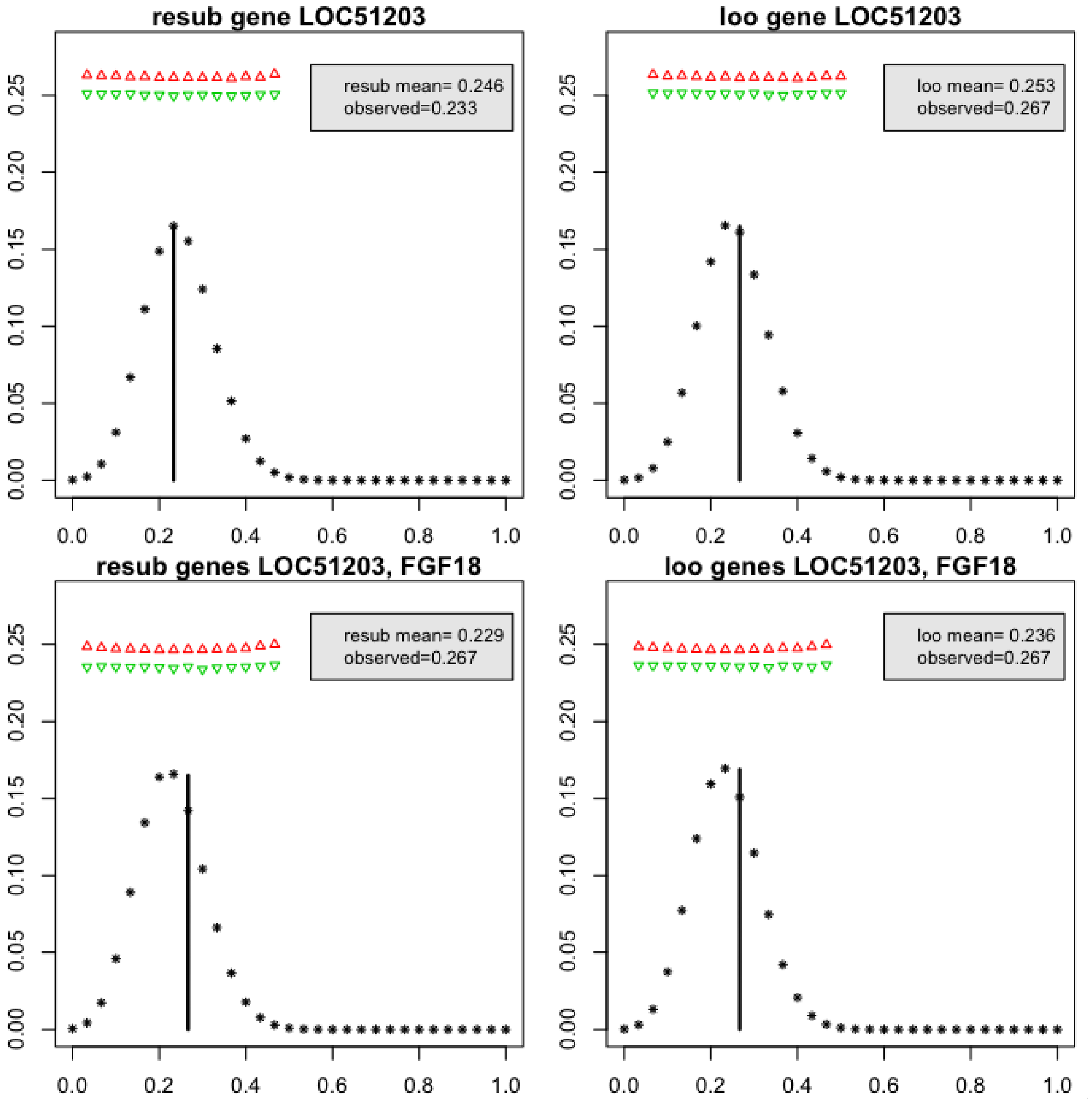

Fig. 10. The 95% upper conditional bounds and regression of actual error given the resubstitution and leave-one-out error estimates in the univariate case (top row) and bivariate case (bottom row), for distributional parameters estimated from gene-expression data (see text). The marginal probability mass function for the error estimators in each case is also plotted for reference. The observed error estimates in each case are printed and indicated by a vertical bar, and the expected error estimates based on the estimated distributions are also printed. Legend key: 95% upper conditional bound (△), regression (▽), probability mass function (∘).

D. Conclusion

This chapter contributes to the analytical study of classification error estimation for LDA under a Gaussian model, a subject with a long history in Pattern Recognition and Statistics. It presents, for what is believed to be the first time, the analytical formulation for the joint sampling distribution of the actual and estimated errors of a classification rule. Here, we considered the resubstitution and leave-one-out error estimators; we remark however that the same methodology could in principle be employed to derive similar results for other error estimators. We provide here exact results in the univariate case, and suggest a simple method to obtain an accurate approximation in the multivariate case. We also showed how these results can be applied in the computation of condition bounds and the regression of the actual error, given the observed error estimate. In contrast to asymptotic results, the analysis presented here is applicable to finite training data. In particular, it applies in the small-sample settings commonly found in genomics and proteomics applications.

In practice the unknown parameters of the Gaussian distributions, which figure in the expressions, are not known and need to be estimated. Using the usual maximum-likelihood estimates for such parameters and plugging them into the theoretical exact expressions provides a sample-based approximation to the joint distribution, and also sample-based methods to estimate upper conditional bounds on the actual error; this approach was employed in the numerical example based on gene-expression data of Section C. As the ML estimators are consistent and all expressions are smooth, these sample-based approximations will converge to the actual values as sample size increases without bound.

CHAPTER III

ANALYTIC STUDY OF PERFORMANCE OF ERROR ESTIMATORS FOR LINEAR DISCRIMINANT ANALYSIS THROUGH RMS – UNIVARIATE MODEL

In this chapter we derive exact analytical expressions for the bias, variance, and RMS for the resubstitution and leave-one-out error estimators in the case of linear discriminant analysis in the univariate Gaussian model. Sample sizes for the two classes need not be the same. The mean resubstitution and leave-one-out errors are represented by probabilities involving bivariate Gaussian distributions. Their second moments and cross-moments with the actual error are represented by 4-variate Gaussian distributions. From these, the bias, variance, and RMS for resubstitution and leave-one-out as estimators of the actual error can be computed. At the end, one practical use of these results on the gene-expression data is discussed.

A. Criteria of Performance of Error Estimation

The widely-adopted metrics for performance of an error estimator $\hat{\epsilon}$ of the actual classifier error ϵ are the:

- Bias:

$$\mathrm{Bias}[\hat{\epsilon}] = E[\hat{\epsilon}] - E[\epsilon] \tag{3.1}$$

- Deviation Variance

$$\mathrm{Var}_d[\hat{\epsilon}] = \mathrm{Var}(\hat{\epsilon} - \epsilon) = \mathrm{Var}(\epsilon) + \mathrm{Var}(\hat{\epsilon}) - 2\mathrm{Cov}(\epsilon, \hat{\epsilon}) \tag{3.2}$$

- RMS:

$$\mathrm{RMS}[\hat{\epsilon}] = \sqrt{E[(\epsilon - \hat{\epsilon})^2]} = \sqrt{E[\epsilon^2] + E[\hat{\epsilon}^2] - 2E[\epsilon\hat{\epsilon}]} \tag{3.3}$$

The bias and the deviation variance measure respectively the average centrality and dispersion of the error estimator in relation to the actual error. The ideal estimator is unbiased and has minimum variance. However, the usual bias-variance dilemma applies; for example, the resubstitution error estimator generally has small variance but is often optimistically biased, whereas the the leave-one-out error estimator is nearly unbiased, but generally has large variance. As can be easily checked, the RMS combines these two complementary criteria into a single metric:

$$\text{RMS}[\hat{\epsilon}] = \sqrt{\text{Bias}[\hat{\epsilon}]^2 + \text{Var}_d[\hat{\epsilon}]} \tag{3.4}$$

In fact, this implies that any one of the three criteria can be obtained by knowledge of the other two. In particular, the variance of deviation is given by:

$$\text{Var}_d[\hat{\epsilon}] = \text{RMS}[\hat{\epsilon}]^2 - \text{Bias}[\hat{\epsilon}]^2 \tag{3.5}$$

From the above discussion, it becomes clear that the bias, variance, and RMS can be obtained with the knowledge of the first moments $E[\epsilon]$ and $E[\hat{\epsilon}]$, the second moments $E[\epsilon^2]$ and $E[\hat{\epsilon}^2]$, and the cross moment $E[\epsilon\hat{\epsilon}]$. In this section, we write down these moments in terms of probabilities involving the discriminant $W(\bar{X}_0, \bar{X}_1, X)$. Note that all the formulas in this section are not exclusive to the Gaussian case, but apply in general. We will write all equations for the resubstitution estimator; the corresponding equations for the leave-one-out estimator can be obtained by simply replacing $W(\bar{X}_0, \bar{X}_1, X_i)$ by $W^{(i)}(\bar{X}_0, \bar{X}_1, X_i)$ throughout.

1. First Moment of the Actual Error

We restate (3.6) here:

$$E[\epsilon] = \alpha_0 P(W(\bar{X}_0, \bar{X}_1, X) \leq 0 \mid X \in \Pi_0) + \alpha_1 P(W(\bar{X}_0, \bar{X}_1, X) > 0 \mid X \in \Pi_1)$$

2. Second Moment of the Actual Error

Here, we have restated the theorem proved in [24] to find the second moments of true error. We will employ this theorem it in the context of LDA. From (1.4), we have that:

$$E[\epsilon^2] = E\left[(\alpha_0\epsilon^0 + \alpha_1\epsilon^1)^2\right] = \alpha_0^2 E[\epsilon^0\epsilon^0] + 2\alpha_0\alpha_1 E[\epsilon^0\epsilon^1] + \alpha_1^2 E[\epsilon^1\epsilon^1] \tag{3.6}$$

It follows that

$$\begin{aligned} E[\epsilon^0\epsilon^0] &= E[P(W(\bar{X}_0,\bar{X}_1,X)\leq 0|X \in \Pi_0,\bar{X}_0,\bar{X}_1)P(W(\bar{X}_0,\bar{X}_1,X')\leq 0|X'\in\Pi_0,\bar{X}_0,\bar{X}_1)] \\ &= E[P(W(\bar{X}_0,\bar{X}_1,X)\leq 0, W(\bar{X}_0,\bar{X}_1,X')\leq 0|X,X'\in\Pi_0,\bar{X}_0,\bar{X}_1)] \\ &= P(W(\bar{X}_0,\bar{X}_1,X)\leq 0, W(\bar{X}_0,\bar{X}_1,X')\leq 0|X,X'\in\Pi_0) \end{aligned} \tag{3.7}$$

Similar expressions obtain for the other terms in (3.6), namely $E[\epsilon^0\epsilon^1]$ and $E[\epsilon^1\epsilon^1]$. In all,

$$\begin{aligned} E[\epsilon^2] = {} & \alpha_0^2 P(W(\bar{X}_0,\bar{X}_1,X)\leq 0, W(\bar{X}_0,\bar{X}_1,X')\leq 0 \mid X,X'\in\Pi_0) \\ & + 2\alpha_0\alpha_1 P(W(\bar{X}_0,\bar{X}_1,X)\leq 0, W(\bar{X}_0,\bar{X}_1,X') > 0 \mid X\in\Pi_0, X'\in\Pi_1) \\ & + \alpha_1^2 P(W(\bar{X}_0,\bar{X}_1,X) > 0, W(\bar{X}_0,\bar{X}_1,X') > 0 \mid X,X'\in\Pi_1) \end{aligned} \tag{3.8}$$

3. First Moment of the Estimated Error

From (1.8), we have that:

$$E[\hat{\epsilon}^r] = \hat{\alpha}_0 E[\hat{\epsilon}_0^r] + \hat{\alpha}_1 E[\hat{\epsilon}_1^r] = \hat{\alpha}_0 P(W(\bar{X}_0,\bar{X}_1,X_1)\leq 0) + \hat{\alpha}_1 P(W(\bar{X}_0,\bar{X}_1,X_{n_0+1}) > 0) \tag{3.9}$$

The corresponding equation for leave-one-out is obtained by replacing $W(\bar{X}_0,\bar{X}_1,X_1)$ and $W(\bar{X}_0,\bar{X}_1,X_{n_0+1})$ by $W^{(1)}(\bar{X}_0,\bar{X}_1,X_1)$ and $W^{(n_0+1)}(\bar{X}_0,\bar{X}_1,X_{n_0+1})$, respectively. Note that $W^{(1)}(\bar{X}_0,\bar{X}_1,X_1)$ is distributed as $W'(\bar{X}_0',\bar{X}_1,X)$ conditioned on $X\in\Pi_0$, where W' and $\bar{X}'$ are the usual W and $\bar{X}_0$ in the case where there are $n_0 - 1$ samples in class 0 and n_1 samples in class 1. An analogous comment applies to $W^{(n_0+1)}(\bar{X}_0, \bar{X}_1, X_{n_0+1})$. By virtue of (3.6), this leads to the well-known fact that $E[\hat{\epsilon}^l] = E[\epsilon_{n-1}]$, provided that $\hat{\alpha}_i = \alpha_i$, for $i = 0, 1$.

4. Second Moment of the Estimated Error

From (1.8), we have that:

$$\begin{aligned}
E[(\hat{\epsilon}^r)^2] &= E\left[(\hat{\alpha}_0\hat{\epsilon}_0^r \;+\; \hat{\alpha}_1\hat{\epsilon}_1^r)^2\right] \\
&= \hat{\alpha}_0^2 E[(\hat{\epsilon}_0^r)^2] + 2\hat{\alpha}_0\hat{\alpha}_1 E[\hat{\epsilon}_0^r\hat{\epsilon}_1^r] + \hat{\alpha}_1^2 E[(\hat{\epsilon}_1^r)^2] \\
&= \hat{\alpha}_0^2\, E\left[\frac{1}{n_0^2}\sum_{i=1}^{n_0}\sum_{j=1}^{n_0} I_{\{W(\bar{X}_0,\bar{X}_1,X_i)\le 0, W(\bar{X}_0,\bar{X}_1,X_j)\le 0\}}\right] \\
&+\; 2\hat{\alpha}_0\hat{\alpha}_1\, E\left[\frac{1}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=n_0+1}^{n_0+n_1} I_{\{W(\bar{X}_0,\bar{X}_1,X_i)\le 0, W(\bar{X}_0,\bar{X}_1,X_j)> 0\}}\right] \\
&+\; \hat{\alpha}_1^2\, E\left[\frac{1}{n_1^2}\sum_{i=n_0+1}^{n_0+n_1}\sum_{j=n_0+1}^{n_0+n_1} I_{\{W(\bar{X}_0,\bar{X}_1,X_i)> 0, W(\bar{X}_0,\bar{X}_1,X_j)> 0\}}\right] \\
&= \frac{\hat{\alpha}_0^2}{n_0}\, P(W(\bar{X}_0,\bar{X}_1,X_1)\le 0) + \frac{\hat{\alpha}_1^2}{n_1}\, P(W(\bar{X}_0,\bar{X}_1,X_{n_0+1})> 0) \\
&+\; \hat{\alpha}_0^2\,\frac{n_0-1}{n_0}\, P(W(\bar{X}_0,\bar{X}_1,X_1)\le 0, W(\bar{X}_0,\bar{X}_1,X_2)\le 0) \\
&+\; \hat{\alpha}_1^2\,\frac{n_1-1}{n_1}\, P(W(\bar{X}_0,\bar{X}_1,X_{n_0+1})> 0, W(\bar{X}_0,\bar{X}_1,X_{n_0+2})> 0) \\
&+\; 2\hat{\alpha}_0\hat{\alpha}_1 P(W(\bar{X}_0,\bar{X}_1,X_1)\le 0, W(\bar{X}_0,\bar{X}_1,X_{n_0+1})> 0)
\end{aligned} \tag{3.10}$$

5. Cross-moment of Actual and Estimated Errors

From (1.4) and (1.8), we have that:

$$\begin{aligned}
E[\epsilon\hat{\epsilon}^r] &= E\left[(\alpha_0\epsilon^0 \;+\; \alpha_1\epsilon^1)(\hat{\alpha}_0\hat{\epsilon}_0^r \;+\; \hat{\alpha}_1\hat{\epsilon}_1^r)\right] \\
&= \alpha_0\hat{\alpha}_0 E[\epsilon^0\hat{\epsilon}_0^r] + \alpha_0\hat{\alpha}_1 E[\epsilon^0\hat{\epsilon}_1^r] + \alpha_1\hat{\alpha}_0 E[\epsilon^1\hat{\epsilon}_0^r] + \alpha_1\hat{\alpha}_1 E[\epsilon^1\hat{\epsilon}_1^r]
\end{aligned} \tag{3.11}$$

It follows from (1.9) that

$$
\begin{aligned}
E[\epsilon^0\hat{\epsilon}_0^r] &= E\left[P(W(\bar{X}_0,\bar{X}_1,X) \leq 0 \mid X \in \Pi_0, \bar{X}_0, \bar{X}_1)\frac{1}{n_0}\sum_{i=1}^{n_0} I_{\{W(\bar{X}_0,\bar{X}_1,X_i)\leq 0\}}\right] \\
&= \frac{1}{n_0}\sum_{i=1}^{n_0} E\left[P(W(\bar{X}_0,\bar{X}_1,X) \leq 0 \mid X \in \Pi_0, \bar{X}_0, \bar{X}_1) I_{\{W(\bar{X}_0,\bar{X}_1,X_i)\leq 0\}}\right] \\
&= \frac{1}{n_0}\sum_{i=1}^{n_0} E\left[P(W(\bar{X}_0,\bar{X}_1,X) \leq 0, W(\bar{X}_0,\bar{X}_1,X_i) \leq 0 \mid X \in \Pi_0, X_i, \bar{X}_0, \bar{X}_1))\right] \\
&= P(W(\bar{X}_0,\bar{X}_1,X) \leq 0, W(\bar{X}_0,\bar{X}_1,X_1) \leq 0 \mid X \in \Pi_0)
\end{aligned}
\tag{3.12}
$$

Similar expressions obtain for the other terms in (3.11), namely $E[\epsilon^0\hat{\epsilon}_1^r]$, $E[\epsilon^1\hat{\epsilon}_0^r]$, and $E[\epsilon^1\hat{\epsilon}_1^r]$. In all,

$$
\begin{aligned}
E[\epsilon\hat{\epsilon}^r] &= \alpha_0\hat{\alpha}_0 P(W(\bar{X}_0,\bar{X}_1,X) \leq 0, W(\bar{X}_0,\bar{X}_1,X_1) \leq 0 \mid X \in \Pi_0) \\
&+ \alpha_0\hat{\alpha}_1 P(W(\bar{X}_0,\bar{X}_1,X) \leq 0, W(\bar{X}_0,\bar{X}_1,X_{n_0+1}) > 0 \mid X \in \Pi_0) \\
&+ \alpha_1\hat{\alpha}_0 P(W(\bar{X}_0,\bar{X}_1,X) > 0, W(\bar{X}_0,\bar{X}_1,X_1) \leq 0 \mid X \in \Pi_1) \\
&+ \alpha_1\hat{\alpha}_1 P(W(\bar{X}_0,\bar{X}_1,X) > 0, W(\bar{X}_0,\bar{X}_1,X_{n_0+1}) > 0 \mid X \in \Pi_1)
\end{aligned}
\tag{3.13}
$$

B. Actual Classification Error

Starting from the expressions obtained in the previous section, in this section we derive the exact expressions for the bias, variance, and the RMS of the resubstitution and leave-one-out for LDA in the univariate Gaussian model. The basic method used in these proofs consists in writing out the W statistics in an appropriate matrix form. Notice that all results are derived for general variances σ_0^2 and σ_1^2 (equal variances are not assumed).

The first and second moments of the actual classification error can be written exactly in the univariate Gaussian case according to the following two theorems. We remark that a special case of Theorem 9 is shown in [1], for the equal-variance case $\sigma_0 = \sigma_1$.

Theorem 9. *Let $X_i \sim N(\mu_0, \sigma^2)$ be i.i.d. observations for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \sigma^2)$ be i.i.d. observations for $i = n_0 + 1, \ldots, n_0 + n_1$ used to derive the classifier in (2.3). Then we have:*

$$E[\epsilon] = \alpha_0 \big[P(Z^I < \mathbf{0}) + P(Z^I \geq \mathbf{0}) \big] + \alpha_1 \big[P(Z^{II} < \mathbf{0}) + P(Z^{II} \geq \mathbf{0}) \big] \tag{3.14}$$

where Z^I and Z^{II} are Gaussian bivariate vectors with means and covariance matrices as follows:

$$\begin{aligned} \mu_{Z^I} &= \begin{bmatrix} \frac{\mu}{2} \\ -\mu \end{bmatrix}, \quad \Sigma_{Z^I} = \begin{pmatrix} (1 + \frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\ \mu_{Z^{II}} &= \begin{bmatrix} \frac{-\mu}{2} \\ \mu \end{bmatrix}, \quad \Sigma_{Z^{II}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \end{aligned} \tag{3.15}$$

where $\mu = \mu_0 - \mu_1$.

Proof. See Appendix. □

Theorem 10. *Let $X_i \sim N(\mu_0, \sigma^2)$ be i.i.d. observations for $i = 1, \ldots, n_0$, and $X_i \sim N(\mu_1, \sigma^2)$ be i.i.d. observations for $i = n_0 + 1, \ldots, n_0 + n_1$ used to derive the classifier in (2.3). Then we have:*

$$\begin{aligned} E[\epsilon^2] &= \alpha_0 \alpha_0 \big[P(Z_0^I < \mathbf{0}) + P(Z_0^I \geq \mathbf{0}) + P(Z_1^I < \mathbf{0}) + P(Z_1^I \geq \mathbf{0}) \big] \\ &+ 2\alpha_0 \alpha_1 \big[P(Z_0^{II} < \mathbf{0}) + P(Z_0^{II} \geq \mathbf{0}) + P(Z_1^{II} < \mathbf{0}) + P(Z_1^{II} \geq \mathbf{0}) \big] \\ &+ \alpha_1 \alpha_1 \big[P(Z_0^{III} < \mathbf{0}) + P(Z_0^{III} \geq \mathbf{0}) + P(Z_1^{III} < \mathbf{0}) + P(Z_1^{III} \geq \mathbf{0}) \big] \end{aligned} \tag{3.16}$$

where Z_i^j, for $i = 0, 1$ and $j = I, \ldots, III$, are 4-variate Gaussian random vectors with *means and covariance matrices as follows:*

$$
\begin{aligned}
\mu_{Z_0^I} &= \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ \frac{\mu}{2} \\ -\mu \end{bmatrix}, \Sigma_{Z_0^I} = \begin{pmatrix} (1+\frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & (1+\frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\
\mu_{Z_1^I} &= \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ -\frac{\mu}{2} \\ \mu \end{bmatrix}, \Sigma_{Z_1^I} = \begin{pmatrix} (1+\frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ . & . & (1+\frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\
\mu_{Z_0^{II}} &= \mu_{Z_0^I}, \quad \Sigma_{Z_0^{II}} = \begin{pmatrix} (1+\frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\
\mu_{Z_1^{II}} &= \mu_{Z_1^I}, \quad \Sigma_{Z_1^{II}} = \begin{pmatrix} (1+\frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix}
\end{aligned}
\tag{3.17}
$$

where $\mu = \mu_0 - \mu_1$, and $\mu_{Z_i^{III}}$ and $\Sigma_{Z_i^{III}}$ are obtained from $\mu_{Z_i^I}$ and $\Sigma_{Z_i^I}$, respectively, by exchanging n_0 and n_1, and σ_0 and σ_1, for $i = 0, 1$.

Proof. See Appendix. □

C. Resubstitution Error Estimator

The first and second moments of the resubstitution error estimator and its cross-moment with the actual classification error can be written exactly in the univariate Gaussian case according to the following three theorems, respectively.

Theorem 11. *Let $X_i \sim N(\mu_0, \sigma^2)$ be i.i.d. observations for $i = 1, \dots, n_0$, and $X_i \sim N(\mu_1, \sigma^2)$ be i.i.d. observations for $i = n_0 + 1, \dots, n_0 + n_1$ used to derive the classifier in (2.3). Then we have:*

$$E[\epsilon_r] = \hat{\alpha}_0 \left[P(Z^I < \mathbf{0}) + P(Z^I \geq \mathbf{0})\right] + \hat{\alpha}_1 \left[P(Z^{II} < \mathbf{0}) + P(Z^{II} \geq \mathbf{0})\right] \tag{3.18}$$

where Z^I and Z^{II} are Gaussian bivariate vectors with means and covariance matrices as follows:

$$\begin{aligned} \mu_{Z^I} &= \begin{bmatrix} \frac{\mu}{2} \\ -\mu \end{bmatrix}, \quad \Sigma_{Z^I} = \begin{pmatrix} (1 - \frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\ \mu_{Z^{II}} &= \begin{bmatrix} \frac{-\mu}{2} \\ \mu \end{bmatrix}, \quad \Sigma_{Z^{II}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1 - \frac{3}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \end{aligned} \tag{3.19}$$

where $\mu = \mu_0 - \mu_1$.

Proof. Similar to Theorem 9. □

Theorem 12. *Let $X_i \sim N(\mu_0, \sigma^2)$ be i.i.d. observations for $i = 1, \dots, n_0$, and $X_i \sim N(\mu_1, \sigma^2)$ be i.i.d. observations for $i = n_0 + 1, \dots, n_0 + n_1$ used to derive the classifier in (2.3). Then we have:*

$$\begin{aligned} E[(\hat{\epsilon}^r)^2] = {} & \frac{\hat{\alpha}_0^2}{n_0} \left[P(Z^I < 0) + P(Z^I \geq \mathbf{0})\right] + \frac{\hat{\alpha}_1^2}{n_1} \left[P(Z^{II} < \mathbf{0}) + P(Z^{II} \geq \mathbf{0})\right] \\ & + \hat{\alpha}_0^2 \frac{n_0 - 1}{n_0} \left[P(Z_0^{III} < \mathbf{0}) + P(Z_0^{III} \geq \mathbf{0}) + P(Z_1^{III} < \mathbf{0}) + P(Z_1^{III} \geq \mathbf{0})\right] \\ & + \hat{\alpha}_1^2 \frac{n_1 - 1}{n_1} \left[P(Z_0^{IV} < \mathbf{0}) + P(Z_0^{IV} \geq \mathbf{0}) + P(Z_1^{IV} < \mathbf{0}) + P(Z_1^{IV} \geq \mathbf{0})\right] \\ & + 2\hat{\alpha}_0\hat{\alpha}_1 \left[P(Z_0^{V} < \mathbf{0}) + P(Z_0^{V} \geq \mathbf{0}) + P(Z_1^{V} < \mathbf{0}) + P(Z_1^{V} \geq \mathbf{0})\right] \end{aligned} \tag{3.20}$$

where Z^I and Z^{II} are defined in Theorem 11, and Z_i^j, for $i = 0, 1$ and $j = III, IV, V$, are 4-variate Gaussian random vectors with means and covariances matrices as follows:

$$\mu_{Z_0^{III}} = \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ \frac{\mu}{2} \\ -\mu \end{bmatrix}, \ \Sigma_{Z_0^{III}} = \begin{pmatrix} (1-\frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{3\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & (1-\frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix}$$

$$\mu_{Z_1^{III}} = \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ -\frac{\mu}{2} \\ \mu \end{bmatrix}, \ \Sigma_{Z_1^{III}} = \begin{pmatrix} (1-\frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{3\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ . & . & (1-\frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix}$$

$$\mu_{Z_0^{IV}} = \mu_{Z_0^{III}}, \ \ \Sigma_{Z_0^{IV}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1-\frac{3}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{3\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1-\frac{3}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix}$$

$$\mu_{Z_1^{IV}} = \mu_{Z_1^{III}}, \ \ \Sigma_{Z_1^{IV}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1-\frac{3}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} + \frac{3\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1-\frac{3}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix}$$

$$\mu_{Z_0^{V}} = \mu_{Z_0^{III}}, \ \ \Sigma_{Z_0^{V}} = \begin{pmatrix} (1-\frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1-\frac{3}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix}$$

$$\mu_{Z_1^V} = \mu_{Z_1^{III}}, \; \Sigma_{Z_1^V} = \begin{pmatrix} (1-\frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1-\frac{3}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \tag{3.21}$$

where $\mu = \mu_0 - \mu_1$.

Proof. Similar to Theorem 10. □

Theorem 13. *Let* $X_i \sim N(\mu_0, \sigma^2)$ *be i.i.d. observations for* $i = 1, \ldots, n_0$, *and* $X_i \sim N(\mu_1, \sigma^2)$ *be i.i.d. observations for* $i = n_0 + 1, \ldots, n_0 + n_1$ *used to derive the classifier in (2.3). Then we have:*

$$\begin{aligned} E[\epsilon\hat{\epsilon}^r] = \; & \alpha_0\hat{\alpha}_0 \left[P(Z_0^I < \mathbf{0}) + P(Z_0^I \geq \mathbf{0}) + P(Z_1^I < \mathbf{0}) + P(Z_1^I \geq \mathbf{0})\right] \\ + \; & \alpha_0\hat{\alpha}_1 \left[P(Z_0^{II} < \mathbf{0}) + P(Z_0^{II} \geq \mathbf{0}) + P(Z_1^{II} < \mathbf{0}) + P(Z_1^{II} \geq \mathbf{0})\right] \\ + \; & \alpha_1\hat{\alpha}_0 \left[P(Z_0^{III} < \mathbf{0}) + P(Z_0^{III} \geq \mathbf{0}) + P(Z_1^{III} < \mathbf{0}) + P(Z_1^{III} \geq \mathbf{0})\right] \\ + \; & \alpha_1\hat{\alpha}_1 \left[P(Z_0^{IV} < \mathbf{0}) + P(Z_0^{IV} \geq \mathbf{0}) + P(Z_1^{IV} < \mathbf{0}) + P(Z_1^{IV} \geq \mathbf{0})\right] \end{aligned} \tag{3.22}$$

where Z_i^j, *for* $i = 0, 1$ *and* $j = I, \ldots, IV$, *are 4-variate Gaussian random vectors with means and covariances as follows:*

$$\mu_{Z_0^I} = \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ \frac{\mu}{2} \\ -\mu \end{bmatrix}, \; \Sigma_{Z_0^I} = \begin{pmatrix} (1+\frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & (1-\frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix}$$

$$\mu_{Z_1^I} = \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ -\frac{\mu}{2} \\ \mu \end{bmatrix}, \; \Sigma_{Z_1^I} = \begin{pmatrix} (1+\frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ . & . & (1-\frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix}$$

$$\begin{aligned}
\mu_{Z_0^{II}} &= \mu_{Z_0^{I}}\,, \quad \Sigma_{Z_0^{II}} = \begin{pmatrix} (1+\frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1-\frac{3}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\
\mu_{Z_1^{II}} &= \mu_{Z_1^{I}}\,, \quad \Sigma_{Z_1^{II}} = \begin{pmatrix} (1+\frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1-\frac{3}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\
\mu_{Z_0^{III}} &= \begin{bmatrix} \frac{-\mu}{2} \\ \mu \\ \frac{\mu}{2} \\ -\mu \end{bmatrix}, \ \Sigma_{Z_0^{III}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ . & . & (1-\frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\
\mu_{Z_1^{III}} &= \begin{bmatrix} \frac{-\mu}{2} \\ \mu \\ -\frac{\mu}{2} \\ \mu \end{bmatrix}, \ \Sigma_{Z_1^{III}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & (1-\frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\
\mu_{Z_0^{IV}} &= \mu_{Z_0^{III}}\,, \ \Sigma_{Z_0^{IV}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1-\frac{3}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\
\mu_{Z_1^{IV}} &= \mu_{Z_1^{III}}\,, \ \Sigma_{Z_1^{IV}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1-\frac{3}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix}
\end{aligned} \tag{3.23}$$

where $\mu = \mu_0 - \mu_1$.

Proof. See Appendix. □

D. Leave-one-out Error Estimator

By virtue of the relation $E[\hat{\epsilon}_i^l] = E[\epsilon_{i,n_i-1}]$, for $i = 0, 1$, the first moment of the leave-one-out error estimator can be obtained by using Theorem 9, while replacing α_i by $\hat{\alpha}_i$ and n_i by $n_i - 1$, for $i = 0, 1$. As for the second moment of the leave-one-out error estimator and its cross-moment with the actual classification error, they can be written exactly in the univariate Gaussian case according to the following two theorems, respectively.

Theorem 14. *Let* $X_i \sim N(\mu_0, \sigma^2)$ *be i.i.d. observations for* $i = 1, \ldots, n_0$*, and* $X_i \sim N(\mu_1, \sigma^2)$ *be i.i.d. observations for* $i = n_0 + 1, \ldots, n_0 + n_1$ *used to derive the classifier in (2.3). Then we have:*

$$
\begin{aligned}
E[(\hat{\epsilon}^l)^2] = {} & \frac{\hat{\alpha}_0^2}{n_0}\left[P(Z^I < \mathbf{0}) + P(Z^I \geq \mathbf{0})\right] + \frac{\hat{\alpha}_1^2}{n_1}\left[P(Z^{II} < \mathbf{0}) + P(Z^{II} \geq \mathbf{0})\right] \\
& + \hat{\alpha}_0^2 \frac{n_0 - 1}{n_0}\left[P(Z_0^{III} < \mathbf{0}) + P(Z_0^{III} \geq \mathbf{0}) + P(Z_1^{III} < \mathbf{0}) + P(Z_1^{III} \geq \mathbf{0})\right] \\
& + \hat{\alpha}_1^2 \frac{n_1 - 1}{n_1}\left[P(Z_0^{IV} < \mathbf{0}) + P(Z_0^{IV} \geq \mathbf{0}) + P(Z_1^{IV} < \mathbf{0}) + P(Z_1^{IV} \geq \mathbf{0})\right] \\
& + 2\hat{\alpha}_0\hat{\alpha}_1\left[P(Z_0^V < \mathbf{0}) + P(Z_0^V \geq \mathbf{0}) + P(Z_1^V < \mathbf{0}) + P(Z_1^V \geq \mathbf{0})\right]
\end{aligned}
\tag{3.24}
$$

where Z^I *and* Z^{II} *are defined in Theorem 9, but with* n_i *replaced by* $n_i - 1$*, for* $i = 0, 1$*, and* Z_i^j*, for* $i = 0, 1$ *and* $j = III, IV, V$*, are 4-variate Gaussian random vectors with means and covariance matrices as follows:*

$$\mu_{Z_0^{III}} = \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ \frac{\mu}{2} \\ -\mu \end{bmatrix}, \mu_{Z_1^{III}} = \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ -\frac{\mu}{2} \\ \mu \end{bmatrix}, \quad \begin{array}{ll} \mu_{Z_0^{IV}} = \mu_{Z_0^{III}}, & \mu_{Z_1^{IV}} = \mu_{Z_1^{III}} \\ \mu_{Z_0^{V}} = \mu_{Z_0^{III}}, & \mu_{Z_1^{V}} = \mu_{Z_1^{III}} \end{array}$$

$$\Sigma_{Z_0^{III}} = \begin{pmatrix} (1+\frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} & \frac{(-3n_0+2)\sigma_0^2}{4(n_0-1)^2} + \frac{\sigma_1^2}{4n_1} & -\frac{n_0\sigma_0^2}{2(n_0-1)^2} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} & -\frac{n_0\sigma_0^2}{2(n_0-1)^2} - \frac{\sigma_1^2}{2n_1} & \frac{(n_0-2)\sigma_0^2}{(n_0-1)^2} + \frac{\sigma_1^2}{n_1} \\ . & . & (1+\frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} \end{pmatrix}$$

$$\Sigma_{Z_1^{III}} = \begin{pmatrix} (1+\frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} & -\frac{(-3n_0+2)\sigma_0^2}{4(n_0-1)^2} - \frac{\sigma_1^2}{4n_1} & \frac{n_0\sigma_0^2}{2(n_0-1)^2} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} & \frac{n_0\sigma_0^2}{2(n_0-1)^2} + \frac{\sigma_1^2}{2n_1} & -\frac{(n_0-2)\sigma_0^2}{(n_0-1)^2} - \frac{\sigma_1^2}{n_1} \\ . & . & (1+\frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} \end{pmatrix}$$

$$\Sigma_{Z_0^{IV}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} & \frac{\sigma_0^2}{4n_0} + \frac{(-3n_1+2)\sigma_1^2}{4(n_1-1)^2} & -\frac{\sigma_0^2}{2n_0} - \frac{n_1\sigma_1^2}{2(n_1-1)^2} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} & -\frac{\sigma_0^2}{2n_0} - \frac{n_1\sigma_1^2}{2(n_1-1)^2} & \frac{\sigma_0^2}{n_0} + \frac{(n_1-2)\sigma_1^2}{(n_1-1)^2} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix}$$

$$\Sigma_{Z_1^{IV}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} & -\frac{\sigma_0^2}{4n_0} - \frac{(-3n_1+2)\sigma_1^2}{4(n_1-1)^2} & \frac{\sigma_0^2}{2n_0} + \frac{n_1\sigma_1^2}{2(n_1-1)^2} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} & \frac{\sigma_0^2}{2n_0} + \frac{n_1\sigma_1^2}{2(n_1-1)^2} & -\frac{\sigma_0^2}{n_0} - \frac{(n_1-2)\sigma_1^2}{(n_1-1)^2} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix}$$

$$\Sigma_{Z_0^{V}} = \begin{pmatrix} (1+\frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix}$$

$$\Sigma_{Z_1^V} = \begin{pmatrix} (1+\frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix} \tag{3.25}$$

where $\mu = \mu_0 - \mu_1$. $\diamond$

Proof. Similar to Theorem 10. □

Theorem 15. *Let* $X_i \sim N(\mu_0, \sigma^2)$ *be i.i.d. observations for* $i = 1, \dots, n_0$, *and* $X_i \sim N(\mu_1, \sigma^2)$ *be i.i.d. observations for* $i = n_0 + 1, \dots, n_0 + n_1$ *used to derive the classifier in (2.3). Then we have:*

$$\begin{aligned} E[\epsilon\hat{\epsilon}^l] = \; & \alpha_0\hat{\alpha}_0 \left[P(Z_0^I < \mathbf{0}) + P(Z_0^I \geq \mathbf{0}) + P(Z_1^I < \mathbf{0}) + P(Z_1^I \geq \mathbf{0})\right] \\ + \; & \alpha_0\hat{\alpha}_1 \left[P(Z_0^{II} < \mathbf{0}) + P(Z_0^{II} \geq \mathbf{0}) + P(Z_1^{II} < \mathbf{0}) + P(Z_1^{II} \geq \mathbf{0})\right] \\ + \; & \alpha_1\hat{\alpha}_0 \left[P(Z_0^{III} < \mathbf{0}) + P(Z_0^{III} \geq \mathbf{0}) + P(Z_1^{III} < \mathbf{0}) + P(Z_1^{III} \geq \mathbf{0})\right] \\ + \; & \alpha_1\hat{\alpha}_1 \left[P(Z_0^{IV} < \mathbf{0}) + P(Z_0^{IV} \geq \mathbf{0}) + P(Z_1^{IV} < \mathbf{0}) + P(Z_1^{IV} \geq \mathbf{0})\right] \end{aligned} \tag{3.26}$$

where Z_i^j, *for* $i = 0, 1$ *and* $j = I, \dots, IV$, *are 4-variate Gaussian random vectors with means and covariance matrices as follows:*

$$\mu_{Z_0^I} = \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ \frac{\mu}{2} \\ -\mu \end{bmatrix}, \mu_{Z_1^I} = \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ -\frac{\mu}{2} \\ \mu \end{bmatrix}, \mu_{Z_0^{III}} = \begin{bmatrix} \frac{-\mu}{2} \\ \mu \\ \frac{\mu}{2} \\ -\mu \end{bmatrix}, \mu_{Z_1^{III}} = \begin{bmatrix} \frac{-\mu}{2} \\ \mu \\ -\frac{\mu}{2} \\ \mu \end{bmatrix}, \quad \begin{matrix} \mu_{Z_0^{II}} = \mu_{Z_0^I}, & \mu_{Z_1^{II}} = \mu_{Z_1^I} \\ \mu_{Z_0^{IV}} = \mu_{Z_0^{III}}, & \mu_{Z_1^{IV}} = \mu_{Z_1^{III}} \end{matrix}$$

$$\Sigma_{Z_0^I} = \begin{pmatrix} (1+\frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & (1+\frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{(n_0-1)} + \frac{\sigma_1^2}{n_1} \end{pmatrix}$$

$$\Sigma_{Z_1^I} = \begin{pmatrix} (1+\frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ . & . & (1+\frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{(n_0-1)} + \frac{\sigma_1^2}{n_1} \end{pmatrix}$$

$$\Sigma_{Z_0^{II}} = \begin{pmatrix} (1+\frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_0 1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix}$$

$$\Sigma_{Z_1^{II}} = \begin{pmatrix} (1+\frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix}$$

$$\Sigma_{Z_0^{III}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ . & . & (1+\frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} \end{pmatrix}$$

$$\Sigma_{Z_1^{III}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & (1+\frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} \\ . & . & . & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} \end{pmatrix}$$

$$\Sigma_{Z_0^{IV}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1+\frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix}$$

$$\Sigma_{Z_1^{IV}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ . & . & \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ . & . & . & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix} \tag{3.27}$$

where $\mu = \mu_0 - \mu_1$.

Proof. Similar to Theorem 13. □

Figure 11 provides graphs of the basic performance measures for resubstitution and leave-one-out as a function of sample size in the balanced case, $n_0 = n_1 = n$. To generate the results, two Gaussian with different means $\mu_1 = -\mu_0 = 1$ and unequal variances $\sigma_0^2 = 1$, $\sigma_1^2 = 4$ have been employed. The optimal linear classifier error in this example is 0.2335. The different parts of the figure show bias, devaiation variance, correlation coefficient, and RMS.

E. RMS Bounds

When one designs a classifier and reports an error estimate, there is no way to tell how accurate the estimate is because we do not know the true error of the classifier. Knowledge of estimation accuracy rests with the accuracy of the error estimation rule, which is most commonly judged by the RMS. When reporting an estimate, it would be beneficial to state some bound on the RMS. In addition, as in any experimental situation, it would be useful to determine ahead of time the the minimum sample size necessary to obtain a desired degree of estimation accuracy. In this vein, some recommendations on sample size requirements have been provided in the literature [49, 50]. In particular, if one has a bound on the RMS in terms of sample size, then the required sample size for a desired RMS can be obtained. There exist some

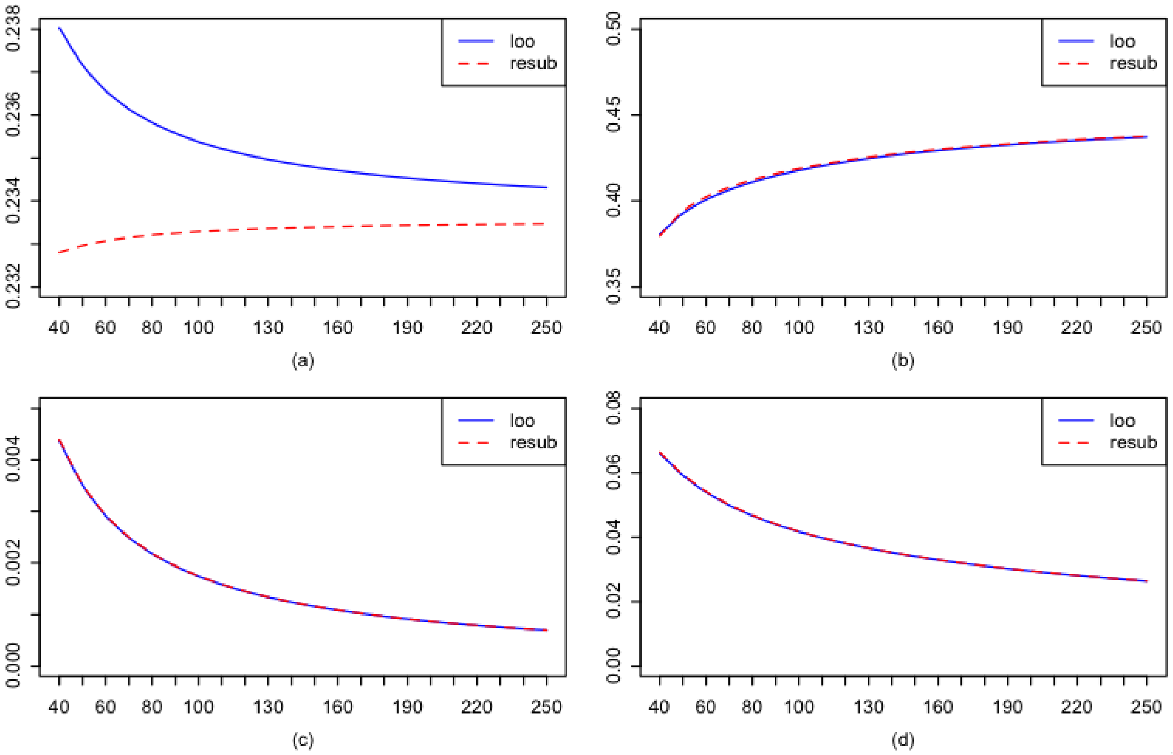

Fig. 11. Performance measures for resubstitution and leave-one-out as a function of sample of sample size for LDA in the univariate model: (a) mean errors, (b) correlation coefficient with actual error, (c) deviation variance, (d) RMS.

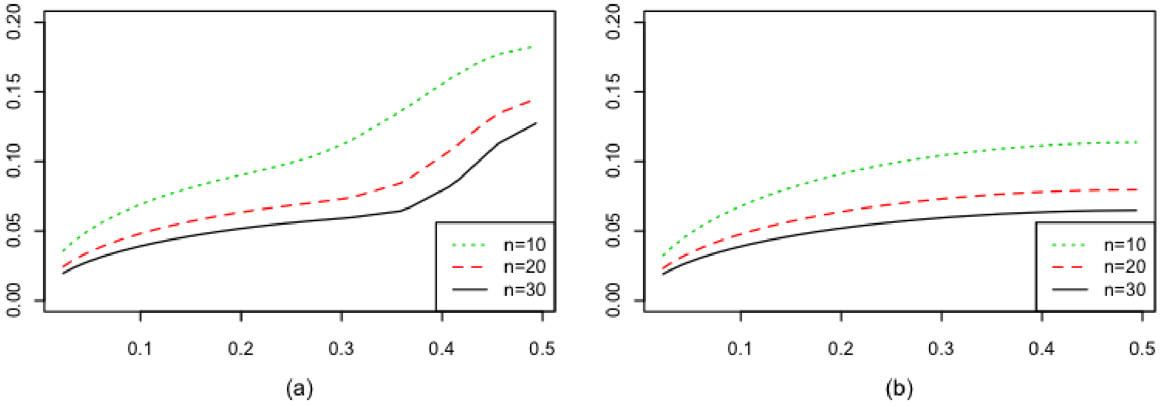

Fig. 12. RMS versus Bayes error in a Gaussian model for (a) leave-one-out, (b) resubstitution.

distribution-free bounds for some classification rules [24, 119, 120]. These bounds tend to be very loose and therefore of limited practical value. For instance, for leave-one-out and the k-nearest-neighbor (kNN) classification rule with random tie-breaking, there exists the following distribution-free bound [24]:

$$\mathrm{RMS}[\hat{\epsilon}^l] \leq \sqrt{\frac{1}{n_0 + n_1}\left(1 + 24\sqrt{\frac{k}{2\pi}}\right)} \tag{3.28}$$

If $k = 3$ and $n_0 + n_1 = 100$, then the bound is approximately 0.419, which means knowledge of the true error is highly uncertain. The problem here is not mainly kNN; rather, it is the distribution-free nature of the bound. Another example is the following resubstitution bound for the histogram rule [24]:

$$\mathrm{RMS}[\hat{\epsilon}^r] \leq \sqrt{\frac{6k}{n_0 + n_1}} \tag{3.29}$$

where k is the maximum number of fixed partitions of the feature space. Taking $k = 10$ and $k = 20$ with $n_0 + n_1 = 100$, then the bounds are 0.77 and 1.09, respectively, both being of no practical value.

Now consider leave-one-out, resubstitution and LDA in the model class we have been considering. Consider two equal univariate Gaussian distributions with means $\mu_1 = -\mu_0 = 1$ and $\sigma_0 = \sigma_1 = 1$. Using the RMS expressions obtained before, the RMS versus Bayes error curves are shown in Fig. 12 for different sample sizes and balanced design, $n_0 = n_1 = n$. Letting ϵ_{bay} denote the Bayes error, we see that RMS is an increasing function of ϵ_{bay}. Letting $\kappa_{\hat{\epsilon}}(n, \tau) = \max_{\epsilon_{bay} \leq \tau} \mathrm{RMS}[\hat{\epsilon}]$ for $n_0 = n_1 = n$ and $\hat{\epsilon} = \hat{\epsilon}^r, \hat{\epsilon}^l$, we have the bounds $\mathrm{RMS}[\hat{\epsilon}^l] \leq \kappa_{\hat{\epsilon}^l}(20, 0.5) = 0.145$ and $\mathrm{RMS}[\hat{\epsilon}^r] \leq \kappa_{\hat{\epsilon}^r}(20, 0.5) = 0.080$ for $n = 20$, and $\mathrm{RMS}[\hat{\epsilon}^l] \leq \kappa_{\hat{\epsilon}^l}(30, 0.5) = 0.127$ and $\mathrm{RMS}[\hat{\epsilon}^r] \leq \kappa_{\hat{\epsilon}^r}(30, 0.5) = 0.065$ for $n = 30$. These are far tighter than the distribution-free bounds in (3.28); indeed, no distribution-free bound is known for LDA.

From a practical perspective, given a desired RMS, the required sample size can be determined. If one desires that the RMS be bounded by η, then one need only find the minimum value of n so that $\kappa_{\hat{\epsilon}}(n, 0.5) \leq \eta$ where $\hat{\epsilon} = \hat{\epsilon}^r, \hat{\epsilon}^l$. Table I shows the required sample size calculated using this scheme for a balanced design ($n_0 = n_1 = n$). Note that the required sample size in Table I does not depend on the actual value of the common variance, a peculiar result of the equal-variance model class being considered. In the univariate case, the number of samples needed to achieve a given $\kappa_{\hat{\epsilon}^l}(n, 0.5)$ is much higher than $\kappa_{\hat{\epsilon}^r}(n, 0.5)$, which is evident in Fig. 12, owing to the abrupt increase of RMS$[\hat{\epsilon}^l]$ for large ϵ_{bay}. While RMS$[\hat{\epsilon}^l] \approx$ RMS$[\hat{\epsilon}^r]$ when $\epsilon_{bay} \leq 0.35$, since we do not know the true error, the bound for RMS$[\hat{\epsilon}^l]$ must take into account the possibility $\epsilon_{bay} > 0.35$. It is instructive to compare the sample sizes determined from Table I with those determined from (3.28) and (3.29) to achieve a given RMS, say 0.1. From (3.28), in the case of kNN with $k = 3$ and leave-one-out, we need $n_0 = n_1 = 875$, whereas $n_0 = n_1 = 67$ from Table I in the univariate LDA case. From (3.29) for resubstitution and in case of the histogram rule with $k = 10$ and $k = 20$, we need 3000 and 6000 sample points in each class, respectively, whereas from Table I we need 13 sample points in each class for univariate LDA and resubstitution.

Table I. Minimum sample size, n, ($n_0 = n_1 = n$) for desired $\kappa(n, 0.5)$ in univariate case.

$\kappa(n, 0.5)$	resub	loo
0.050	51	793
0.060	36	403
0.070	26	230
0.080	20	143
0.090	16	95
0.100	13	67

1. Implementation for Gene-expression Classification

In this section, we demonstrate the practical use of RMS bounds in the case of classification using gene-expression data from a breast-cancer study that analyzed 295 gene-expression microarrays containing a total of 25760 transcripts on each [118]. Discrimination is between good versus bad prognosis. Here we design of a classifier based on a single gene. Using resubstitution, from Table I, we need 20 sample points for each class to have $\kappa_{\hat{\epsilon}^r}(n, 0.5) = 0.08$. This bound does not apply to leave-one-out; indeed, $\kappa_{\hat{\epsilon}^l}(20, 0.5) > 0.13$. However, as explained previously, if it happens that $\epsilon_{bay} < 0.35$ then $\text{RMS}[\hat{\epsilon}^l] \approx \text{RMS}[\hat{\epsilon}^r]$, so that $\kappa_{\hat{\epsilon}^l}(20, 0.35) \approx \kappa_{\hat{\epsilon}^r}(20, 0.35) < \kappa_{\hat{\epsilon}^r}(20, 0.5) = 0.08$ also. This example will elucidate this situation because we will have an accurate estimate of the true error. We consider the total of 295 gene-expression profiles for 70 genes from the 295 microarrays as the population and draw a random sample of size 40 with $n_0 = n_1 = 20$. Using the 40 sample points selected, we applied the t-test to find the differentially expressed genes among the 70 genes. Results of the t-test on the sample showed 35 genes to be differentially expressed among the 70 genes. Then the Shapiro-Wilk test (using the R statistical software) was applied on these 35 genes to test the normality of each gene at significance level 0.95. Note that to do so, only the 40 points taken randomly from the whole population were considered, so as to reflect the situation that no additional data are available in practice. The test did not reject the Gaussianity assumption of 26 genes out of the 35 genes previously selected by t-test. Then F-test for equality of variances of both classes was performed on these 26 selected genes to test the equality of variances of each gene across the classes. The result of F-test reduced the number of genes to 13. In sum, these 13 genes are those that show significant different expressions between two classes (by t-test), are close to normal (by Shapiro-Wilk test), and have close to equal variances in the two classes (by F-test). Since we take into account the validity of the classifier, through RMS, as well as its goodness, through estimated error, we call this whole procedure of selecting the genes validity-goodness feature selection. The genes selected using this scheme are shown in Table II. The last column of this table shows the hold out estimate using 190 hold-out points selected from the 255 remaining sample points to reflect the equal prior probability of the classes, as was done for training. With 190 hold-out points, one can expect the hold-out estimate to be very accurate. Comparing the values of hold-out in these examples with those of the estimators themselves, we conclude that both resubstitution and leave-one-out have reasonably estimated the true error. We would certainly have expected this owing to the RMS bound on resubstitution and, as we see the true errors are less than 0.35, so that the Bayes errors must also be less than 0.35, in hindsight we expect this from leave-one-out. In practice, of course, we do not have a population based evaluation of the true error, so that a conservative approach requires taking $\kappa_{\hat{\epsilon}^l}(n, 0.5)$ as the bound.

Table II. Genes selected using the validity-goodness model selection criterion.

genes	resubs error	loo error	hold-out
Contig46218_RC	0.225	0.225	0.260
NM_016359	0.200	0.200	0.211
Contig28552_RC	0.300	0.300	0.250
Contig32125_RC	0.350	0.375	0.358
AB037863	0.275	0.275	0.331
NM_020974	0.275	0.275	0.255
Contig55377_RC	0.225	0.225	0.233
Contig25991	0.325	0.325	0.315
NM_006101	0.325	0.325	0.282
NM_003239	0.325	0.325	0.293
NM_01644	0.325	0.325	0.298
NM_001809	0.225	0.250	0.173
NM_004702	0.225	0.225	0.239

F. Conclusion

Because the error of a classifier characterizes its predictive capacity, which represents the scientific content of the classifier, the salient epistemological problem in pattern recognition is error-estimator performance. When one has access to large samples, the issue is not so severe because the data can be split into training and test data, and moreover, training-data error estimators tend to have good large-sample performance, as demonstrated in this chapter for resubstitution and leave-one-out. Current high-throughput technologies often produce high-dimensional data with a small number of replicates. Hence, the efficacy of classifiers derived from such data sets requires direct performance analysis. In this chapter we have provided analytic representation for the main performance criteria: bias, variance, and RMS for resubstitution and leave-one-out for LDA in a univariate Gaussian model. More such studies will be necessary if we are to gain critical understanding of classifier performance in the context of small samples. The second part of the study will address the corresponding multivariate model.

CHAPTER IV

ANALYTIC STUDY OF PERFORMANCE OF ERROR ESTIMATORS FOR LINEAR DISCRIMINANT ANALYSIS THROUGH RMS – MULTIVARIATE MODEL

In this chapter, we derive double asymptotic (in sample size and dimension) analytical expressions for the first moments, second moments, and cross-moments with the actual error for the resubstitution and leave-one-out error estimators in the case of linear discriminant analysis (LDA) in the multivariate Gaussian model under the assumption of a common known covariance matrix. Sample sizes for the two classes need not be the same. Such asymptotic results generally provide good small sample approximations and this is demonstrated in the present situation via numerical comparisons. From the asymptotic moment representations, we directly obtain double asymptotic expressions for the bias, variance, and RMS of the error estimators.

A. Double Asymptotic Approximation

1. Previous Work

In [121], Raudys proposed an approximation to the expected actual classification error:

$$E[\epsilon_0] = P(W(\bar{X}_0, \bar{X}_1, X) \leq 0 \mid X \in \Pi_0) \approx \Phi\left(-\frac{E[W(\bar{X}_0, \bar{X}_1, X) \mid X \in \Pi_0]}{\sqrt{\mathrm{Var}(W(\bar{X}_0, \bar{X}_1, X) \mid X \in \Pi_0)}}\right) \tag{4.1}$$

in which $\Phi(.)$ is the standard normal cumulative function. To obtain the corresponding approximation to $E[\epsilon_1]$, it suffices to modify the argument of Φ by replacing Π_0 by Π_1 and multiplying by -1. In the case $n_0 = n_1 = n$, then $E[\epsilon] = E[\epsilon_0] = E[\epsilon_1]$.

Using (4.1) in this case, Raudys obtained in [33] the approximation:

$$E[\epsilon] \approx \Phi\left(-\frac{\delta}{2}\frac{1}{\sqrt{1+\frac{1}{n}+\frac{2p}{n\delta^2}+\frac{p}{n^2\delta^2}}}\right) \tag{4.2}$$

where $\delta^2 = (\mu_0 - \mu_1)^T\Sigma^{-1}(\mu_0 - \mu_1)$. In [116], Raudys pointed out, without exhibiting an explicit proof, that this approximation is asymptotically exact under the double asymptotic condition $n \to \infty$, $p \to \infty$, $n/p \to$ constant. Under these conditions, the following asymptotically-equivalent approximation results:

$$E[\epsilon] \approx \Phi\left(-\frac{\delta}{2}\frac{1}{\sqrt{1+\frac{2p}{n\delta^2}}}\right) \tag{4.3}$$

To obtain the approximation for the expectation of the resubstitution error, (4.1) is modified by replacing X by X_1:

$$E[\hat{\epsilon}_0^r] = P(W(\bar{X}_0,\bar{X}_1,X_1) \le 0) \approx \Phi\left(-\frac{E[W(\bar{X}_0,\bar{X}_1,X_1)]}{\sqrt{\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X_1))}}\right) \tag{4.4}$$

To obtain the corresponding approximation to $E[\hat{\epsilon}_1^r]$, it suffices to modify the argument of Φ by replacing X_1 by X_{n_0+1} and multiplying by -1. In the case $n_0 = n_1 = n$, then $E[\hat{\epsilon}^r] = E[\hat{\epsilon}_0^r] = E[\hat{\epsilon}_1^r]$, and (4.4) leads to the following approximation:

$$E[\hat{\epsilon}^r] \approx \Phi\left(-\frac{\delta}{2}\frac{1+\frac{2p}{n\delta^2}}{\sqrt{1+\frac{1}{n}+\frac{2p}{n\delta^2}}}\right) \tag{4.5}$$

This expression is equivalent to the one published by Raudys in [115, 116], under the double asymptotic condition $n \to \infty$, $p \to \infty$, $n/p \to$ constant, namely:

$$E[\hat{\epsilon}^r] \approx \Phi\left(-\frac{\delta}{2}\sqrt{1+\frac{2p}{n\delta^2}}\right) \tag{4.6}$$

We will prove in the following subsections that all the approximations discussed above are asymptotically exact, as $n_0 \to \infty$, $n_1 \to \infty$, $p \to \infty$, $p/n_0 \to \lambda_0$, $p/n_1 \to \lambda_1$ —

which Serdobolskii calls “Kolmogorov asymptotic conditions” in [122].

Wyman and his colleagues [123] used Monte-Carlo simulations to compare different expressions for expectation of true error. The expressions they considered for this comparison were those proposed by Raudys [121], Efron [47], Anderson [46], Okamoto [43], Sayre [48], Deev [124], and [95]. They concluded that: “*A simple and relatively obscure asymptotic expansion derived by Raudys (Tech. Cybern. 4, 168-174, 1972) is found to yield better approximation than the well-known asymptotic expansions*”.

With all ambiguity on the origin of Kolmogorov asymptotic analysis, this approach has been vigorously followed in Soviet-Union [50, 115, 116, 121, 122, 124–128]. The finite-sample approximations obtained via these asymptotic expressions have been shown to be remarkably accurate in small-sample cases [123, 129]. More recently, this kind of asymptotic approach has been used successfully to analyze the performance of popular multiuser detection algorithms such as CDMA [130, 131] . There the assumption is that in a K-user channel with spreading gain N, both K and N go to infinity while their ratio remains constant. In this context, the assumption of increasing dimension of the system has been called a “large-system limit”. One can find its root in the prominent work of Wigner [132]. Recently, Serdobolskii, who was a pioneer on developing the Kolmogorov asymptotic approach in the Soviet Union, has published a book [122] to integrate the main results on this kind of limit that have been developing independently in the eastern and western hemispheres.

In what follows, we will denote convergence in probability under Kolmogorov asymptotic conditions by “ $\underset{n_0,n_1,p\to\infty}{\text{pklim}}$ ”. Similarly, “ $\underset{n_0,n_1,p\to\infty}{\text{klim}}$ ” and “$\xrightarrow{K}$” will denote ordinary convergence under the Kolmogorov asymptotic conditions. For simplifying the notations, the following functions are defined that will be used throughout this chapter:

$$f_0\left(n_0,n_1,p,\delta^2\right)=\sqrt{1+\frac{1}{n_1}+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)+\frac{p}{2\delta^2}\left(\frac{1}{n_0^2}+\frac{1}{n_1^2}\right)}, \quad f_1\left(n_0,n_1,p,\delta^2\right)=f_0\left(n_1,n_0,p,\delta^2\right)$$
$$g_0\left(n_0,n_1,p,\delta^2\right)=\sqrt{1+\frac{1}{n_1}+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)+\frac{p}{2\delta^2}\left(\frac{1}{n_1^2}-\frac{1}{n_0^2}\right)}, \quad g_1\left(n_0,n_1,p,\delta^2\right)=g_0\left(n_1,n_0,p,\delta^2\right) \tag{4.7}$$

2. Actual Classification Error

Let us define a sequence of Gaussian discrimination problems defined by the sequence of parameter and sample sizes:

$$(\mu_{p,0},\mu_{p,1},\Sigma_p,n_{p,0},n_{p,1}), \quad p=1,2,\ldots \tag{4.8}$$

where the means and covariance matrix are arbitrary except that the Mahalanobis distance, $\delta = \sqrt{(\mu_{p,0}-\mu_{p,1})^T\Sigma_p^{-1}(\mu_{p,0}-\mu_{p,1})}$, is assumed to be a constant (with slightly more work, this condition can be relaxed to an arbitrary Mahalanobis distance converging to a constant δ as $p \to \infty$, as in [125]). For simplicity of notation, and at no risk of ambiguity, we will omit in the sequel the subscript "p" from the parameters and sample sizes in (4.8).

The assumption that the covariance matrix Σ is known simplifies the analysis, eliminating the need for many of the regularity conditions required by Serdobolskii in [122]. Let

$$\hat{G}_i=E[W(\bar{X}_0,\bar{X}_1,X)\mid\bar{X}_0,\bar{X}_1,X\in\Pi_i], \quad \hat{D}_i=\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X)\mid\bar{X}_0,\bar{X}_1,X\in\Pi_i) \tag{4.9}$$

for $i=0,1$. Then the population-specific classification errors are given by:

$$\epsilon_0 = \Phi\left(-\frac{\hat{G}_0}{\sqrt{\hat{D}_0}}\right), \quad \epsilon_1 = \Phi\left(\frac{\hat{G}_1}{\sqrt{\hat{D}_1}}\right) \tag{4.10}$$

We have the following result:

Theorem 16. *Consider the sequence of Gaussian discrimination problems defined by (4.8). Then*

$$\underset{n_0,n_1,p\to\infty}{\text{klim}} E[\epsilon_0] = \underset{n_0,n_1,p\to\infty}{\text{pklim}} \epsilon_0 = \Phi\left(\frac{-G_0}{\sqrt{D}}\right), \qquad \underset{n_0,n_1,p\to\infty}{\text{klim}} E[\epsilon_1] = \underset{n_0,n_1,p\to\infty}{\text{pklim}} \epsilon_1 = \Phi\left(\frac{G_1}{\sqrt{D}}\right) \tag{4.11}$$

so that

$$\underset{n_0,n_1,p\to\infty}{\text{klim}} E[\epsilon] = \underset{n_0,n_1,p\to\infty}{\text{pklim}} \epsilon = \alpha_0\Phi\left(-\frac{G_0}{\sqrt{D}}\right) + \alpha_1\Phi\left(\frac{G_1}{\sqrt{D}}\right) \tag{4.12}$$

where

$$\begin{aligned} G_0 &= \underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{G}_0] = \frac{1}{2}(\delta^2 + \lambda_1 - \lambda_0), \quad G_1 = \underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{G}_1] = -\frac{1}{2}(\delta^2 + \lambda_0 - \lambda_1) \\ D &= \underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{D}_0] = \underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{D}_1] = \delta^2 + \lambda_0 + \lambda_1 \end{aligned} \tag{4.13}$$

Proof. See Appendix. □

We remark that (4.12) is equivalent to the specialization of Deev's formula [116] to the case where the covariance matrix is known.

Theorem 16 suggests the following finite-sample approximation:

$$E[\epsilon_0] \approx \Phi\left(-\frac{E[\hat{G}_0]}{\sqrt{E[\hat{D}_0]}}\right) = \Phi\left(-\frac{E[W(\bar{X}_0,\bar{X}_1,X) \mid X \in \Pi_0]}{\sqrt{E[\text{Var}(W(\bar{X}_0,\bar{X}_1,X) \mid \bar{X}_0,\bar{X}_1,X \in \Pi_0)]}}\right) \tag{4.14}$$

To obtain the corresponding approximation to $E[\epsilon_1]$, it suffices to replace $\hat{G}_0$ by $\hat{G}_1$, $\hat{D}_0$ by $\hat{D}_1$, and Π_0 by Π_1, and multiply the argument of both Φ functions by -1. Evaluating the expectation in the numerator and denominator of (4.14) yields

$$E[\epsilon_0] \approx \Phi\left(-\frac{\delta}{2}\,\frac{1 + \frac{p}{\delta^2}\left(\frac{1}{n_1} - \frac{1}{n_0}\right)}{\sqrt{1 + \frac{p}{\delta^2}\left(\frac{1}{n_1} + \frac{1}{n_0}\right)}}\right) \tag{4.15}$$

with the corresponding approximation for $E[\epsilon_1]$ obtained by simply exchanging n_0

and n_1. This approximation is asymptotically exact, as shown by Theorem 16. However, in the case $n_0 = n_1 = n$, (4.15) reduces to (4.3) and not (4.2). The reason is that, if one compares (4.14) to Raudys' formula (4.1), one observes that the denominators differ by the term:

$$\begin{aligned}&\mathrm{Var}[E(W(\bar{X}_0,\bar{X}_1,X)|\bar{X}_0,\bar{X}_1,X\in\Pi_0)] = \\ &\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X)|X\in\Pi_0)-E[\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X)|\bar{X}_0,\bar{X}_1,X\in\Pi_0)]=\frac{\delta^2}{n_1}+\frac{p}{n_0^2}+\frac{p}{n_1^2}\overset{K}{\to}0\end{aligned} \tag{4.16}$$

Hence, the finite-sample approximations obtained by (4.1) and (4.14) differ, but are asymptotically equivalent. By Theorem 16, this also proves that Raudy's approximation (4.2) is indeed asymptotically exact. For moderate n_0/p and n_1/p, the term (4.16) becomes close to zero, and (4.1) and (4.14) yield very similar values.

The next expression is the finite-sample approximation obtained with Raudys' formula (4.1) in the general case $n_0 \neq n_1$, which has not been available before:

$$E[\epsilon_0] \approx \Phi\left(-\frac{\delta}{2}\,\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0}\right)}{f_0\left(n_0,n_1,p,\delta^2\right)}\right) \tag{4.17}$$

which of course reduces to (4.2) when $n_0 = n_1 = n$. If we remove the terms which tend to zero under Kolmogorov asymptotic conditions, then (4.17) becomes:

$$E[\epsilon_0] \approx \Phi\left(-\frac{\delta}{2}\,\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0}\right)}{\sqrt{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}}\right) \tag{4.18}$$

i.e., the same as (4.15), which reduces to (4.3) when $n_0 = n_1 = n$. Also notice that (4.18) corresponds to replacing λ_0 by p/n_0 and λ_1 by p/n_1 in (4.11), as it should. To obtain the corresponding approximations for $E[\epsilon_1]$, it suffices to exchange n_0 and n_1 in (4.17) and (4.18).

3. Resubstitution Error Estimator

Consider the expectation of the resubstitution error estimator $E[\hat{\epsilon}^r]$. Let

$$\begin{aligned} \epsilon_0^r &= P(W(\bar{X}_0, \bar{X}_1, X_1) \le 0 | \bar{X}_0, \bar{X}_1) \\ \epsilon_1^r &= P(W(\bar{X}_0, \bar{X}_1, X_{n_0+1}) > 0 | \bar{X}_0, \bar{X}_1) \end{aligned} \tag{4.19}$$

Note that ϵ_i^r is different from the class-specific resubstitution error $\hat{\epsilon}_i^r$, for $i = 0, 1$. However, it is clear that $E[\epsilon_i^r] = E[\hat{\epsilon}_i^r]$, for $i = 0, 1$. In particular,

$$E[\hat{\epsilon}^r] = \hat{\alpha}_0 E[\epsilon_0^r] + \hat{\alpha}_1 E[\epsilon_1^r] \tag{4.20}$$

Let

$$\begin{aligned} \hat{G}_0^r &= E[W(\bar{X}_0, \bar{X}_1, X_1) \mid \bar{X}_0, \bar{X}_1], & \hat{G}_1^r &= E[W(\bar{X}_0, \bar{X}_1, X_{n_0+1}) \mid \bar{X}_0, \bar{X}_1] \\ \hat{D}_0^r &= Var(W(\bar{X}_0, \bar{X}_1, X_1) \mid \bar{X}_0, \bar{X}_1), & \hat{D}_1^r &= Var(W(\bar{X}_0, \bar{X}_1, X_{n_0+1}) \mid \bar{X}_0, \bar{X}_1) \end{aligned} \tag{4.21}$$

Then

$$\epsilon_0^r = \Phi\left(-\frac{\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}\right), \quad \epsilon_1^r = \Phi\left(\frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}\right) \tag{4.22}$$

Theorem 17. *Consider the sequence of Gaussian discrimination problems defined by (4.8). Then*

$$\underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{\epsilon}^r] = \underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{\epsilon}_0^r] = \underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{\epsilon}_1^r] = \underset{n_0,n_1,p\to\infty}{\text{pklim}} \epsilon_0^r = \underset{n_0,n_1,p\to\infty}{\text{pklim}} \epsilon_1^r = \Phi\left(\frac{-G}{\sqrt{D}}\right) \tag{4.23}$$

where

$$\begin{aligned} G &= \underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{G}_0^r] = -\underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{G}_1^r] = \frac{1}{2}(\delta^2 + \lambda_0 + \lambda_1) \\ D &= \underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{D}_0^r] = \underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{D}_1^r] = \delta^2 + \lambda_0 + \lambda_1 \end{aligned} \tag{4.24}$$

Proof. See Appendix. □

Theorem 17 suggests the following finite-sample approximation:

$$E[\hat{\epsilon}_0^r] \approx \Phi\left(-\frac{E[\hat{G}_0^r]}{\sqrt{E[\hat{D}_0^r]}}\right) = \Phi\left(-\frac{E[W(\bar{X}_0,\bar{X}_1,X_1)]}{\sqrt{E[\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X_1)\mid\bar{X}_0,\bar{X}_1)]}}\right) \tag{4.25}$$

To obtain the corresponding approximation to $E[\hat{\epsilon}_1^r]$, it suffices to replace $\hat{G}_0^r$ by $\hat{G}_1^r$, $\hat{D}_0^r$ by $\hat{D}_1^r$, and X_1 by X_{n_0+1}, and multiply the argument of both Φ functions by -1. Evaluating the expectation in the numerator and denominator of (4.25) yields

$$E[\hat{\epsilon}_0^r] \approx \Phi\left(-\frac{\delta}{2\sqrt{1-\frac{1}{n_0}}}\sqrt{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}\right) \tag{4.26}$$

with the corresponding approximation for $E[\hat{\epsilon}_1^r]$ obtained by exchanging n_0 and n_1. Theorem 17 shows this approximation is asymptotically exact. If $n_0 = n_1 = n$, then (4.26) reduces to

$$E[\hat{\epsilon}_0^r] \approx \Phi\left(-\frac{\delta}{2\sqrt{1-\frac{1}{n}}}\sqrt{1+\frac{2p}{n\delta^2}}\right) \tag{4.27}$$

which is not the same as (4.5) or (4.6). Once again, the reason is that, if one compares (4.25) to Raudys' formula (4.4), one observes that the denominators differ by the term:

$$\begin{aligned}\mathrm{Var}[E(W(\bar{X}_0,\bar{X}_1,X_1)|\bar{X}_0,\bar{X}_1] &= \mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X_1)) - E[\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X_1)|\bar{X}_0,\bar{X}_1]\\ &= \delta^2\left(\frac{1}{n_0}+\frac{1}{n_1}\right)+\frac{p}{2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^2 \xrightarrow{K} 0\end{aligned} \tag{4.28}$$

Hence, the finite-sample approximations obtained by (4.4) and (4.25) differ, but are asymptotically equivalent. Furthermore, both are asymptotically equivalent to (4.6). Incidentally, this proves that both (4.5) and (4.6) are asymptotically exact. For moderate values of n_0, n_1, n_0/p, and n_1/p, the term (4.28) becomes close to zero, and in fact all three approximations give very similar results.

The next expression is the finite-sample approximation obtained with Raudys' formula (4.4) in the general case $n_0 \neq n_1$, which has not been available before:

$$E[\hat{\epsilon}_0^r] \approx \Phi\left(-\frac{\delta}{2}\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}{g_0\left(n_0,n_1,p,\delta^2\right)}\right) \tag{4.29}$$

which of course reduces to (4.5) when $n_0 = n_1 = n$. To obtain the corresponding approximation for $E[\epsilon_1]$, it suffices to exchange n_0 and n_1 in (4.29). If we remove the terms which tend to zero under Kolmogorov asymptotic conditions, then (4.29) and (4.26) both become:

$$E[\hat{\epsilon}^r] \approx E[\hat{\epsilon}_0^r] \approx E[\hat{\epsilon}_1^r] \approx \Phi\left(-\frac{\delta}{2}\sqrt{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}\right) \tag{4.30}$$

which reduces to (4.6) when $n_0 = n_1 = n$. Also notice that (4.30) corresponds to replacing λ_0 by p/n_0 and λ_1 by p/n_1 in (4.23), as it should.

4. Leave-one-out Error Estimator

By virtue of the relation $E[\hat{\epsilon}_{i,n_i}^l] = E[\epsilon_{i,n_i-1}]$, for $i = 0, 1$, the expectation of the leave-one-out error estimator can be obtained by using the results of Section 2, while replacing α_i by $\hat{\alpha}_i$ and n_i by $n_i - 1$, for $i = 0, 1$.

B. Second-order Double Asymptotic Approximation

Here we extend the double asymptotic method to obtain results for the double asymptotic joint distribution of the pair of random variables $(W(\bar{X}_0, \bar{X}_1, X), W(\bar{X}_0, \bar{X}_1, X'))$, which allows one to obtain finite-sample approximations to the second and cross moments of actual and estimated errors, and therefore the bias, variance, and RMS performance measures.

1. Second-order Approximations

We start by considering the entension of equations (4.1) and (4.4) to second moments. Consider the standard bivariate Gaussian distribution function

$$\Phi(a,b;\rho) = \int_{-\infty}^{a}\int_{-\infty}^{b} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho)^2)}\left(x^2+y^2-2\rho xy\right)\right\} dx\,dy \tag{4.31}$$

This corresponds to the distribution function of a joint bivariate Gaussian vector with zero means, unit variances, and correlation coefficient ρ. Note that $\Phi(a,\infty;\rho) = \Phi(a)$ and $\Phi(a,b;0) = \Phi(a)\Phi(b)$. For simplicity of notation, we write $\Phi(a,a;\rho)$ as $\Phi(a;\rho)$. The rectangular-area probabilities involving any jointly Gaussian pair of variables (X,Y) can be written in terms of the standard bivariate Gaussian distribution function:

$$P\left(X \le c, Y \le d\right) = \Phi\left(\frac{c-\mu_X}{\sigma_X}, \frac{d-\mu_Y}{\sigma_Y}; \rho_{XY}\right) \tag{4.32}$$

where $\mu_X = E[X]$, $\mu_Y = E[Y]$, $\sigma_X = \sqrt{\mathrm{Var}(X)}$, $\sigma_Y = \sqrt{\mathrm{Var}(Y)}$, and ρ_{XY} is the correlation coefficient between X and Y.

Using (4.32), we obtain the second-order extension of Raudys' formula (4.1):

$$\begin{aligned} E[\epsilon_0^2] &= P(W(\bar{X}_0,\bar{X}_1,X) \le 0, W(\bar{X}_0,\bar{X}_1,X') \le 0 \mid X, X' \in \Pi_0) \\ &\approx \Phi\left(-\frac{E[W(\bar{X}_0,\bar{X}_1,X)|X\in\Pi_0]}{\sqrt{\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X)\mid X\in\Pi_0)}}; \frac{\mathrm{Cov}(W(\bar{X}_0,\bar{X}_1,X),W(\bar{X}_0,\bar{X}_1,X')|X,X'\in\Pi_0)}{\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X)|X\in\Pi_0)}\right) \end{aligned} \tag{4.33}$$

In the general case $n_0 \neq n_1$, evaluation of the terms in (4.33) yields

$$E[\epsilon_0^2] \approx \Phi\left(-\frac{\delta}{2}\,\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0}\right)}{f_0\left(n_0,n_1,p,\delta^2\right)}; \frac{\frac{1}{n_1}+\frac{p}{2\delta^2}\left(\frac{1}{n_0^2}+\frac{1}{n_1^2}\right)}{f_0^2\left(n_0,n_1,p,\delta^2\right)}\right) \tag{4.34}$$

Equation (4.34) is the second-order extension of (4.17). Similarly, it can be shown that

$$E[\epsilon_0\epsilon_1] \approx \Phi\left(-\frac{\delta}{2}\,\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0}\right)}{f_0\left(n_0,n_1,p,\delta^2\right)}\right)\Phi\left(-\frac{\delta}{2}\,\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}-\frac{1}{n_1}\right)}{f_1\left(n_0,n_1,p,\delta^2\right)}\right) \tag{4.35}$$

The corresponding approximation for $E[\epsilon_1^2]$ is obtained from $E[\epsilon_0^2]$ by exchanging n_0 and n_1.

A key fact is that by removing the terms that tend to zero under Kolmogorov asymptotic conditions the covariance term in (4.34) becomes zero, and the pair of random variables $(W(\bar{X}_0,\bar{X}_1,X),W(\bar{X}_0,\bar{X}_1,X'))$ become independent. This suggests the approximation

$$E[\epsilon_0^2] \approx \left[\Phi\left(-\frac{\delta}{2}\,\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0}\right)}{\sqrt{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}}\right)\right]^2 \tag{4.36}$$

Equation (4.36) is simply the square of the approximation (4.18). The corresponding approximations for $E[\epsilon_0\epsilon_1]$ and $E[\epsilon_1^2]$ are obtained similarly.

To obtain the approximation for the second moment of the resubstitution error, (4.33) is modified by replacing X and X' by X_1 and X_2, respectively:

$$\begin{aligned} E[(\hat{\epsilon}_0^r)^2] &= P(W(\bar{X}_0,\bar{X}_1,X_1)\le 0, W(\bar{X}_0,\bar{X}_1,X_2)\le 0) \\ &\approx \Phi\left(-\frac{E[W(\bar{X}_0,\bar{X}_1,X_1)]}{\sqrt{\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X_1))}};\frac{\mathrm{Cov}(W(\bar{X}_0,\bar{X}_1,X_1),W(\bar{X}_0,\bar{X}_1,X_2))}{\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X_1))}\right) \end{aligned} \tag{4.37}$$

In the general case $n_0 \neq n_1$, (4.37) gives

$$E[(\hat{\epsilon}_0^r)^2] \approx \Phi\left(-\frac{\delta}{2}\,\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}{g_0\left(n_0,n_1,p,\delta^2\right)};\frac{\frac{1}{n_1}+\frac{p}{2\delta^2}\left(\frac{1}{n_1^2}-\frac{1}{n_0^2}\right)}{g_0^2\left(n_0,n_1,p,\delta^2\right)}\right) \tag{4.38}$$

The corresponding approximation for $E[(\hat{\epsilon}_1^r)^2]$ is obtained from $E[(\hat{\epsilon}_0^r)^2]$ by exchanging n_0 and n_1. Similarly, it can be shown that

$$E[\hat{\epsilon}_0^r \hat{\epsilon}_1^r] \approx \Phi\left(-\frac{\delta}{2}\,\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}{g_0\left(n_0,n_1,p,\delta^2\right)}, -\frac{\delta}{2}\,\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}{g_1\left(n_0,n_1,p,\delta^2\right)}; \frac{\frac{1}{n_0}+\frac{1}{n_1}+\frac{p}{2\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_0^2}\right)^2}{g_0\left(n_0,n_1,p,\delta^2\right)g_1\left(n_0,n_1,p,\delta^2\right)}\right) \tag{4.39}$$

Throwing out the terms that tend to zero under Kolmogorov asymptotic conditions in (4.38) gives the approximation

$$E[(\hat{\epsilon}_0^r)^2] \approx \left[\Phi\left(-\frac{\delta}{2}\sqrt{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}\right)\right]^2 \tag{4.40}$$

Equation (4.40) is simply the square of the approximation (4.30). The corresponding approximations for $E[\hat{\epsilon}_0^r \hat{\epsilon}_1^r]$ and $E[(\hat{\epsilon}_1^r)^2]$ are obtained similarly.

The approximation for the cross-moment between actual and resubstitution errors is

$$\begin{aligned} E[\epsilon_0 \hat{\epsilon}_0^r] &= P(W(\bar{X}_0, \bar{X}_1, X) \le 0, W(\bar{X}_0, \bar{X}_1, X_1) \le 0 \mid X \in \Pi_0) \approx \\ &\Phi\left(\frac{-E[W(\bar{X}_0,\bar{X}_1,X)|X\in\Pi_0]}{\sqrt{\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X)|X\in\Pi_0))}}, \frac{-E[W(\bar{X}_0,\bar{X}_1,X_1)]}{\sqrt{\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X_1))}};\right. \\ &\left.\frac{\mathrm{Cov}(W(\bar{X}_0,\bar{X}_1,X), W(\bar{X}_0,\bar{X}_1,X_1) \mid X \in \Pi_0)}{\sqrt{\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X \mid X \in \Pi_0))}\sqrt{\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X_1))}}\right) \end{aligned} \tag{4.41}$$

In the general case $n_0 \ne n_1$, (4.41) gives

$$E[\epsilon_0 \hat{\epsilon}_0^r] \approx \Phi\left(-\frac{\delta}{2}\,\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0}\right)}{f_0\left(n_0,n_1,p,\delta^2\right)}, -\frac{\delta}{2}\,\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}{g_0\left(n_0,n_1,p,\delta^2\right)}; \frac{\frac{1}{n_1}+\frac{p}{2\delta^2}\left(\frac{1}{n_1^2}-\frac{1}{n_0^2}\right)}{f_0\left(n_0,n_1,p,\delta^2\right)g_0\left(n_0,n_1,p,\delta^2\right)}\right) \tag{4.42}$$

Similarly, it can be shown that

$$E[\epsilon_0 \hat{\epsilon}_1^r] \approx \Phi\left(-\frac{\delta}{2}\,\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0}\right)}{f_0\left(n_0,n_1,p,\delta^2\right)}, -\frac{\delta}{2}\,\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}{g_1\left(n_0,n_1,p,\delta^2\right)}; \frac{\frac{1}{n_1}+\frac{p}{2\delta^2}\left(\frac{1}{n_1^2}-\frac{1}{n_0^2}\right)}{f_0\left(n_0,n_1,p,\delta^2\right)g_1\left(n_0,n_1,p,\delta^2\right)}\right) \tag{4.43}$$

The corresponding approximations for $E[\epsilon_1 \hat{\epsilon}_0^r]$, and $E[\epsilon_1 \hat{\epsilon}_1^r]$ are obtained from $E[\epsilon_0 \hat{\epsilon}_1^r]$ and $E[\epsilon_0 \hat{\epsilon}_0^r]$ by exchanging n_0 and n_1, respectively .

Throwing out the terms that tend to zero under Kolmogorov asymptotic conditions in (4.42) gives the approximation

$$E[\epsilon_0 \hat{\epsilon}_0^r] \approx \Phi\left(-\frac{\delta}{2}\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0}\right)}{\sqrt{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}}\right)\Phi\left(-\frac{\delta}{2}\sqrt{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}\right) \tag{4.44}$$

Equation (4.44) is simply the product of the approximations in (4.18) and (4.30). Corresponding approximations for $E[\epsilon_0 \hat{\epsilon}_1^r]$, $E[\epsilon_1 \hat{\epsilon}_0^r]$, and $E[\epsilon_1 \hat{\epsilon}_1^r]$ are obtained similarly.

To obtain the approximation for the second moment of the leave-one-out error $E[(\hat{\epsilon}_0^l)^2]$, (4.37) is modified by replacing $W(\bar{X}_0, \bar{X}_1, X_1)$ by $W^{(1)}(\bar{X}_0, \bar{X}_1, X_1)$ and $W(\bar{X}_0, \bar{X}_1, X_2)$ by $W^{(2)}(\bar{X}_0, \bar{X}_1, X_2)$. In the general case $n_0 \neq n_1$, this gives

$$E[(\hat{\epsilon}_0^l)^2] \approx \Phi\left(-\frac{\delta}{2}\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0-1}\right)}{f_0\left(n_0-1, n_1, p, \delta^2\right)}; \frac{\frac{1}{n_1}+\frac{p}{2\delta^2}\left(\frac{1}{n_1^2}+\frac{2}{(n_0-1)^4}-\frac{(n_0-2)^2}{(n_0-1)^4}\right)}{f_0^2\left(n_0-1, n_1, p, \delta^2\right)}\right) \tag{4.45}$$

The corresponding approximation for $E[(\hat{\epsilon}_1^l)^2]$ is obtained from $E[(\hat{\epsilon}_0^l)^2]$ by exchanging n_0 and n_1, respectively. Similarly,

$$\begin{aligned} E[\hat{\epsilon}_0^l \hat{\epsilon}_1^l] \approx \Phi\Bigg(&-\frac{\delta}{2}\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0-1}\right)}{f_0\left(n_0-1, n_1, p, \delta^2\right)}, -\frac{\delta}{2}\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}-\frac{1}{n_1-1}\right)}{f_1\left(n_0, n_1-1, p, \delta^2\right)}; \\ &\frac{\frac{1}{n_0}+\frac{1}{n_1}+\frac{p}{2\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^2}{f_0\left(n_0-1, n_1, p, \delta^2\right) f_1\left(n_0, n_1-1, p, \delta^2\right)}\Bigg) \end{aligned} \tag{4.46}$$

The corresponding approximation for $E[(\hat{\epsilon}_1^l)^2]$ is obtained from $E[(\hat{\epsilon}_0^l)^2]$ by exchanging n_0 and n_1, respectively .

Throwing out the terms that tend to zero under Kolmogorov asymptotic conditions in (4.38) gives the approximation

$$E[(\hat{\epsilon}_0^l)^2] \approx \left[\Phi\left(-\frac{\delta}{2}\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{(n_0-1)}\right)}{\sqrt{1+\frac{p}{\delta^2}\left(\frac{1}{(n_0-1)}+\frac{1}{n_1}\right)}}\right)\right]^2 \tag{4.47}$$

Equation (4.47) is simply the square of the approximation (4.18), with n_0 replaced by $n_0 - 1$. The corresponding approximations for $E[\hat{\epsilon}_0^l \hat{\epsilon}_1^l]$ and $E[(\hat{\epsilon}_1^l)^2]$ are obtained similarly.

The approximation for the cross-moment $E[\epsilon_0 \hat{\epsilon}_0^l]$ between actual and leave-one-out errors is obtained by replacing $W(\bar{X}_0, \bar{X}_1, X_1)$ by $W^{(1)}(\bar{X}_0, \bar{X}_1, X_1)$ in (4.41). The corresponding approximations for $E[\epsilon_0 \hat{\epsilon}_1^l]$, $E[\epsilon_1 \hat{\epsilon}_0^l]$, and $E[\epsilon_1 \hat{\epsilon}_1^l]$ are entirely similar. When $n_0 \neq n_1$, this gives

$$\begin{aligned}
E[\epsilon_0\hat{\epsilon}_0^l] &\approx \Phi\left(-\frac{\delta}{2}\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0}\right)}{f_0\left(n_0,n_1,p,\delta^2\right)}, -\frac{\delta}{2}\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0-1}\right)}{f_0\left(n_0-1,n_1,p,\delta^2\right)}; \frac{\frac{1}{n_1}+\frac{p}{2\delta^2}\left(\frac{1}{n_1^2}-\frac{1}{n_0^2}\right)}{f_0\left(n_0,n_1,p,\delta^2\right)f_0\left(n_0-1,n_1,p,\delta^2\right)}\right)\\
E[\epsilon_0\hat{\epsilon}_1^l] &\approx \Phi\left(-\frac{\delta}{2}\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0}\right)}{f_0\left(n_0,n_1,p,\delta^2\right)}, -\frac{\delta}{2}\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}-\frac{1}{n_1-1}\right)}{f_1\left(n_0,n_1-1,p,\delta^2\right)}; \frac{\frac{1}{n_1}+\frac{p}{2\delta^2}\left(\frac{1}{n_1^2}-\frac{1}{n_0^2}\right)}{f_0\left(n_0,n_1,p,\delta^2\right)f_1\left(n_0,n_1-1,p,\delta^2\right)}\right)
\end{aligned} \tag{4.48}$$

The corresponding approximations for $E[\epsilon_1 \hat{\epsilon}_0^l]$, and $E[\epsilon_1 \hat{\epsilon}_1^l]$ are obtained from $E[\epsilon_0 \hat{\epsilon}_1^l]$ and $E[\epsilon_0 \hat{\epsilon}_0^l]$ by exchanging n_0 and n_1, respectively .

Throwing out the terms that tend to zero under Kolmogorov asymptotic conditions in (4.42) gives the approximation

$$E[\epsilon_0\hat{\epsilon}_0^l] \approx \Phi\left(-\frac{\delta}{2}\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0}\right)}{\sqrt{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}}\right)\Phi\left(-\frac{\delta}{2}\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0-1}\right)}{\sqrt{1+\frac{p}{\delta^2}\left(\frac{1}{n_0-1}+\frac{1}{n_1}\right)}}\right) \tag{4.49}$$

Equation (4.44) is simply the product of the approximations in (4.18) and itself with n_0 replaced by $n_0 - 1$. The approximations for $E[\epsilon_0 \hat{\epsilon}_1^l]$, $E[\epsilon_1 \hat{\epsilon}_0^l]$ and $E[\epsilon_1 \hat{\epsilon}_1^l]$ are obtained similarly.

We will prove in the following subsections that all the second-order approximations discussed above are asymptotically exact under Kolmogorov asymptotic conditions.

2. Actual Classification Error

Note that the populations specific errors satisfy

$$\epsilon_0^2 = \left[\Phi\left(-\frac{\hat{G}_0}{\sqrt{\hat{D}_0}}\right)\right]^2, \epsilon_0\epsilon_1 = \Phi\left(-\frac{\hat{G}_0}{\sqrt{\hat{D}_0}}\right)\Phi\left(\frac{\hat{G}_1}{\sqrt{\hat{D}_1}}\right), \epsilon_1^2 = \left[\Phi\left(\frac{\hat{G}_1}{\sqrt{\hat{D}_1}}\right)\right]^2 \tag{4.50}$$

where $\hat{G}_i$ and $\hat{D}_i$ were defined in (4.9). Using the results of Theorem 16, we obtain:

Theorem 18. *Consider the sequence of Gaussian discrimination problems defined by (4.8). Then*

$$\underset{n_0,n_1,p\to\infty}{\text{klim}} E[\epsilon_0^2] = \underset{n_0,n_1,p\to\infty}{\text{pklim}} \epsilon_0^2 = \left[\Phi\left(\frac{-G_0}{\sqrt{D}}\right)\right]^2, \qquad \underset{n_0,n_1,p\to\infty}{\text{klim}} E[\epsilon_1^2] = \underset{n_0,n_1,p\to\infty}{\text{pklim}} \epsilon_1^2 = \left[\Phi\left(\frac{G_1}{\sqrt{D}}\right)\right]^2$$
$$\underset{n_0,n_1,p\to\infty}{\text{klim}} E[\epsilon_0\epsilon_1] = \underset{n_0,n_1,p\to\infty}{\text{pklim}} \epsilon_0\epsilon_1 = \Phi\left(-\frac{G_0}{\sqrt{D_0}}\right)\Phi\left(\frac{G_1}{\sqrt{D_1}}\right) \tag{4.51}$$

so that

$$\underset{n_0,n_1,p\to\infty}{\text{klim}} E[\epsilon^2] = \underset{n_0,n_1,p\to\infty}{\text{pklim}} \epsilon^2 = \left(\underset{n_0,n_1,p\to\infty}{\text{klim}} E[\epsilon]\right)^2 = \left[\alpha_0\Phi\left(-\frac{G_0}{\sqrt{D}}\right) + \alpha_1\Phi\left(\frac{G_1}{\sqrt{D}}\right)\right]^2 \tag{4.52}$$

where G_0, G_1 and D are the same as in (4.13).

Theorem 18 suggests the following finite-sample approximation:

$$E[\epsilon_0^2] \approx \left[\Phi\left(-\frac{E[\hat{G}_0]}{\sqrt{E[\hat{D}_0]}}\right)\right]^2 = \left[\Phi\left(-\frac{E[W(\bar{X}_0,\bar{X}_1,X) \mid X \in \Pi_0]}{\sqrt{E[\text{Var}(W(\bar{X}_0,\bar{X}_1,X) \mid \bar{X}_0,\bar{X}_1, X \in \Pi_0)]}}\right)\right]^2 \tag{4.53}$$

with similar approximations for $E[\epsilon_0\epsilon_1]$ and $[\epsilon_1^2]$ derived from (4.51). These approximations are asymptotically exact, as shown by Theorem 18. Recalling (4.15), we see that (4.53) yields (4.36), showing that both (4.36) and (4.34) are asymptotically exact under the Kolmogorov limit. For moderate n_0/p and n_1/p, the two approximations yield very similar results.

An asymptotically exact approximation to the full second moment $E[\epsilon^2]$ is obtained from (4.52) upon replacing λ_0 by p/n_0 and λ_1 by p/n_1.

3. Resubstitution Error Estimator

In this section, we are interested in the second moment of the resubstitution error estimator $E[(\hat{\epsilon}^r)^2]$ and the cross-moment with the actual classification error $E[\epsilon\hat{\epsilon}^r]$. Let

$$\begin{aligned}\epsilon_{00}^r &= P(W(\bar{X}_0,\bar{X}_1,X_1) \le 0, W(\bar{X}_0,\bar{X}_1,X_2) \le 0 \mid \bar{X}_0,\bar{X}_1)\\ \epsilon_{01}^r &= P(W(\bar{X}_0,\bar{X}_1,X_1) \le 0, W(\bar{X}_0,\bar{X}_1,X_{n_0+1}) > 0 \mid \bar{X}_0,\bar{X}_1)\\ \epsilon_{11}^r &= P(W(\bar{X}_0,\bar{X}_1,X_{n_0+1}) > 0, W(\bar{X}_0,\bar{X}_1,X_{n_0+2}) > 0 \mid \bar{X}_0,\bar{X}_1)\end{aligned} \tag{4.54}$$

Note that $E[(\hat{\epsilon}_0^r)^2] = E[\epsilon_{00}^r]$, $E[\hat{\epsilon}_0^r\hat{\epsilon}_1^r] = E[\epsilon_{01}^r]$ and $E[(\hat{\epsilon}_1^r)^2] = E[\epsilon_{11}^r]$. From representation of $E[(\hat{\epsilon}^r)^2]$ given in (3.10), it follows that

$$E[(\hat{\epsilon}^r)^2] = \frac{\hat{\alpha}_0^2}{n_0}E[\epsilon_0^r] + \frac{\hat{\alpha}_1^2}{n_1}E[\epsilon_1^r] + \hat{\alpha}_0^2\,\frac{n_0-1}{n_0}E[\epsilon_{00}^r] + \hat{\alpha}_1^2\,\frac{n_1-1}{n_1}E[\epsilon_{11}^r] + 2\hat{\alpha}_0\hat{\alpha}_1 E[\epsilon_{01}^r] \tag{4.55}$$

where ϵ_0^r and ϵ_1^r are defined in (4.19).

Let

$$\begin{aligned}\hat{H}_0^r &= \mathrm{Cov}(W(\bar{X}_0,\bar{X}_1,X_1), W(\bar{X}_0,\bar{X}_1,X_2) \mid \bar{X}_0,\bar{X}_1)\\ \hat{H}_{01}^r &= \mathrm{Cov}(W(\bar{X}_0,\bar{X}_1,X_1), W(\bar{X}_0,\bar{X}_1,X_{n_0+1}) \mid \bar{X}_0,\bar{X}_1)\\ \hat{H}_1^r &= \mathrm{Cov}(W(\bar{X}_0,\bar{X}_1,X_{n_0+1}), W(\bar{X}_0,\bar{X}_1,X_{n_0+2}) \mid \bar{X}_0,\bar{X}_1)\end{aligned} \tag{4.56}$$

$$\epsilon_{00}^r = \Phi\left(\frac{-\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}\,;\frac{\hat{H}_0^r}{\hat{D}_0^r}\right),\ \epsilon_{11}^r = \Phi\left(\frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}\,;\frac{\hat{H}_1^r}{\hat{D}_1^r}\right),\ \epsilon_{01}^r = \Phi\left(\frac{-\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}\,;\frac{\hat{H}_0^r}{\hat{D}_0^r}\right) - \Phi\left(\frac{-\hat{G}_0^r}{\sqrt{\hat{D}_0^r}},\frac{-\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}\,;\frac{\hat{H}_{01}^r}{\hat{D}_{01}^r}\right) \tag{4.57}$$

where $\hat{G}_i^r$ and $\hat{D}_i^r$ were defined in (4.21).

Theorem 19. *For the sequence of Gaussian discrimination problems defined by (4.8),*

$$\begin{aligned} &\underset{n_0,n_1,p\to\infty}{\text{klim}} E[\epsilon_{00}^r] = \underset{n_0,n_1,p\to\infty}{\text{klim}} E[\epsilon_{01}^r] = \underset{n_0,n_1,p\to\infty}{\text{klim}} E[\epsilon_{11}^r] \\ &= \underset{n_0,n_1,p\to\infty}{\text{pklim}}\ \epsilon_{00}^r = \underset{n_0,n_1,p\to\infty}{\text{pklim}}\ \epsilon_{01}^r = \underset{n_0,n_1,p\to\infty}{\text{pklim}}\ \epsilon_{11}^r = \Phi\left(-\frac{G}{\sqrt{D}}\,;\frac{H}{D}\right) = \left[\Phi\left(-\frac{G}{\sqrt{D}}\right)\right]^2 \end{aligned} \tag{4.58}$$

and

$$\underset{n_0,n_1,p\to\infty}{\text{klim}} E[(\hat{\epsilon}^r)^2] = \left(\underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{\epsilon}^r]\right)^2 = \left[\Phi\left(-\frac{G}{\sqrt{D}}\right)\right]^2 \tag{4.59}$$

where G and D are given in (4.24) and $H = \underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{H}_0^r] = \underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{H}_1^r] = 0\,.$

Proof. See Appendix. □

Theorem 19 suggests the following finite-sample approximation:

$$\begin{aligned} E[(\hat{\epsilon}_0^r)^2] &\approx \Phi\left(-\frac{E[\hat{G}_0^r]}{\sqrt{E[\hat{D}_0^r]}}\,;\,\frac{E[\hat{H}_0^r]}{E[\hat{D}_0^r]}\right) \\ &= \Phi\Bigg(-\frac{E[W(\bar{X}_0,\bar{X}_1,X_1)]}{\sqrt{E[\text{Var}(W(\bar{X}_0,\bar{X}_1,X_1)\mid\bar{X}_0,\bar{X}_1)]}}; \\ &\qquad \frac{E[\text{Cov}(W(\bar{X}_0,\bar{X}_1,X_1),W(\bar{X}_0,\bar{X}_1,X_2)\mid\bar{X}_0,\bar{X}_1)]}{E[\text{Var}(W(\bar{X}_0,\bar{X}_1,X_1)\mid\bar{X}_0,\bar{X}_1)]}\Bigg) \end{aligned} \tag{4.60}$$

with corresponding approximations to $E[\hat{\epsilon}_0^r\hat{\epsilon}_1^r]$ and $E[(\hat{\epsilon}_1^r)^2]$ being obtained from (4.57). These approximations are asymptotically exact, as shown by Theorem 19. Eq. (4.60) yields

$$E[(\hat{\epsilon}_0^r)^2] \approx \Phi\left(-\frac{\delta}{2\sqrt{1-\frac{1}{n_0}}}\sqrt{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}\,;-\frac{1}{n_0-1}\right) \tag{4.61}$$

If one throws out extra terms that tend to zero under the Kolmogorov limit, this reduces to (4.40), showing that both (4.40) and (4.38) are asymptotically exact under the Kolmogorov limit. For moderate n_0/p and n_1/p, the three approximations yield very similar results.

An asymptotically exact approximation to the full second moment $E[(\hat{\epsilon}^r)^2]$ is obtained from (4.59) upon replacing λ_0 by p/n_0 and λ_1 by p/n_1.

To find the cross-expectation between true error and resubstitution, we can use the representation of $E[\epsilon\hat{\epsilon}^r]$ given in (3.13) in conjunction with the independence of testing and training samples to show $E[\epsilon^i\hat{\epsilon}_j^r] = E[\epsilon^i\epsilon_j^r]$ for $i, j = 0, 1$. Thus,

$$\begin{aligned}
E[\epsilon\hat{\epsilon}^r] &= \alpha_0\hat{\alpha}_0 E[\epsilon^0\epsilon_0^r] + \alpha_0\hat{\alpha}_1 E[\epsilon^0\epsilon_1^r] + \alpha_1\hat{\alpha}_0 E[\epsilon^1\epsilon_0^r] + \alpha_1\hat{\alpha}_1 E[\epsilon^1\epsilon_1^r] \\
&= \alpha_0\hat{\alpha}_0 E\left[\Phi\left(\frac{-\hat{G}_0}{\sqrt{\hat{D}_0}}\right)\Phi\left(\frac{-\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}\right)\right] + \alpha_0\hat{\alpha}_1 E\left[\Phi\left(\frac{-\hat{G}_0}{\sqrt{\hat{D}_0}}\right)\Phi\left(\frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}\right)\right] \\
&+ \alpha_1\hat{\alpha}_0 E\left[\Phi\left(\frac{\hat{G}_1}{\sqrt{\hat{D}_1}}\right)\Phi\left(\frac{-\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}\right)\right] + \alpha_1\hat{\alpha}_1 E\left[\Phi\left(\frac{\hat{G}_1}{\sqrt{\hat{D}_1}}\right)\Phi\left(\frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}\right)\right]
\end{aligned} \tag{4.62}$$

where where $\hat{G}_i$ and $\hat{D}_i$ were defined in (4.9), and $\hat{G}_i^r$ and $\hat{D}_i^r$ were defined in (4.21). Using the results of Theorems 16 and 17, the following result immediately follows.

Theorem 20. *For the sequence of Gaussian discrimination problems defined by (4.8),*

$$\begin{aligned}
\operatorname*{klim}_{n_0,n_1,p\to\infty} E[\epsilon_0\hat{\epsilon}_0^r] = \operatorname*{klim}_{n_0,n_1,p\to\infty} E[\epsilon_0\hat{\epsilon}_1^r] = \operatorname*{pklim}_{n_0,n_1,p\to\infty} \epsilon_0\epsilon_0^r = \operatorname*{pklim}_{n_0,n_1,p\to\infty} \epsilon_0\epsilon_1^r = \Phi\left(\frac{-G_0}{\sqrt{D}}\right)\Phi\left(\frac{-G}{\sqrt{D}}\right) \\
\operatorname*{klim}_{n_0,n_1,p\to\infty} E[\epsilon_1\hat{\epsilon}_0^r] = \operatorname*{klim}_{n_0,n_1,p\to\infty} E[\epsilon_1\hat{\epsilon}_1^r] = \operatorname*{pklim}_{n_0,n_1,p\to\infty} \epsilon_1\epsilon_0^r = \operatorname*{pklim}_{n_0,n_1,p\to\infty} \epsilon_1\epsilon_1^r = \Phi\left(\frac{G_1}{\sqrt{D}}\right)\Phi\left(\frac{-G}{\sqrt{D}}\right)
\end{aligned} \tag{4.63}$$

so that

$$\operatorname*{klim}_{n_0,n_1,p\to\infty} E[\epsilon\hat{\epsilon}^r] = \left(\operatorname*{klim}_{n_0,n_1,p\to\infty} E[\epsilon]\right)\left(\operatorname*{klim}_{n_0,n_1,p\to\infty} E[\hat{\epsilon}^r]\right) = \Phi\left(\frac{-G}{\sqrt{D}}\right)\left[\alpha_0\Phi\left(\frac{-G_0}{\sqrt{D}}\right) + \alpha_1\Phi\left(\frac{G_1}{\sqrt{D}}\right)\right] \tag{4.64}$$

where G_0, G_1, G and D are the same as in (4.13) and (4.24).

Theorem 20 suggests the following finite-sample approximation:

$$
\begin{aligned}
E[\epsilon_0 \hat{\epsilon}_0^r] &\approx \Phi\left(-\frac{E[\hat{G}_0]}{\sqrt{E[\hat{D}_0]}}\right)\Phi\left(-\frac{E[\hat{G}_0^r]}{\sqrt{E[\hat{D}_0^r]}}\right) \\
&= \Phi\left(-\frac{E[W(\bar{X}_0,\bar{X}_1,X) \mid X \in \Pi_0]}{\sqrt{E[\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X) \mid \bar{X}_0,\bar{X}_1,X \in \Pi_0)]}}\right) \qquad (4.65) \\
&\times\ \Phi\left(-\frac{E[W(\bar{X}_0,\bar{X}_1,X_1)]}{\sqrt{E[\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X_1) \mid \bar{X}_0,\bar{X}_1)]}}\right)
\end{aligned}
$$

with corresponding approximations to $E[\epsilon_0 \hat{\epsilon}_1^r]$, $E[\epsilon_0 \hat{\epsilon}_1^r]$, and $E[\epsilon_0 \hat{\epsilon}_1^r]$ being obtained from (4.62). By Theorem 20, these approximations are asymptotically exact. Eq. (4.65) yields

$$
E[\epsilon_0 \hat{\epsilon}_0^r] \approx \Phi\left(-\frac{\delta}{2}\,\frac{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}-\frac{1}{n_0}\right)}{\sqrt{1+\frac{p}{\delta^2}\left(\frac{1}{n_1}+\frac{1}{n_0}\right)}}\right)\Phi\left(-\frac{\delta}{2\sqrt{1-\frac{1}{n_0}}}\sqrt{1+\frac{p}{\delta^2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}\right) \qquad (4.66)
$$

If one throws out extra terms that tend to zero under the Kolmogorov limit, this reduces to (4.44), showing that both (4.44) and (4.42) are asymptotically exact under the Kolmogorov limit. For moderate n_0/p and n_1/p, the three approximations yield very similar results.

An asymptotically exact approximation to the full second moment $E[\epsilon\hat{\epsilon}^r]$ is obtained from (4.64) upon replacing λ_0 by p/n_0 and λ_1 by p/n_1.

4. Leave-one-out Error Estimator

In theorem 16, we showed that $\underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\epsilon_0, n] = \Phi\left(-\frac{G_0}{\sqrt{D}}\right)$. It follows that

$$
\underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\hat{\epsilon}_{0,n_0}^l] = \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\epsilon_{0,n_0-1}] = \underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ \hat{\epsilon}_{0,n}^l = \Phi\left(-\frac{G_0}{\sqrt{D}}\right) \qquad (4.67)
$$

A similar fact applies to $\underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\hat{\epsilon}_{1,n_1}^l]$. On the other hand, if $(X_p, Y_p) \xrightarrow{P} (X, Y)$, then $X_p Y_p \xrightarrow{P} XY$, by the Continuous Mapping Theorem [133]. Thus, we have the following result.

Theorem 21. *For the sequence of Gaussian discrimination problems defined by (4.8),*

$$\begin{aligned}
\underset{n_0,n_1,p\to\infty}{\text{klim}} E[(\hat{\epsilon}_0^l)^2] &= \underset{n_0,n_1,p\to\infty}{\text{pklim}} (\hat{\epsilon}_0^l)^2 = \left[\Phi\left(-\frac{G_0}{\sqrt{D}}\right)\right]^2 \\
\underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{\epsilon}_0^l\hat{\epsilon}_1^l] &= \underset{n_0,n_1,p\to\infty}{\text{pklim}} \hat{\epsilon}_0^l\hat{\epsilon}_1^l = \Phi\left(-\frac{G_0}{\sqrt{D_0}}\right)\Phi\left(\frac{G_1}{\sqrt{D_1}}\right) \qquad (4.68) \\
\underset{n_0,n_1,p\to\infty}{\text{klim}} E[(\hat{\epsilon}_1^l)^2] &= \underset{n_0,n_1,p\to\infty}{\text{pklim}} (\hat{\epsilon}_1^l)^2 = \left[\Phi\left(\frac{G_1}{\sqrt{D}}\right)\right]^2
\end{aligned}$$

so that

$$\underset{n_0,n_1,p\to\infty}{\text{klim}} E[(\hat{\epsilon}^l)^2] = \underset{n_0,n_1,p\to\infty}{\text{pklim}} (\hat{\epsilon}^l)^2 = \left(\underset{n_0,n_1,p\to\infty}{\text{klim}} E[\hat{\epsilon}^l]\right)^2 = \frac{1}{\lambda_0+\lambda_1}\left[\lambda_0\Phi\left(\frac{-G_0}{\sqrt{D}}\right) + \lambda_1\Phi\left(\frac{G_1}{\sqrt{D}}\right)\right]^2 \qquad (4.69)$$

where G_0, G_1 and D are the same as in (4.13).

Similar expressions are obtained for $E[\epsilon\hat{\epsilon}^l]$. An asymptotically exact approximation to the full second moment $E[(\hat{\epsilon}_0^l)^2]$ is obtained by replacing λ_0 by p/n_0 and λ_1 by p/n_1. However, the fact that $E[\hat{\epsilon}_{0,n_0}^l] = E[\epsilon_{0,n_0-1}]$ and $E[\hat{\epsilon}_{1,n_1}^l] = E[\epsilon_{1,n_1-1}]$ suggests that a more precise approximation is to replace λ_0 by $\frac{p}{n_0-1}$ and λ_1 by $\frac{p}{n_1-1}$, which results in an expression equivalent to (4.47).

Figures 13–15 provide graphical demonstration of the basic performance measures using the asymptotically-exact approximations for resubstitution and leave-one-out, as a function of total sample size, the balanced case $n_0 = n_1 = n$ being assumed throughout, so that the x-axis represents $2n$. Monte-Carlo approximations are also displayed to illustrate the finite-sample accuracy of the approximations. Two Gaussians with different means and equal covariance matrix have been employed such that the Mahalanobis distance $\delta^2 = 4$ corresponds to Bayes error = 0.1586. Figure 16 displays a plot of the RMS of resubstitution and leave-one-out as functions of both sample size and dimensionality and again assuming $n_0 = n_1 = n$. The Gaussian distributions used here have means $\mu_0 = -\mathbf{1}_{p\times 1}$ and $\mu_1 = \mathbf{1}_{p\times 1}$ with equal covariance matrices in which the diagonal elements are 1 and off-diagonal elements are ρ. Notice that here we have not fixed the Bayes error. This allows one to determine the minimum value of RMS in terms of both sample size and dimensionality, shown by the pink line. Notice that for each sample size, the RMS decreases as a function of p and then increases for increasing p. We refer to this phenomenon as RMS peaking.

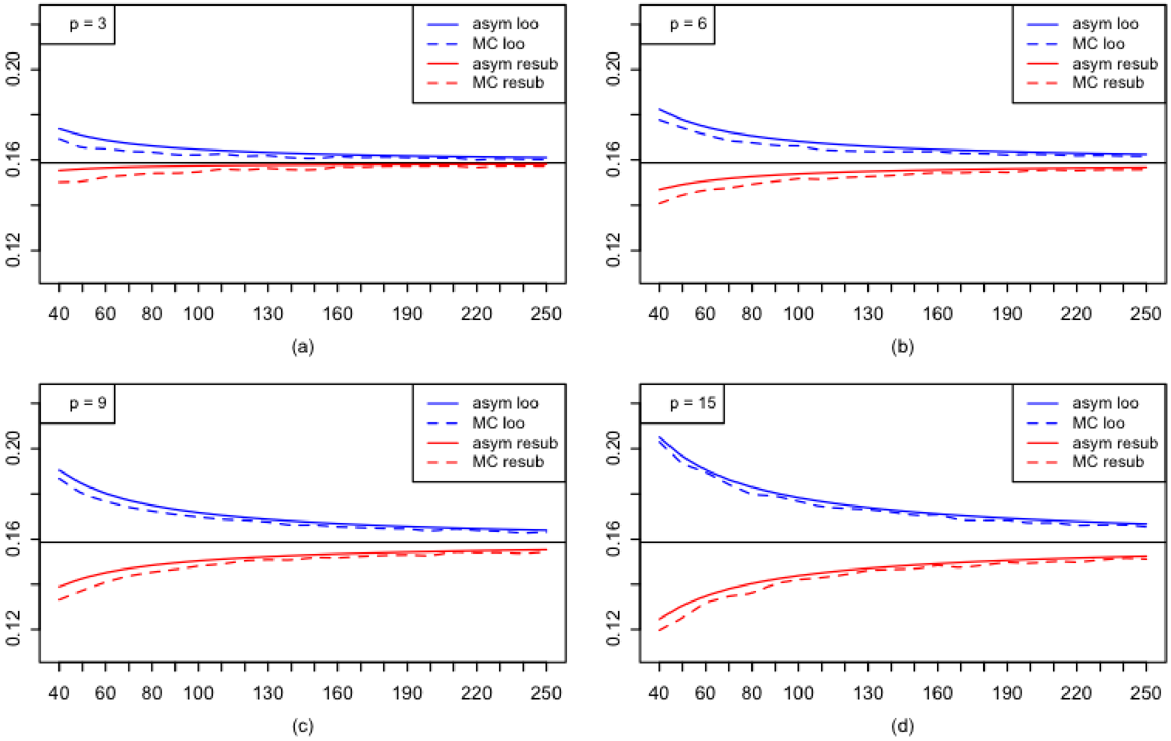

Fig. 13. Comparison of expectation for resubstitution and leave-one-out using the double asymptotic approximation with Monte Carlo estimates as a function of sample size for dimensions p = 3, 6, 9, and 15 (Bayes error 0.1586).

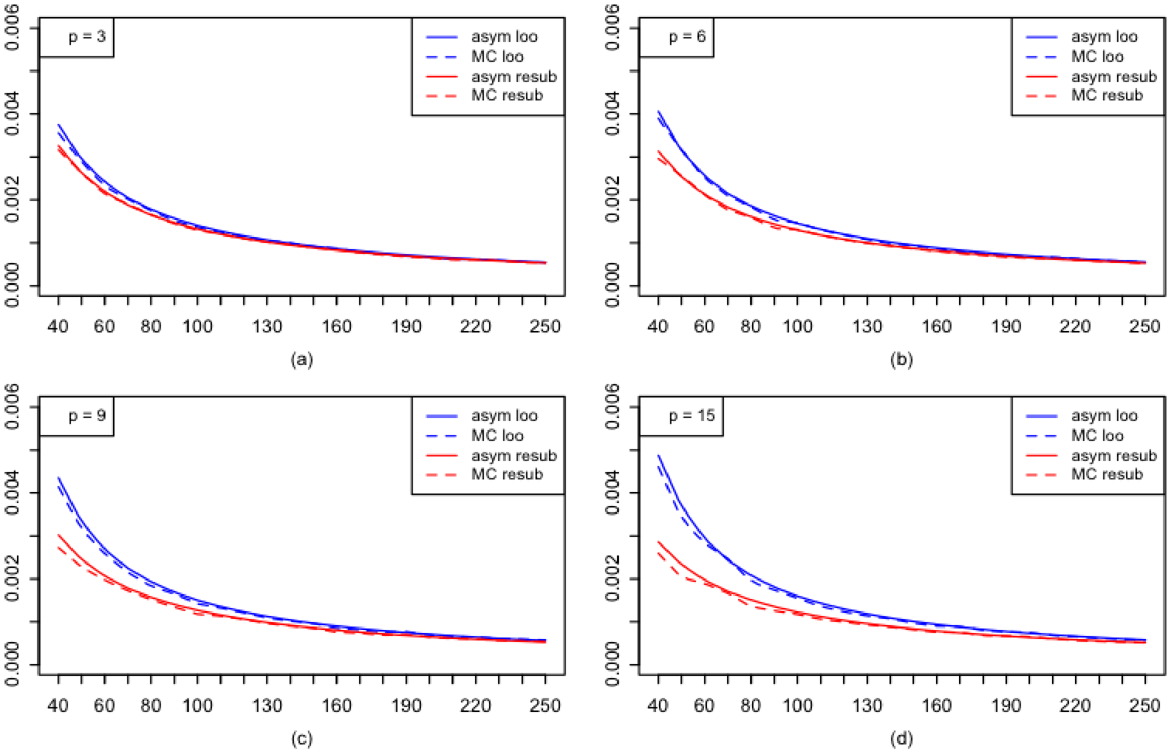

Fig. 14. Comparison of deviation variance for resubstitution and leave-one-out using the double asymptotic approximation with Monte Carlo estimates as a function of sample size for dimensions p = 3, 6, 9, and 15 (Bayes error 0.1586).

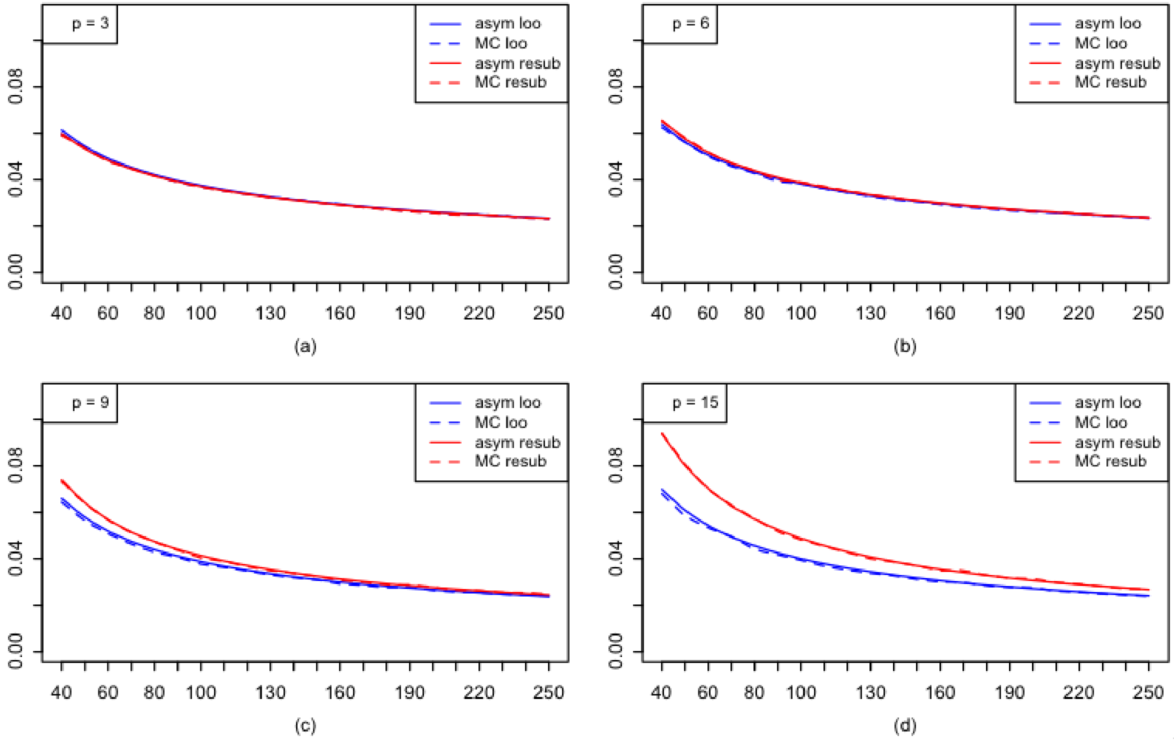

Fig. 15. Comparison of RMS for resubstitution and leave-one-out using the double asymptotic approximation with Monte Carlo estimates as a function of sample size for dimensions p = 3, 6, 9, and 15 (Bayes error 0.1586).

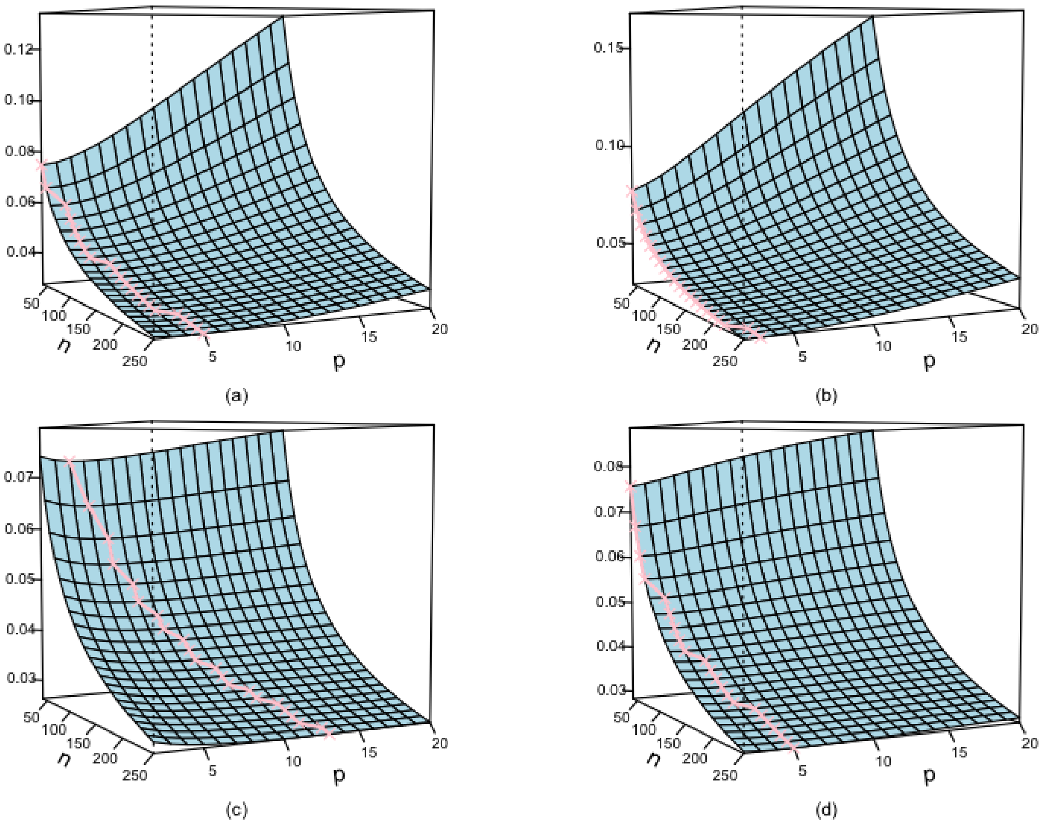

Fig. 16. Demonstration of RMS peaking phenomenon: (a) resubstitution, $\rho = 0.3$; (b) resubstitution, $\rho = 0.5$; (c) leave-one-out, $\rho = 0.3$; (d) leave-one-out, $\rho = 0.5$.

C. Asymptotic Performance of Error Estimation

In this section we state the consequences of the Theorems 16–21 to the limiting values of bias, variance, and RMS of resubstitution and leave-one-out error estimators under Kolmogorov asymptotic conditions.

From Theorems 16 and 17, we conclude that the asymptotic bias of resubstitution is given by (for the sake of simplicity, we consider here the asymptotically balanced case $\lambda_1 = \lambda_0 = \lambda$):

$$\operatorname*{klim}_{n_0,n_1,p\to\infty} \mathrm{Bias}[\hat{\epsilon}^r] = \Phi\left(-\frac{\delta}{2}\frac{1}{\sqrt{1+\frac{2\lambda}{\delta^2}}}\right) - \Phi\left(-\frac{\delta}{2}\sqrt{1+\frac{2\lambda}{\delta^2}}\right) < 0 \tag{4.70}$$

Therefore, resubstitution has an optimistic asymptotic bias. Recalling that under the Kolmogorov limit we have $n_0/p, n_1/p \to 1/\lambda$, we observe that this bias disappears as the sample sizes n_0, n_1 grow much faster than the dimensionality p. In fact, this is also true if the opposite happens and the dimensionality grows much faster than the sample sizes; however, this corresponds to the no-information case $\operatorname*{klim}_{n_0,n_1,p\to\infty} E[\epsilon] = \operatorname*{klim}_{n_0,n_1,p\to\infty} E[\hat{\epsilon}^r] = \frac{1}{2}$.

As for the asymptotic bias of leave-one-out, since $E[\hat{\epsilon}^l_{i,n_i}] = E[\epsilon_{i,n_i-1}]$, for $i = 0, 1$, $\operatorname*{klim}_{n_0,n_1,p\to\infty} \mathrm{Bias}[\hat{\epsilon}^l]=0$. This is true also in the unbalanced case $\lambda_0 \neq \lambda_1$.

A perhaps surprising consequence of Theorems 16–21 is that all variances and covariances are asymptotically zero, i.e.,

$$\operatorname*{klim}_{n_0,n_1,p\to\infty} \mathrm{Var}(\epsilon) = \operatorname*{klim}_{n_0,n_1,p\to\infty} \mathrm{Var}(\hat{\epsilon}^r) = \operatorname*{klim}_{n_0,n_1,p\to\infty} \mathrm{Var}(\hat{\epsilon}^l) = \operatorname*{klim}_{n_0,n_1,p\to\infty} \mathrm{Cov}(\epsilon,\hat{\epsilon}^r) = \operatorname*{klim}_{n_0,n_1,p\to\infty} \mathrm{Cov}(\epsilon,\hat{\epsilon}^l) = 0 \tag{4.71}$$

and this is true also in the unbalanced case $\lambda_0 \neq \lambda_1$. Consequently, the deviation variances are also asymptotically null, i.e.,

$$\operatorname*{klim}_{n_0,n_1,p\to\infty} \mathrm{Var}_d[\hat{\epsilon}^r] = \operatorname*{klim}_{n_0,n_1,p\to\infty} \mathrm{Var}_d[\hat{\epsilon}^l] = 0 \tag{4.72}$$

Hence, $\underset{n_0,n_1,p\to\infty}{\text{klim}}\text{RMS}[\hat{\epsilon}^r] = |\text{Bias}[\hat{\epsilon}^r]|$ whereas $\underset{n_0,n_1,p\to\infty}{\text{klim}}\text{RMS}[\hat{\epsilon}^l] = 0$. The asymptotic RMS of leave-one-out is therefore exactly zero under any limiting rates λ_0 and λ_1 between sample sizes and dimensionality.

D. RMS Bounds

As we considered RMS bounds in the first part of this study for the univariate case, we consider them for the multivariate case, where we must keep in mind that in the present case the expressions for RMS are asymptotic. In intensive simulation studies, we have observed in the multivariate case that the finite sample approximations obtained for $\text{RMS}[\hat{\epsilon}^r]$ and $\text{RMS}[\hat{\epsilon}^l]$ are very accurate when $0 < \epsilon_{bay} < 0.3$ for all dimensions, but while they retain good accuracy when $0.3 < \epsilon_{bay} < 0.5$ for high dimensions, accuracy deteriorates in this high-Bayes-error setting for low dimensions. This can be partially explained by noticing the fact that the finite sample approximations obtained from the Kolmogorov asymptotic conditions are inherently suitable for cases where the dimension is comparable to the sample size. In analogy to how we proceeded in the univariate case in the first part of this study, we note that RMS is an increasing function of the Bayes error, ϵ_{bay}, and we let $\kappa_{\hat{\epsilon}}(n,p,\tau) = \max_{\epsilon_{bay}\leq\tau}\text{RMS}[\hat{\epsilon}]$ for $n_0 = n_1 = n$ and $\hat{\epsilon} = \hat{\epsilon}^r, \hat{\epsilon}^l$. The desired RMS bound is given by $\kappa_{\hat{\epsilon}}(n,p,0.5)$. Note that $\kappa_{\hat{\epsilon}}(n,p,0.5) = \lim_{\delta^2\to 0}\text{RMS}[\hat{\epsilon}]$. Letting $\delta^2 \to 0$ in our asymptotic expressions for the RMS of resubstitution and leave-one-out yields the approximate bounds for finite samples:

$$\text{RMS}[\hat{\epsilon}^r] \leq \kappa_{\hat{\epsilon}^r}(n,p,0.5) \approx \sqrt{\frac{1}{4}+\left(\frac{1}{2n}-1\right)\left(\Phi\left(-\sqrt{\frac{p}{2n}}\right)-\left[\Phi\left(-\sqrt{\frac{p}{2n}}\right)\right]^2\right)} \tag{4.73}$$

$$\text{RMS}[\hat{\epsilon}^l] \leq \kappa_{\hat{\epsilon}^l}(n,p,0.5) \approx \sqrt{\frac{1}{8n}+\Phi\left(-\sqrt{\frac{p}{8n^3}},-\sqrt{\frac{p}{8n^3}};\frac{1}{n}\right)-\frac{1}{8}} \tag{4.74}$$

Based upon the preceding comments, these will be very accurate in high dimensions and less so for small dimensinos.

It can be seen from (4.73) and (4.74) that the bound for leave-one-out is much less affected by dimensionality than the bound for resubstitution. This is because the terms involving p in (4.74) are functions of $\sqrt{p/n^3}$, whereas the corresponding terms in (4.73) are functions of $\sqrt{p/n}$. This difference in sensitivity to dimension between resubstitution and leave-one-out is not specific to the bound $\kappa_{\hat{\epsilon}}(n, p, 0.5)$ but holds in the whole range of δ^2. This phenomenon can be seen in Fig 16.

To find the necessary number of samples to insure a given RMS, one can find the minimum n to satisfy (4.73) and (4.74). Table III shows the minimum number of sample points needed for resubstitution and leave-one-out to achieve a given value of $\kappa_{\hat{\epsilon}}(n, p, 0.5)$. In this table we have considered different dimensions for resubstitution and only two dimensions for leave-one-out. The reason, as mentioned before, is that leave-one-out is much less affected by dimensionality. To test the applicability (robustness) of the expressions in (4.73) and (4.74), and the necessary sample sizes determined from these expressions, we have examined the effect of estimating the covariance matrix, defined in the definition of discriminant, on $\kappa_{\hat{\epsilon}}(n, p, 0.5)$, which has been obtained under the assumption of a known covariance matrix. This has been accomplished by using the required sample sizes in Table III in the Monte-Carlo estimation of $\kappa_{\hat{\epsilon}}(n, p, 0.5)$ when the covariance matrix is estimated from the data. The results are shown in Table III by the values in parentheses. Comparing these values with the given values of $\kappa_{\hat{\epsilon}}(n, p, 0.5)$ on the left-hand side of the table reveals that (4.73) and (4.74), and the sample sizes determined therefrom, can be reliably used in practice. A key observation regarding Table III is that the required sample size for resubstitution increases significantly as the dimension increases, whereas for leave-one-out the increase is slight, an observation consistent with the RMS peaking phenomenon seen in Fig. 16. As a final point, since the bounds are determined by $\epsilon_{bay} = 0.5$ and the finite sample RMS approximations are less accurate for $0.3 < \epsilon_{bay} < 0.5$ for low dimensions, in Table III we see that the accuracy of the results improves as the dimension increases.

Table III. Minimum sample size, n, $(n_0 = n_1 = n)$ for desired $\kappa_{\hat{\epsilon}}(n, p, 0.5)$. The values in parentheses are the Monte-Carlo estimation of $\kappa_{\hat{\epsilon}}(n, p, 0.5)$ when covariance matrix is estimated from data.

	resub					loo	
$\kappa_{\hat{\epsilon}}(n, p, 0.5)$	$p=2$	$p=3$	$p=4$	$p=6$	$p=10$	$p=3$	$p=10$
0.05	114 (0.043)	145 (0.045)	177 (0.045)	240 (0.047)	367 (0.048)	88 (0.054)	92 (0.048)
0.06	79 (0.051)	101 (0.053)	123 (0.054)	167 (0.056)	254 (0.058)	62 (0.065)	65 (0.056)
0.07	58 (0.060)	74 (0.062)	90 (0.063)	122 (0.066)	187 (0.067)	46 (0.076)	49 (0.065)
0.08	44 (0.069)	57 (0.070)	69 (0.073)	93 (0.075)	142 (0.078)	36 (0.083)	38 (0.074)
0.09	35 (0.076)	45 (0.080)	54 (0.083)	74 (0.085)	112 (0.088)	29 (0.091)	31 (0.083)
0.10	28 (0.087)	36 (0.090)	44 (0.092)	60 (0.095)	91 (0.098)	24 (0.101)	25 (0.091)

1. Implementation for Gene-expression Classification

In this section, we consider three-gene classification using the same gene-expression profiles used in chapter III. To have $\kappa_{\hat{\epsilon}^r}(n, p, 0.5) \approx 0.1$, we need $n_0 = n_1 = 36$. This sample size makes $\kappa_{\hat{\epsilon}^l}(n, p, 0.5) < 0.1$. Proceeding analogously to chapter III, using the 72 sample points selected, we applied the t-test to each gene to find significant differences between the good prognosis class and bad prognosis class. 53 of the 70 genes in the study had p-value less than 0.05. We chose the 9 genes showing the most significant differences among the two classes. Among these genes we picked the genes Contig28552_RC, NM_003981 and NM_020188, shown to be close to normal by the Shapiro-Wilk test and to have close to equal covariance matrices between classes by Box's M test. It is known that Box's M test performs well when the number of sample points in each class exceeds 9 and the dimension is less than 5 [134]. The significance level for all tests is 95%. The estimated errors using these three genes are $\hat{\epsilon}^r$ = 0.153 and $\hat{\epsilon}^l$ = 0.167, with hold-out giving a good approximation of the true error to be 0.164. Comparing the values of hold-out in these examples with those of the estimators themselves, we conclude that both resubstitution and leave-one-out have reasonably estimated the true error. Figure 17 shows the designed classifier. This example demonstrates how one can use Table III and combine it with the proper assumptions to get to a reliable estimation of the true error.

E. Conclusion

Using the double asymptotic method of Kolmogorov, we have derived double asymptotic (in sample size and dimension) representations for the second moments and cross-moments with the actual error for resubstitution and leave-one-out in a multivariate Gaussian model. From these, the bias, variance, and RMS for resubstitution and leave-one-out as estimators of the actual error can be computed. Such asymptotic results have historically been shown to provide good small sample approximations and this has been demonstrated in the present situation via numerical comparisons. As has generally been historically the case, the results for known covariance matrix have been obtained prior to those for unknown covariance matrix, the latter typically being

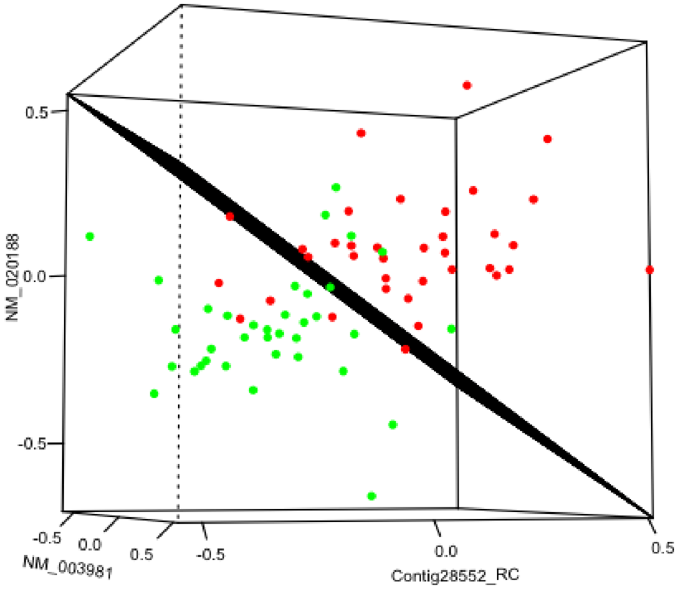

Fig. 17. The designed classifier for good-prognosis (green) vs. bad-prognosis (red) using the minimum number of samples to get a given RMS. The three genes selected are Contig28552_RC, NM_003981 and NM_020188.

significantly more difficult. Obtaining corresponding results with unknown covariance matrix is the next logical step in the line of this research.

REFERENCES

[1] M. Hills, “Allocation rules and their error rates,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 28, no. 1, pp. 1–31, 1966.

[2] B. Welch, “Note on discriminant functions,” *Biometrika*, vol. 31, pp. 218–220, 1939.

[3] P. Hoel and R. Peterson, “A solution to the problem of optimum allocation,” *Annals of Mathematical Statistics*, vol. 20, pp. 433–438, 1949.

[4] T. W. Anderson, *Introduction to Multivariate Statistical Analysis.* New York: Wiley, 1958.

[5] H. Raiffa, “Statistical decision theory approach to item selection for dichotomous test and criterion variables,” in *Studies in Item Analysis and Prediction*, H. Solomon, Ed. Palo Alto, CA: Stanford University Press, 1961, pp. 187–220.

[6] R. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[7] ——, “The precision of discriminant function,” *Annals of Eugenics*, vol. 10, pp. 422–429, 1940.

[8] B. C. Sutradhar, “Discrimination of observations into one of two t-populations,” *Biometrics*, vol. 46, pp. 827–835, 1990.

[9] R. K. Amoh and K. Kocherlakota, “Errors of misclassification associated with the inverse gaussian distribution,” *Communications in Statistics*, vol. 15, pp. 589–612, 1986.

[10] S. E. Khattabi and F. Streit, “Further results on identification when the parameters are partially unknown,” in *Statistical Data Analysis and Inference*, Y. Dodge, Ed. Amsterdam: North-Holland, 1989, pp. 347–352.

[11] A. Batsidis and K. Zografos, “Discrimination of observations into one of two elliptic populations based on monotone training samples,” *Metrika*, vol. 64, pp. 221–241, 2006.

[12] A. Azzalini and A. Capitanio, “Statistical application of the multivariate skew normal distribution,” *Journal of Royal Statistical Society B*, vol. 65, pp. 367–389, 1999.

[13] J. A. Anderson, “Logistic discrimination,” in *Handbook of Statistics*, P. R. Krishnaiah and N. K. L, Eds. Amsterdam: North-Holland, 1982, pp. 169–191.

[14] A. Albert and J. A. Anderson, “Probit and logistic discriminant functions,” *Communications in Statistics - Theory and Methods*, vol. 10, pp. 641–657, 1981.

[15] H. R. Warner, A. F. Toronto, L. G. Veasey, and R. Stephenson, “A mathematical approach to medical diagnosis,” *Journal of the American Medical Association*, vol. 177, pp. 177–183, 1961.

[16] S. E. Fienberg, *The Analysis of Cross-Classified Categorial Data.* Cambridge, MA: MIT, 1980.

[17] G. F. Hughes, “On the mean accuracy of statistical pattern recognizers,” *IEEE Transactions on Information Theory*, vol. 14, pp. 55–63, 1968.

[18] W. G. Cochran and C. E. Hopkins, “Some classification problems with multivariate qualitative data,” *Biometrics*, vol. 17, pp. 10–32, 1961.

[19] D. F. Specht, "Generation of polynomial discriminant functions for pattern recognition," *IEEE Transactions on Electronic Computers*, vol. 16, pp. 308–319, 1967.

[20] D. Hand, *Discrimination and Classification.* Chichester, UK: John Wiley, 1981.

[21] J. Aitchison and C. G. G. Aitken, "Multivariate binary discrimination by the kernel method," *Biometrika*, vol. 63, pp. 413–420, 1976.

[22] D. M. Titterington, "Analysis of incomplete multivariate binary data by the kernel method," *Biometrika*, vol. 64, pp. 455–460, 1977.

[23] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification.* New York: John Wiley, 2001.

[24] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition.* New York: Springer, 1996.

[25] A. Wald, "On a statistical problem arising in the classification of an individual into one of two groups," *Annals of Mathematical Statistics*, vol. 15, pp. 145–162, 1944.

[26] T. Anderson, "Classification by multivariate analysis," *Psychometrika*, vol. 16, pp. 31–50, 1951.

[27] S. van Vuuren and H. Hermansky, "Data-driven design of rasta-like filters," in *Eurospeech*, 1997.

[28] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, CA, 1992, pp. 13–16.

[29] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711–720, 1997.

[30] D. Swets and J. Weng, “Using discriminant eigenfeatures for image retrieval,” *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 18, pp. 891–896, 1996.

[31] S. Kim, E. R. Dougherty, I. Shmulevich, K. R. Hess, S. R. Hamilton, J. M. Trent, G. N. Fuller, and W. Zhang, “Identification of combination gene sets for glioma classification,” *Molecular Cancer Therapeutics*, vol. 1, pp. 1229–1236, 2002.

[32] H. Somura, N. Iizuka, T. Tamesa, K. Sakamoto, T. Hamaguchi, R. Tsunedomi, H. Yamada-Okabe, M. Sawamura, M. Eramoto, T. Miyamoto, Y. Hamamoto, and M. Oka, “A three-gene predictor for early intrahepatic recurrence of hepatocellular carcinoma after curative hepatectomy,” *Oncology Reports*, vol. 19, pp. 489–495, 2008.

[33] S. Raudys, “On determining training sample size of a linear classifier,” *Computer Systems*, vol. 28, pp. 79–87, 1967, in Russian.

[34] S. John, “Errors in discrimination,” *Annals of Mathematical Statistics*, vol. 32, pp. 1125–1144, 1961.

[35] M. Moran, “On the expectation of errors of allocation associated with a linear discriminant function,” *Biometrika*, vol. 62, no. 1, pp. 141–148, 1975.

[36] G. T. Toussaint, “Bibliography on estimation of misclassification,” *IEEE Transactions on Information Theory*, vol. 20, pp. 472–479, 1974.

[37] R. A. Schiavo and D. J. Hand, “Ten more years of error rate research,” *International Statistical Review*, vol. 68, no. 3, pp. 295–310, 2000.

[38] G. J. McLachlan, “The bias of the apparent error in discriminant analysis,” *Biometrika*, vol. 63, pp. 239–244, 1976.

[39] M. J. Sorum, “Estimating the conditional probability of misclassification,” *Technometrics*, vol. 13, pp. 333– 343, 1971.

[40] ——, “Estimating the expected and the optimal probabilities of misclassification,” *Technometrics*, vol. 14, pp. 935– 943, 1972.

[41] J. Kittler and P. DeVijver, “Statistical properties of error estimators in performance assessment of recognition systems,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 4, no. 2, pp. 215–220, 1982.

[42] K. Fukunaga and R. R. Hayes, “Estimation of classifier performance,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 1087–1101, 1989.

[43] M. Okamoto, “An asymptotic expansion for the distribution of the linear discriminant function,” *Annals of Mathematical Statistics*, vol. 34, pp. 1286–1301, 1963, Correction: *Annals of Mathematical Statistics*, vol. 39, pp. 1358–1359, 1968.

[44] G. J. McLachlan, “An asymptotic expansion of the expectation of the estimated error rate in discriminant analysis,” *Australian Journal of Statistics*, vol. 15, no. 3, pp. 210–214, 1973.

[45] ——, “Estimation of the errors of misclassification on the criterion of asymptotic mean square error,” *Technometrics*, vol. 16, no. 2, pp. 255–260, 1974.

[46] T. Anderson, “An asymptotic expansion of the distribution of the studentized classification statistic w,” *The Annals of Statistics*, vol. 1, pp. 964–972, 1973.

[47] B. Efron, “The efficiency of logistic regression compared to normal discriminant analysis,” *Journal of the American Statistical Association*, vol. 70, pp. 892–898, 1975.

[48] ——, “The distributions of the actual error rates in linear discriminant analysis,” *Journal of the American Statistical Association*, vol. 75, pp. 201–205, 1980.

[49] S. Raudys and A. K. Jain, “Small sample size effects in statistical pattern recognition: Recommendations for practitioners,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 252–264, 1991.

[50] S. Raudys and V. Pikelis, “On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 2, pp. 242–252, 1980.

[51] D. Foley, “Considerations of sample and feature size,” *IEEE Transactions on Information Theory*, vol. IT-18, pp. 618–626, 1972.

[52] R. A. Fisher, *Statistical Methods for Research Workers*, 14th ed. Edinburgh: Oliver & Boyd, 1925, the quotation is from the preface to the first (1925) edition.

[53] J. K. Martin and D. S. Hirschberg, “Small sample statistics for classification error rates II: Confidence intervals and significance tests,” University of California, Irvine, CA, Tech. Rep. 96-22, 1996.

[54] D. Hand, “Recent advances in error rate estimation,” *Pattern Recognition Letters*, vol. 4, pp. 335–346, 1986.

[55] C. H. S. Michiels and S. Koscielny, “Prediction of cancer outcome with microarrays: A multiple random validation strategy,” *The Lancet*, vol. 365, pp. 488–492, 2005.

[56] A. Dupuy and R. Simon, “Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting,” *Journal of the National Cancer Institute*, vol. 99, pp. 147–157, 2008.

[57] O. Gevaert, F. D. Smet, T. V. Gorp, N. Pochet, K. Engelen, F. Amant, B. D. Moor, D. Timmerman, and I. Vergote, “Expression profiling to predict the clinical behaviour of ovarian cancer fails independent evaluation,” *BMC Cancer*, vol. 8, pp. 1–10, 2008.

[58] E. R. Dougherty, J. Hua, and M. Bittner, “Validation of computational methods in genomics,” *Current Genomics*, vol. 8, pp. 1–19, 2007.

[59] C. Smith, “Some examples of discrimination,” *Annals of Eugenics*, vol. 18, pp. 272–282, 1947.

[60] P. Lachenbruch and M. Mickey, “Estimation of error rates in discriminant analysis,” *Technometrics*, vol. 10, pp. 1–11, 1968.

[61] T. Cover, “Learning in pattern recognition,” in *Methodologies of Pattern Recognition*, S. Watanabe, Ed. New York: Academic Press, 1969, pp. 111–132.

[62] G. Toussaint and R. Donaldson, “Algorithms for recognizing contour-traced hand-printed characters,” *IEEE Transactions on Computers*, vol. 19, pp. 541–546, 1970.

[63] M. Stone, “Cross-validatory choice and assessment of statistical predictions,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, pp.

111–147, 1974.

[64] T. Nagahata, M. Onda, M. Emi, H. Nagai, K. Tsumagari, and et al., "Expression profiling to predict postoperative prognosis for estrogen receptor-negative breast cancers by analysis of 25,344 genes on a cdna microarray," *Cancer Science*, vol. 95, pp. 218–225, 2004.

[65] S. Chiaretti, X. Li, R. Gentleman, A. Vitale, M. Vignetti, and et al., "Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival," *Blood*, vol. 103, no. 7, pp. 2771–2778, 2004.

[66] F. De Smet, N. Pochet, K. Engelen, T. Van Gorp, K. M. P. Van Hummelen, F. Amant, D. Timmerman, B. De Moor, and I. Vergote, "Predicting the clinical behavior of ovarian cancer from gene expression profiles," *International Journal of Gynecological Cancer*, vol. 16, pp. 147–151, 2006.

[67] H. Somura, N. Iizuka, T. Tamesa, K. Sakamoto, T. Hamaguchi, R. Tsunedomi, H. Yamada-Okabe, M. Sawamura, M. Eramoto, T. Miyamoto, Y. Hamamoto, and M. Oka, "A three-gene predictor for early intrahepatic recurrence of hepatocellular carcinoma after curative hepatectomy," *Oncology Reports*, vol. 19, no. 2, pp. 489–495, 2008.

[68] M. Shirahata, K. Iwao-Koizumi, S. Saito, N. Ueno, M. Oda, N. Hashimoto, J. Takahashi, and K. Kato, "Gene expression-based molecular diagnostic system for malignant gliomas is superior to histological diagnosis," *Clinical Cancer Research*, vol. 13, no. 24, pp. 7341–7356, 2007.

[69] C. Rimkus, J. Friederichs, A. Boulesteix, J. Theisen, J. Mages, K. Becker, H. Nekarda, R. Rosenberg, K. Janssen, and J. Siewert, "Microarray-based pre-

diction of tumor response to neoadjuvant radiochemotherapy of patients with locally advanced rectal cancer," *Clinical Gastroenterology and Hepatology*, vol. 6, no. 1, pp. 53–61, 2008.

[70] A. Wald, "Contributions to the theory of statistical estimation and testing hypotheses," *Annals of Mathematical Statistics*, vol. 10, pp. 299–326, 1945.

[71] R. V. Mises, "On the classification of observation data into distinct groups," *Annals of Mathematical Statistics*, vol. 16, pp. 68–73, 1945.

[72] C. R. Rao, "A general theory of discrimination when the information about alternative population distributions is based on samples," *Annals of Mathematical Statistics*, vol. 25, pp. 651–670, 1945.

[73] ——, "A classification problem in which information about alternative distributions is based on samples," *Annals of Mathematical Statistics*, vol. 13, pp. 213–223, 1962.

[74] T. W. Anderson and R. R. Bahadur, "Classification into two multivariate normal distributions with different covariance matrices," *Annals of Mathematical Statistics*, vol. 33, pp. 420–431, 1962.

[75] R. Sitgreaves, "On the distribution of two random matrices used in classification procedures," *Annals of Mathematical Statistics*, vol. 23, pp. 263–270, 1952.

[76] H. Harter, "On the distribution of Wald's classification statistics," *Annals of Mathematical Statistics*, vol. 22, pp. 58–67, 1951.

[77] S. Raudys and A. Jain, "The distribution of Wald's classification statistic when the dispersion matrix is known," *Sankhya*, vol. 21, pp. 371–376, 1991.

[78] S. John, “On some classification problems–I,” *Sankhya*, vol. 22, pp. 301–308, 1960.

[79] ——, “On some classification problems,” *Sankhya*, vol. 22, pp. 309–316, 1960.

[80] A. Bowker, “A representation of hotelling’s t^2 and anderson’s classification statistic w in terms of simple statistics,” in *Studies in Item Analysis and Prediction*, H. Solomon, Ed. Palo Alto, CA: Stanford University Press, 1961, pp. 285–292.

[81] R. Sitgreaves, “Some results on the distribution of the W-classification,” in *Studies in Item Analysis and Prediction*, H. Solomon, Ed. Palo Alto, CA: Stanford University Press, 1961, pp. 241–251.

[82] A. Jain and W. Waller, “On the optimal number of features in the classification of multivariate gaussian data,” *Pattern Recognition*, vol. 10, pp. 365–374, 1978.

[83] A. Bowker and R. Sitgreaves, “An asymptotic expansion for the distribution function of the w-classification statistic,” in *Studies in Item Analysis and Prediction*, H. Solomon, Ed. Palo Alto, CA: Stanford University Press, 1961, pp. 292–310.

[84] M. Okamoto, “Correction to: An asymptotic expansion for the discriminant function,” *Annals of Mathematical Statistics*, vol. 39, pp. 1358–1359, 1968.

[85] D. Kabe, “Some results on the distribution of two random matrices used in classification procedures,” *Annals of Mathematical Statistics*, vol. 34, pp. 181–185, 1963.

[86] S. D. Gupta, “Optimum classification rules for classification into two multivariate normal populations,” *Annals of Mathematical Statistics*, vol. 36, pp.

1174–1184, 1965.

[87] D. Teichroew and R. Sitgreaves, "Some operating characteristics of linear discriminant functions," in *Discriminant Analysis and Applications*, T. Cacoullos, Ed. New York: Academic Press, 1973, pp. 365–374.

[88] A. Z. Memon and M. Okamoto, "The classification statistic $w*$ in covariate discriminant analysis," *Annals of Mathematical Statistics*, vol. 41, pp. 1491–1499, 1970.

[89] W. G. Cochran, "Comparison of two methods of handling covariates in discriminatory analysis," *Annals of the Institute of Statistical Mathematics*, vol. 16, pp. 43–53, 1964.

[90] W. G. Cochran and C. I. Bliss, "Discriminant functions with covariance," *Annals of Mathematical Statistics*, vol. 19, pp. 151–176, 1948.

[91] G. J. McLachlan, "Asymptotic results for discriminant analysis when the initial samples are misclassified," *Technometrics*, vol. 14, no. 2, pp. 415–422, 1972.

[92] ——, "An asymptotic expansion for the variance of the errors of misclassification of the linear discriminant function," *Australian Journal of Statistics*, vol. 14, no. 3, pp. 68–72, 1972.

[93] M. J. Schervish, "Asymptotic expansions for the means and variances of error rates," *Biometrika*, vol. 68, pp. 295–299, 1981.

[94] N. Glick, "Asymptotic error rates of the w and z statistics when the training observations are dependent," *Pattern Recognition*, vol. 19, pp. 467–471, 1986.

[95] Y. Kharin, “The investigation of risk for statistical classifiers using minimum estimators,” *Theory of Probability and Its Applications*, vol. 28, pp. 623–630, 1984.

[96] V. L. Brailovskiy and A. L. Lunts, “The multiparameter recognition problem and its solution,” *Engineering Cybernetics*, vol. 1, pp. 13–22, 1964, in Russian.

[97] M. J. Sorum, “Estimating the expected probability of misclassification for a rule based on the linear discriminant function: Univariate normal case,” *Technometrics*, vol. 15, pp. 329–339, 1973.

[98] A. Davison and P. Hall, “On the bias and variability of bootstrap and cross-validation estimates of error rate in discrimination problems,” *Biometrika*, vol. 79, no. 2, pp. 279–284, 1992.

[99] S. Varma and R. Simon, “Bias in error estimation when using cross-validation for model selection,” *BMC Bioinformatics*, vol. 7:91, 2006.

[100] S. Wehberg, “A comparison of nonparametric error rate estimation methods in classification problems,” *Biometrical Journal*, vol. 46, no. 1, pp. 35–47, 2004.

[101] E. R. Dougherty, C. Sima, J. Hua, B. Hanczar, and U. Braga-Neto, “Performance of error estimators for classification,” *Current Bioinformatics*, vol. 5, no. 1, pp. 53–67, 2010.

[102] E. Dougherty and U. Braga-Neto, “Epistemology of computational biology: Mathematical models and experimental prediction as the basis of their validity,” *Journal of Biological Systems*, vol. 14, no. 1, pp. 65–90, 2006.

[103] Q. Xu, J. Hua, U. Braga-Neto, Z. Xiong, E. Suh, and E. Dougherty, “Confidence intervals for the true classification error conditioned on the estimated error,”

Technology in Cancer Research and Treatment, vol. 5, no. 6, pp. 579–590, 2006.

[104] C. Sima and E. Dougherty, “Bolstered error estimation provides superior feature-set ranking for small samples,” *Bioinformatics*, vol. 21, no. 7, pp. 1046–1054, 2005.

[105] B. Hanczar, J. Hua, and E. Dougherty, “Decorrelation of the true and estimated classifier errors in high-dimensional settings,” *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 347, 2007, article ID 38473.

[106] S. M. Weiss, “Small sample error rate estimation for k-nn classifiers,” *Theory of Probability and Its Applications*, vol. 13, pp. 285–289, 1991.

[107] S. Ganeshanandama and W. J. Krzanowskib, “Error-rate estimation in two-group discriminant analysis using the linear discriminant function,” *Journal of Statistical Computation and Simulation*, vol. 36, pp. 157–175, 1990.

[108] S. Snapinn and J. Knoke, “Classification error rate estimators evaluated by unconditional mean squared error,” *Technometrics*, vol. 26, pp. 371–378, 1984.

[109] ——, “An evaluation of smoothed classification error-rate estimators,” *Technometrics*, vol. 27, no. 2, pp. 199–206, 1985.

[110] D. Hand, “Common errors in data analysis: The apparent error rate of classification rules,” *Psychological Medicine*, vol. 13, pp. 201–203, 1983.

[111] J. T. Page, “Error-rate estimation in discriminant analysis,” *Technometrics*, vol. 27, pp. 189–198, 1985.

[112] N. Glick, “Additive estimators for probabilities of correct classification,” *Pattern Recognition*, vol. 10, pp. 211–222, 1978.

[113] V. L. Brailovskiy, "An object recognition algorithm with many parameters and its applications," *Engineering Cybernetics*, vol. 1, pp. 22–30, 1964, in Russian.

[114] K. Fukunaga and R. R. Hayes, "Effects of sample size in classifier design," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 873–885, 1989.

[115] S. Raudys, "Comparison of the estimates of the probability of misclassification," in *Proc. International Joint Conference on Pattern Recognition*, Kyoto, Japan, 1978, pp. 280–282.

[116] S. Raudys and D. M. Young, "Results in statistical discriminant analysis: A review of the former soviet union literature," *Journal of Multivariate Analysis*, vol. 89, pp. 1–35, 2004.

[117] A. Genz and F. Bretz, "Methods for the computation of multivariate t-probabilities," *Journal of Statistical Computation and Simulation*, vol. 11, no. 1, pp. 950–971, 2002.

[118] M. van de Vijver, Y. He, L. vant Veer, H. Dai, A. A. M. Hart, and et al., "A gene-expression signature as a predictor of survival in breast cancer," *The New England Journal of Medicine*, vol. 347, pp. 1999–2009, 2002.

[119] L. Devroye and T. Wagner, "Distribution-free inequalities for the deleted and hold-out error estimates," *IEEE Transactions on Information Theory*, vol. 25, pp. 202–207, 1979.

[120] ——, "Distribution-free performance bounds for potential function rules," *IEEE Transactions on Information Theory*, vol. 25, pp. 601–604, 1979.

[121] S. Raudys, “On the amount of a priori information in designing the classification algorithm,” *Technical Cybernetics*, vol. 4, pp. 168–174, 1972, in Russian.

[122] V. Serdobolskii, *Multivariate Statistical Analysis: A High-Dimensional Approach.* Dordrecht: Kluwer, 2000.

[123] F. Wyman, D. Young, and D. Turner, “A comparison of asymptotic error rate expansions for the sample linear discriminant function,” *Pattern Recognition*, vol. 23, no. 7, pp. 775–783, 1990.

[124] A. Deev, “Asymptotic expansions for distributions of statistics w, m, and w* in discriminant analysis,” *Statistical Methods of Classification*, vol. 31, pp. 6–57, 1972, in Russian.

[125] ——, “Representation of statistics of discriminant analysis and asymptotic expansion when space dimensions are comparable with sample size,” *Doklady Akademii Nauk SSSR*, vol. 195, pp. 759–762, 1970, in Russian.

[126] L. D. Meshalkin and V. I. Serdobolskii, “Errors in the classification of multivariate observations,” *Theory of Probability and its Applications*, vol. 23, pp. 741–750, 1978.

[127] S. Raudys, “On dimensionality, sample size and classification error of nonparametric linear classification algorithms,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 667–671, 1997.

[128] Y. Fujikoshi, “Error bounds for asymptotic approximations of the linear discriminant function when the sample sizes and dimensionality are large,” *Journal of Multivariate Analysis*, vol. 73, pp. 1–17, 2000.

[129] V. Pikelis, "Comparison of methods of computing the expected classification errors," *Automatic and Remote Control*, vol. 5, no. 7, pp. 59–63, 1976.

[130] S. Verdu and S. Shamai, "Spectral efficiency of cdma with random spreading," *IEEE Transaction on Information Theory*, vol. 45, pp. 622–640, 1999.

[131] D. Tse and S. Verdu, "Optimum asymptotic multiuser efficiency of randomly spread cdma," *IEEE Transaction on Information Theory*, vol. 46, pp. 2718–2722, 2000.

[132] E. P. Wigner, "On the distribution of the roots of certain symmetric matrices," *Annals of Mathematics*, vol. 67, pp. 325–327, 1958.

[133] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley, 1995.

[134] G. A. F. Seber, *Multivariate Observations*, 1st ed. New York: Wiley, 1984.

[135] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd ed. New York: Wiley, 1984.

[136] O. J. Dunn, "Some expected values for probabilities of correct classification in discriminant analysis," *Technometrics*, vol. 13, pp. 345–353, 1971.

[137] R. Kan, "From moments of sum to moments of product," *Journal of Multivariate Analysis*, vol. 99, pp. 542 – 554, 2008.

[138] P. Sen and J. Singer, *Large Sample Methods in Statistics.* New York: Chapman and Hall, 1993.

APPENDIX A

PROOFS IN CHAPTER II

Proof of Theorem 1.

We give the proof for the case $P\left(\hat{\varepsilon}_r = 0, \hat{\mu} \in (a,b), \hat{\mu}_0 > \hat{\mu}_1\right)$, the other cases being entirely similar. Note that the event corresponding to direction of classification, $\hat{\mu}_0 > \hat{\mu}_1$ in this case, how affect the different situations that corresponds to $\hat{\varepsilon}_r = 0$. From the expression for the univariate discriminant

$$W(x) \;=\; (x - \hat{\mu})\,(\hat{\mu}_0 - \hat{\mu}_1)$$

and noting the the definition of apparent error, it follows that

$$\begin{aligned}
&P(\hat{\varepsilon}_r = 0, \hat{\mu} \in (a,b), \hat{\mu}_0 > \hat{\mu}_1)\\
&= P(W(X_1) \geq 0, \ldots, W(X_{n_0}) \geq 0, W(X_{n_0+1}) < 0, \ldots, W(X_{n_0+n_1}) < 0, \hat{\mu} \in (a,b), \hat{\mu}_0 > \hat{\mu}_1)\\
&= P\left(W(X_1) > 0, \ldots, W(X_{n_0}) > 0, W(X_{n_0+1}) < 0, \ldots, W(X_{n_0+n_1}) < 0, \hat{\mu} > a, -\hat{\mu} > -b, \hat{\mu}_0 - \hat{\mu}_1 > 0\right)\\
&= P(X_1 - \hat{\mu} > 0, \ldots, X_{n_0} - \hat{\mu} > 0, \hat{\mu} - X_{n_0+1} > 0, \ldots, \hat{\mu} - X_{n_0+n_1} > 0, \hat{\mu}_0 - \hat{\mu}_1 > 0, \hat{\mu} > a, -\hat{\mu} > -b, \hat{\mu}_0 - \hat{\mu}_1 > 0)\\
&+ P(X_1 - \hat{\mu} < 0, \ldots, X_{n_0} - \hat{\mu} < 0, \hat{\mu} - X_{n_0+1} < 0, \ldots, \hat{\mu} - X_{n_0+n_1} < 0, \hat{\mu}_0 - \hat{\mu}_1 < 0, \hat{\mu} > a, -\hat{\mu} > -b, \hat{\mu}_0 - \hat{\mu}_1 > 0)\\
&= P(Z_1 > 0)
\end{aligned}$$

since $P(\ldots, \hat{\mu}_0 - \hat{\mu}_1 < 0, \ldots, \hat{\mu}_0 - \hat{\mu}_1 > 0) = 0$, with the vector Z_1 being given by:

$$Z_1 \;=\; \left[2(X_1 - \hat{\mu}), \ldots, 2(X_{n_0} - \hat{\mu}), 2(\hat{\mu} - X_{n_0+1}), \ldots, 2(\hat{\mu} - X_{n_0+n_1}), \hat{\mu}_0 - \hat{\mu}_1, \hat{\mu} - a, -\hat{\mu} + b\right]^T$$

Vector Z_1 is a linear combination of the vector of observations $X = [X_1, \ldots, X_{n_0+n_1}]$, namely, $Z_1 = AX - \mathbf{c}$, where $\mathbf{c}$ is determined as follows:

$$\mathbf{c} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 2a \\ -2b \end{pmatrix}_{(n_0+n_1+3)\times 1} \tag{A.1}$$

matrix A is a function of n_0 and n_1, a and b determined as follows:

$$A = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix}_{(n_0+n_1+3)\times(n_0+n_1)} \tag{A.2}$$

where

$$A_1 = \begin{pmatrix} \left(2-\frac{1}{n_0}\right) & -\frac{1}{n_0} & \dots & -\frac{1}{n_0} & -\frac{1}{n_1} & \dots & -\frac{1}{n_1} \\ -\frac{1}{n_0} & \left(2-\frac{1}{n_0}\right) & \dots & -\frac{1}{n_0} & -\frac{1}{n_1} & \dots & -\frac{1}{n_1} \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \\ -\frac{1}{n_0} & -\frac{1}{n_0} & \dots & \left(2-\frac{1}{n_0}\right) & -\frac{1}{n_1} & \dots & -\frac{1}{n_1} \\ & & & & & & \end{pmatrix}_{n_0\times(n_0+n_1)} \tag{A.3}$$

$$A_2 = \begin{pmatrix} \frac{1}{n_0} & \dots & \frac{1}{n_0} & \left(\frac{1}{n_1}-2\right) & \frac{1}{n_1} & \dots & \frac{1}{n_1} \\ \frac{1}{n_0} & \dots & \frac{1}{n_0} & \frac{1}{n_1} & \left(\frac{1}{n_1}-2\right) & \dots & \frac{1}{n_1} \\ & \vdots & & \vdots & \vdots & \ddots & \\ \frac{1}{n_0} & \dots & \frac{1}{n_0} & \frac{1}{n_1} & \frac{1}{n_1} & \dots & \left(\frac{1}{n_1}-2\right) \\ & & & & & & \end{pmatrix}_{n_1\times(n_0+n_1)} \tag{A.4}$$

$$A_3 = \begin{pmatrix} \frac{1}{n_0} & \cdots & \frac{1}{n_0} & -\frac{1}{n_1} & \cdots & -\frac{1}{n_1} \\ \frac{1}{n_0} & \cdots & \frac{1}{n_0} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \\ \frac{-1}{n_0} & \cdots & \frac{-1}{n_0} & \frac{-1}{n_1} & \cdots & \frac{-1}{n_1} \\ & & & & & \end{pmatrix}_{3\times(n_0+n_1)} . \tag{A.5}$$

Therefore, Z is a Gaussian random vector, with mean $\mu_Z = A\mu_X - \mathbf{c}$ and covariance $\Sigma_Z = A\Sigma_X A^T$. Substituting the values of $\mu_X = [\mu_0 \mathbf{1}_{n_0}, \mu_1 \mathbf{1}_{n_1}]^T$ and $\Sigma_X = \text{diag}(\mathbf{1}_{n_0}, \mathbf{1}_{n_1})$ results in (2.10) and (2.11). □

Proof of Theorem 5.

We give the proof for the case $\hat{\varepsilon}_r = 0$. The case $\hat{\varepsilon}_r > 0$ is obtained by using the same argument employed in connection with Theorems 2, 4, and 7. From Theorem 3 and the proof of Theorem 1, we observe that

$$P\left(\hat{\varepsilon}_r = 0, \hat{\mu} > a, \hat{\mu}_0 > \hat{\mu}_1\right) = P(Z > \mathbf{0}) \tag{A.6}$$

where

$$Z = \left[X_1 - \hat{\mu}, \ldots, X_{n_0} - \hat{\mu}_0, \hat{\mu} - X_{n_0+1}, \ldots, \hat{\mu} - X_{n_0+n_1}, \hat{\mu}_0 - \hat{\mu}_1, 2(\hat{\mu} - a)\right] \tag{A.7}$$

is a Gaussian random vector of size $n_0 + n_1 + 2$, with mean μ_Z given by:

$$\mu_Z = \begin{bmatrix} (\mu_0 - \mu_1)\mathbf{1}_{n_0+n_1+1} \\ \\ (\mu_0 + \mu_1) - 2a \end{bmatrix} \tag{A.8}$$

and covariance matrix Σ_Z given by

$$(\Sigma_Z)_{ij} = \begin{cases} (4n_0 - 3)\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & i, j = 1, \ldots, n_0, i = j \\ -3\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & i, j = 1, \ldots, n_0, i \neq j \\ \frac{\sigma_0^2}{n_0} + (4n_1 - 3)\frac{\sigma_1^2}{n_1}, & i, j = n_0 + 1, \ldots, n_0 + n_1, i = j \\ \frac{\sigma_0^2}{n_0} - 3\frac{\sigma_1^2}{n_1}, & i, j = n_0 + 1, \ldots, n_0 + n_1, i \neq j \\ \frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1}, & \begin{cases} i = n_0 + n_1 + 2, j = 1, \ldots, n_0 + n_1 + 1 \\ j = n_0 + n_1 + 2, i = 1, \ldots, n_0 + n_1 + 1 \end{cases}, \\ \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & \text{otherwise} \end{cases} \tag{A.9}$$

Let $Z = [Y, W]$, where Y is the vector containing the first n_0+n_1+1 components of Z, and $W = 2(\hat{\mu} - a)$. Note that

$$\begin{aligned} p\left(\hat{\varepsilon}_r = 0, \hat{\mu} = a, \hat{\mu}_0 > \hat{\mu}_1\right) &= P\left(\hat{\varepsilon}_r = 0, \hat{\mu}_0 > \hat{\mu}_1 \,|\, \hat{\mu} = a\right) p(\hat{\mu} = a) \\ &= P(Y > \mathbf{0} \,|\, \hat{\mu} = a) \, p(\hat{\mu} = a) \\ &= P(Y > \mathbf{0} \,|\, W = 0) \, p(\hat{\mu} = a) \end{aligned} \tag{A.10}$$

Now, it is a well-known fact (e.g. see Theorem 2.5.1 in [135]) that the distribution of vector Y given W is again Gaussian, with mean $\mu_Y - \frac{\mu_W}{\sigma_W^2}\Sigma_{YW}$, and covariance matrix $\Sigma_Y - \frac{1}{\sigma_W^2}\Sigma_{YW}\Sigma_{YW}^T$. In addition, $p(\hat{\mu} = a)$ is a Gaussian density with mean $\frac{\mu_0+\mu_1}{2}$ and variance $\frac{1}{4}(\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1})$. The computation of $p\left(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = a, \hat{\mu}_0 < \hat{\mu}_1\right)$ is entirely similar. □

Proof of Theorem 6.

We give the proof for the case $P\left(\hat{\varepsilon}_l = 0, \hat{\mu} \in (a, b), \hat{\mu}_0 > \hat{\mu}_1\right)$, the other cases being entirely similar. The univariate discriminant where the i-th sample is left out is given by

$$W^{(i)}(x) = (x - \hat{\mu}^{(i)})\,\hat{\nu}^{(i)}$$

where $\hat{\mu}^{(i)}$ and $\hat{\nu}^{(i)}$ are the average and difference, respectively, of sample means when the i-th sample is left out. Let us define the event intersection of the events $\mathbf{A} = \{\hat{\mu} - a > 0\} \cap \{-\hat{\mu} + b > 0\} \cap \{\hat{\mu}_0 - \hat{\mu}_1 > 0\}$. We have that:

$$P(W^{(1)}(X_1) \geq 0,\ldots, W^{(n_0)}(X_{n_0}) \geq 0, W^{(n_0+1)}(X_{n_0+1}) < 0,\ldots, W^{(n_0+n_1)}(X_{n_0+n_1}) < 0, \mathbf{A})$$

$$= P\big(X_1 - \hat{\mu}^{(1)} \geq 0,\, \hat{\nu}^{(1)} \geq 0, X_2 - \hat{\mu}^{(2)} \geq 0,\, \hat{\nu}^{(2)} \geq 0,\, \ldots,\, X_{n_0} - \hat{\mu}^{(n_0)} \geq 0,\, \hat{\nu}^{(n_0)} \geq 0,$$

$$\hat{\mu}^{(n_0+1)} - X_{n_0+1} \geq 0,\, \hat{\nu}^{(n_0+1)} \geq 0,\, \ldots,\, \hat{\mu}^{(n_0+n_1)} - X_{n_0+n_1} \geq 0,\, \hat{\nu}^{(n_0+n_1)} \geq 0, \mathbf{A}\big)$$

$$+ P\big(X_1 - \hat{\mu}^{(1)} < 0,\, \hat{\nu}^{(1)} < 0, X_2 - \hat{\mu}^{(2)} \geq 0,\, \hat{\nu}^{(2)} \geq 0,\, \ldots,\, X_{n_0} - \hat{\mu}^{(n_0)} \geq 0,\, \hat{\nu}^{(n_0)} \geq 0,$$

$$\hat{\mu}^{(n_0+1)} - X_{n_0+1} \geq 0,\, \hat{\nu}^{(n_0+1)} \geq 0,\, \ldots,\, \hat{\mu}^{(n_0+n_1)} - X_{n_0+n_1} \geq 0,\, \hat{\nu}^{(n_0+n_1)} \geq 0, \mathbf{A}\big)$$

$$\vdots$$

$$+ P\big(X_1 - \hat{\mu}^{(1)} < 0,\, \hat{\nu}^{(1)} < 0, X_2 - \hat{\mu}^{(2)} < 0,\, \hat{\nu}^{(2)} < 0,\, \ldots,\, X_{n_0} - \hat{\mu}^{(n_0)} < 0,\, \hat{\nu}^{(n_0)} < 0,$$

$$\hat{\mu}^{(n_0+1)} - X_{n_0+1} < 0,\, \hat{\nu}^{(n_0+1)} < 0,\, \ldots,\, \hat{\mu}^{(n_0+n_1)} - X_{n_0+n_1} < 0,\, \hat{\nu}^{(n_0+n_1)} < 0, \mathbf{A}\big)$$

where in fact the total number of joint probabilities that should be computed is $2^{n_0}2^{n_1}$. Simplification by grouping repeated probabilities results in:

$$P\left(\hat{\varepsilon}_l = 0\right) = \sum_{m=0}^{n_0} \sum_{n=0}^{n_1} \binom{n_0}{m} \binom{n_1}{n} P(Z_{1,m,n} \geq 0)$$

where $Z_{1,m,n} = E_{m,n} Z_1$ in which matrix $Z_1 = AX - \mathbf{c}$ where $\mathbf{c}$ is determined as follows:

$$\mathbf{c} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 2a \\ -2b \end{pmatrix}_{(2n_0+2n_1+3)\times 1} \tag{A.11}$$

and $X = [X_1, \dots, X_{n_0+n_1}]$ and A is:

$$A = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \end{pmatrix}_{(2n_0+2n_1+3)\times(n_0+n_1)} \tag{A.12}$$

where

$$A_1 = \begin{pmatrix} 2(1-\frac{1}{n_0}) & -\frac{1}{n_0} & \dots & -\frac{1}{n_0} & -(\frac{1}{n_1}-\frac{1}{n_0 n_1}) & \dots & -(\frac{1}{n_1}-\frac{1}{n_0 n_1}) \\ -\frac{1}{n_0} & 2(1-\frac{1}{n_0}) & \dots & -\frac{1}{n_0} & -(\frac{1}{n_1}-\frac{1}{n_0 n_1}) & \dots & -(\frac{1}{n_1}-\frac{1}{n_0 n_1}) \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \\ -\frac{1}{n_0} & -\frac{1}{n_0} & \dots & 2(1-\frac{1}{n_0}) & -(\frac{1}{n_1}-\frac{1}{n_0 n_1}) & \dots & -(\frac{1}{n_1}-\frac{1}{n_0 n_1}) \end{pmatrix}_{n_0\times(n_0+n_1)} \tag{A.13}$$

$$A_2 = \begin{pmatrix} 0 & \frac{1}{n_0} & \dots & \frac{1}{n_0} & -(\frac{1}{n_1}-\frac{1}{n_0 n_1}) & -(\frac{1}{n_1}-\frac{1}{n_0 n_1}) & \dots & -(\frac{1}{n_1}-\frac{1}{n_0 n_1}) \\ \frac{1}{n_0} & 0 & \dots & \frac{1}{n_0} & -(\frac{1}{n_1}-\frac{1}{n_0 n_1}) & -(\frac{1}{n_1}-\frac{1}{n_0 n_1}) & \dots & -(\frac{1}{n_1}-\frac{1}{n_0 n_1}) \\ & \vdots & \ddots & \vdots & \vdots & & \vdots & \\ \frac{1}{n_0} & \dots & \frac{1}{n_0} & 0 & -(\frac{1}{n_1}-\frac{1}{n_0 n_1}) & -(\frac{1}{n_1}-\frac{1}{n_0 n_1}) & \dots & -(\frac{1}{n_1}-\frac{1}{n_0 n_1}) \end{pmatrix}_{n_0\times(n_0+n_1)} \tag{A.14}$$

$$A_3 = \begin{pmatrix} (\frac{1}{n_0} - \frac{1}{n_0 n_1}) & \dots & (\frac{1}{n_0} - \frac{1}{n_0 n_1}) & -2(1 - \frac{1}{n_1}) & \frac{1}{n_1} & \dots & \frac{1}{n_1} \\ (\frac{1}{n_0} - \frac{1}{n_0 n_1}) & \dots & (\frac{1}{n_0} - \frac{1}{n_0 n_1}) & \frac{1}{n_1} & -2(1 - \frac{1}{n_1}) & \dots & \frac{1}{n_1} \\ & \vdots & & \vdots & \vdots & \ddots & \\ (\frac{1}{n_0} - \frac{1}{n_0 n_1}) & \dots & (\frac{1}{n_0} - \frac{1}{n_0 n_1}) & \frac{1}{n_1} & \dots & \frac{1}{n_1} & -2(1 - \frac{1}{n_1}) \\ \end{pmatrix}_{n_1 \times (n_0 + n_1)} \tag{A.15}$$

$$A_4 = \begin{pmatrix} (\frac{1}{n_0} - \frac{1}{n_0 n_1}) & \dots & (\frac{1}{n_0} - \frac{1}{n_0 n_1}) & 0 & -\frac{1}{n_1} & -\frac{1}{n_1} & \dots & -\frac{1}{n_1} \\ (\frac{1}{n_0} - \frac{1}{n_0 n_1}) & \dots & (\frac{1}{n_0} - \frac{1}{n_0 n_1}) & -\frac{1}{n_1} & 0 & -\frac{1}{n_1} & \dots & -\frac{1}{n_1} \\ & \vdots & & \vdots & \vdots & \ddots & & \\ (\frac{1}{n_0} - \frac{1}{n_0 n_1}) & \dots & (\frac{1}{n_0} - \frac{1}{n_0 n_1}) & -\frac{1}{n_1} & -\frac{1}{n_1} & \dots & -\frac{1}{n_1} & 0 \\ \end{pmatrix}_{n_1 \times (n_0 + n_1)} \tag{A.16}$$

$$A_5 = \begin{pmatrix} \frac{1}{n_0} & \dots & \frac{1}{n_0} & -\frac{1}{n_1} & \dots & -\frac{1}{n_1} \\ \frac{1}{n_0} & \dots & \frac{1}{n_0} & \frac{1}{n_1} & \dots & \frac{1}{n_1} \\ \frac{-1}{n_0} & \dots & \frac{-1}{n_0} & \frac{-1}{n_1} & \dots & \frac{-1}{n_1} \end{pmatrix}_{3 \times (n_0 + n_1)} \tag{A.17}$$

Therefore, Z_1 is a Gaussian random vector, with mean $\mu_{Z_1} = A\mu_X - \mathbf{c}$ and covariance $\Sigma_Z = A\Sigma_X A^T$. Substituting the values of $\mu_X = [\mu_0 \mathbf{1}_{n_0}, \mu_1 \mathbf{1}_{n_1}]^T$ and $\Sigma_X = \text{diag}(\mathbf{1}_{n_0}, \mathbf{1}_{n_1})$ results in the values of μ_{Z_1} and Σ_{Z_1} stated in the Theorem. □

APPENDIX B

PROOFS IN CHAPTER III

Proof of Theorem 9

Using the fact that the univariate discriminant is given by:

$$W(X) = (X - \bar{X})\left(\bar{X}_0 - \bar{X}_1\right) \tag{B.1}$$

it follows that we have:

$$P(W(\bar{X}_0, \bar{X}_1, X) \leq 0 \mid X \in \Pi_0) = P(X - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 > 0) + P(X - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0)$$

where $\bar{X} = \frac{\bar{X}_0 + \bar{X}_1}{2}$. Expanding $\bar{X}_0$ and $\bar{X}_1$ by $\frac{1}{n_0}\sum_{i=1}^{n_0} X_i$ and $\bar{X}_1 = \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} X_i$, respectively results in:

$$P(W(\bar{X}_0, \bar{X}_1, X) \leq 0 \mid X \in \Pi_0) = P(Z^{\mathrm{I}} < 0) + P(Z^{\mathrm{I}} \geq 0)$$

where vector Z^{I} is $Z^{\mathrm{I}} = AY$ in which $Y = [X, X_1, \ldots, X_{n_0}, X_{n_0+1}, \ldots, X_{n_0+n_1}]^T$ and

$$A = \begin{pmatrix} 1 & -\frac{1}{2n_0} & -\frac{1}{2n_0} & \cdots & -\frac{1}{2n_0} & -\frac{1}{2n_1} & \cdots & -\frac{1}{2n_1} \\ 0 & -\frac{1}{n_0} & -\frac{1}{n_0} & \cdots & -\frac{1}{n_0} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \end{pmatrix}$$

Therefore, Z^{I} is a gaussian random vector with mean $A\mu_Y$ and covariance $A\Sigma_Y A^T$. Substituting the values of $\mu_Y = [\mu_0\mathbf{1}_{n_0+1}, \mu_1\mathbf{1}_{n_1}]^T$ and $\Sigma_Y = \mathrm{diag}(\sigma_0^2\mathbf{1}_{n_0+1}, \sigma_1^2\mathbf{1}_{n_1})$ reduces to the expression stated in Theorem 9. Evaluating the mean and covariance matrix of vector Z^{II} stated in the theorem is entirely similar by considering $P(W(\bar{X}_0, \bar{X}_1, X) > 0 \mid X \in \Pi_1, \bar{X}_0, \bar{X}_1)$.

Proof of Theorem 10

We try to expand the first term in (3.8). Other terms are very similar. Using the univariate representation of classifier in (B.1), we have:

$$
\begin{gathered}
P(W(\bar{X}_0, \bar{X}_1, X) \leq 0, W(\bar{X}_0, \bar{X}_1, X') \leq 0 \mid X, X' \in \Pi_0) = \\
P(X - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0, X' - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0) + \\
P(X - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 > 0, X' - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \geq 0) + \\
P(X - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0, X' - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \geq 0) + \\
P(X - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \geq 0, X' - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0)
\end{gathered}
$$

Expanding $\bar{X}_0$ and $\bar{X}_1$ by $\frac{1}{n_0}\sum_{i=1}^{n_0} X_i$ and $\bar{X}_1 = \frac{1}{n_1}\sum_{i=n_0+1}^{n_0+n_1} X_i$, respectively results in

$$
\begin{gathered}
P(W(\bar{X}_0, \bar{X}_1, X) \leq 0, W(\bar{X}_0, \bar{X}_1, X') \leq 0 \mid X, X' \in \Pi_0) = \\
P(Z_0^{\mathrm{I}} < 0) + P(Z_0^{\mathrm{I}} \geq 0) + P(Z_1^{\mathrm{I}} < 0) + P(Z_1^{\mathrm{I}} \geq 0)
\end{gathered}
$$

where $Z_0^{\mathrm{I}} = A_0 Y$ and $Z_1^{\mathrm{I}} = A_1 Y$ in which $Y = [X, X', X_1, \ldots, X_{n_0}, X_{n_0+1}, \ldots, X_{n_0+n_1}]^T$ and

$$
A_0 = \begin{pmatrix}
1 & 0 & -\frac{1}{2n_0} & -\frac{1}{2n_0} & \cdots & -\frac{1}{2n_0} & -\frac{1}{2n_1} & \cdots & -\frac{1}{2n_1} \\
0 & 0 & -\frac{1}{n_0} & -\frac{1}{n_0} & \cdots & -\frac{1}{n_0} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \\
0 & 1 & -\frac{1}{2n_0} & -\frac{1}{2n_0} & \cdots & -\frac{1}{2n_0} & -\frac{1}{2n_1} & \cdots & -\frac{1}{2n_1} \\
0 & 0 & -\frac{1}{n_0} & -\frac{1}{n_0} & \cdots & -\frac{1}{n_0} & \frac{1}{n_1} & \cdots & \frac{1}{n_1}
\end{pmatrix}
$$

$$
A_1 = \begin{pmatrix}
1 & 0 & -\frac{1}{2n_0} & -\frac{1}{2n_0} & \cdots & -\frac{1}{2n_0} & -\frac{1}{2n_1} & \cdots & -\frac{1}{2n_1} \\
0 & 0 & -\frac{1}{n_0} & -\frac{1}{n_0} & \cdots & -\frac{1}{n_0} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \\
0 & -1 & \frac{1}{2n_0} & \frac{1}{2n_0} & \cdots & \frac{1}{2n_0} & \frac{1}{2n_1} & \cdots & \frac{1}{2n_1} \\
0 & 0 & \frac{1}{n_0} & \frac{1}{n_0} & \cdots & \frac{1}{n_0} & -\frac{1}{n_1} & \cdots & -\frac{1}{n_1}
\end{pmatrix}
$$

Therefore, Z_0^{I} and Z_1^{I} are gaussian random vectors with means $A_0\mu_Y$ and $A_1\mu_Y$ and covariance matrices $A_0\Sigma_Y A_0^T$ and $A_1\Sigma_Y A_1^T$, respectively. Substituting the values of $\mu_Y = [\mu_0 \mathbf{1}_{n_0+2}, \mu_1 \mathbf{1}_{n_1}]^T$ and $\Sigma_Y = \text{diag}(\sigma_0^2 \mathbf{1}_{n_0+2}, \sigma_1^2 \mathbf{1}_{n_1})$ reduces to the expression stated in Theorem 10. Evaluating the means and covariance matrices of $Z_i^{\text{II}} < 0$ and $Z_i^{\text{II}} < 0$, $i = 0, 1$ stated in the theorem is entirely similar by considering the corresponding terms in (3.8).

Proof of Theorem 13

We try to expand the first term in (3.13). Other terms are very similar. Using the univariate representation of classifier in (B.1), we have:

$$\begin{gathered}
P(W(\bar{X}_0, \bar{X}_1, X) \le 0, W(\bar{X}_0, \bar{X}_1, X_1) \le 0 \mid X \in \Pi_0) = \\
P(X - \bar{X} \ge 0, \bar{X}_0 - \bar{X}_1 < 0, X_1 - \bar{X} \ge 0, \bar{X}_0 - \bar{X}_1 < 0) + \\
P(X - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 > 0, X_1 - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \ge 0) + \\
P(X - \bar{X} \ge 0, \bar{X}_0 - \bar{X}_1 < 0, X_1 - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \ge 0) + \\
P(X - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \ge 0, X_1 - \bar{X} \ge 0, \bar{X}_0 - \bar{X}_1 < 0)
\end{gathered}$$

Expanding $\bar{X}_0$ and $\bar{X}_1$ by $\frac{1}{n_0} \sum_{i=1}^{n_0} X_i$ and $\bar{X}_1 = \frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1}$, respectively results in the gaussian vectors Z_0^I and Z_1^I with the means and covariance matrices stated in Theorem 13.

APPENDIX C

PROOFS IN CHAPTER IV

Proof of Theorem 16

Since the classification error ϵ is invariant to any linear transformation, we can use the canonical convenient form proposed by [136], with $\Sigma = I$ and $\mu_1 = -\mu_0 = (\frac{\delta}{2}, 0, \ldots, 0)^T$.

We prove that $\mathrm{Var}(\hat{G}_0) \overset{K}{\rightarrow} 0$. Let $V(i)$ denote the i-th component of vector V. We have

$$\mathrm{Var}(\hat{G}_0) = \mathrm{Var}(E[W(\bar{X}_0, \bar{X}_1, X) | \bar{X}_0, \bar{X}_1, X \in \Pi_0]) = \mathrm{Var}\left(\left(-\frac{\delta}{2} - \bar{X}(1)\right)\bar{a}(1) - \sum_{i=2}^{p} \bar{X}(i)a(i)\right) \tag{C.1}$$

where $\bar{X} = \frac{\bar{X}_0 + \bar{X}_1}{2}$ and $\bar{a} = \bar{X}_0 - \bar{X}_1$ are Gaussian vectors, with mean vectors and covariance

$$\mu_{\bar{X}} = (0, \ldots, 0) \ , \mu_{\bar{a}} = (-\delta, 0, \ldots, 0) \ , \Sigma_{\bar{X}} = \frac{1}{4}\left(\frac{1}{n_0} + \frac{1}{n_1}\right) I_p \ , \Sigma_{\bar{a}} = \left(\frac{1}{n_0} + \frac{1}{n_1}\right) I_p \tag{C.2}$$

Given the independence of the vector components, and using the results of [137] to find the variance of a product of non central gaussian vectors, algebraic manipulation leads to:

$$\mathrm{Var}(\hat{G}_0) = \frac{\delta^2}{n_1} + \frac{p}{2}\left(\frac{1}{n_0^2} + \frac{1}{n_1^2}\right) \overset{K}{\rightarrow} 0 \tag{C.3}$$

as desired. By a simple application of Chebyshev's inequality, it follows that

$$\begin{aligned} \underset{n_0, n_1, p \to \infty}{\mathrm{pklim}} \hat{G}_0 &= \underset{n_0, n_1, p \to \infty}{\mathrm{klim}} E[\hat{G}_0] = \underset{n_0, n_1, p \to \infty}{\mathrm{klim}} E[W(\bar{X}_0, \bar{X}_1, X) \mid X \in \Pi_0] \\ &= \underset{n_0, n_1, p \to \infty}{\mathrm{klim}} \left[\frac{\delta^2}{2} + \frac{p}{2}\left(\frac{1}{n_1} - \frac{1}{n_0}\right)\right] = \frac{1}{2}(\delta^2 + \lambda_1 - \lambda_0) \overset{\Delta}{=} G_0 \end{aligned} \tag{C.4}$$

An analogous argument shows that $\mathrm{Var}(\hat{G}_1)\overset{K}{\to}0$ and

$$\underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ \hat{G}_1 = \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\hat{G}_1] = \frac{1}{2}(\delta^2+\lambda_0-\lambda_1)\overset{\Delta}{=}G_1 \tag{C.5}$$

Now we prove that $\mathrm{Var}(\hat{D}_0)\overset{K}{\to}0$. We have

$$\hat{D}_0 = \mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X)\mid \bar{X}_0,\bar{X}_1,X\in\Pi_0) = \bar{a}^T\Sigma_X\bar{a} = \bar{a}^T\bar{a} = \hat{\delta}^2 \tag{C.6}$$

since $\Sigma_X = I_p$, where $\bar{a}$ is defined as before and $\hat{\delta}^2 = \left(\bar{X}_0-\bar{X}_1\right)^T\left(\bar{X}_0-\bar{X}_1\right)$. Notice that

$$\frac{\hat{\delta}^2}{\frac{1}{n_0}+\frac{1}{n_1}} \sim \chi_1^2\left(\frac{\delta^2}{\frac{1}{n_0}+\frac{1}{n_1}}\right)+\chi_{p-1}^2 \tag{C.7}$$

i.e., the sum of a noncentral and a central independent chi-square random variable, with the noncentrality parameter and degrees of freedom indicated. It follows that

$$\mathrm{Var}(\hat{D}_0)=\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^2\left[\mathrm{Var}\left(\chi_1^2\left(\frac{\delta^2}{\frac{1}{n_0}+\frac{1}{n_1}}\right)\right)+\mathrm{Var}(\chi_{p-1}^2)\right]=4\delta^2\left(\frac{1}{n_0}+\frac{1}{n_1}\right)+2p\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^2\overset{K}{\to}0 \tag{C.8}$$

as desired. By a simple application of Chebyshev's inequality, it follows that

$$\begin{aligned}\underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ \hat{D}_0 &= \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\hat{D}_0] = \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X)\mid \bar{X}_0,\bar{X}_1,X\in\Pi_0)]\\ &= \underset{n_0,n_1,p\to\infty}{\mathrm{klim}}\left[\delta^2+p\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\right] = \delta^2+\lambda_0+\lambda_1\overset{\Delta}{=}D\end{aligned} \tag{C.9}$$

An analogous argument shows that $\mathrm{Var}(\hat{D}_1)\overset{K}{\to}0$ and

$$\underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ \hat{D}_1 = \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\hat{D}_1] = \delta^2+\lambda_0+\lambda_1 = D \tag{C.10}$$

By using the Continuous Mapping Theorem [133], it follows that

$$
\begin{aligned}
\operatorname*{pklim}_{n_0,n_1,p\to\infty} \epsilon_0 &= \operatorname*{pklim}_{n_0,n_1,p\to\infty} \Phi\left(-\frac{\hat{G}_0}{\sqrt{\hat{D}_0}}\right) = \Phi\left(\operatorname*{pklim}_{n_0,n_1,p\to\infty} -\frac{\hat{G}_0}{\sqrt{\hat{D}_0}}\right) = \Phi\left(-\frac{G_0}{\sqrt{D}}\right) \\
\operatorname*{pklim}_{n_0,n_1,p\to\infty} \epsilon_1 &= \operatorname*{pklim}_{n_0,n_1,p\to\infty} \Phi\left(\frac{\hat{G}_1}{\sqrt{\hat{D}_1}}\right) = \Phi\left(\operatorname*{pklim}_{n_0,n_1,p\to\infty} \frac{\hat{G}_1}{\sqrt{\hat{D}_1}}\right) = \Phi\left(\frac{G_1}{\sqrt{D}}\right)
\end{aligned}
\tag{C.11}
$$

Boundedness and continuity of Φ allows one to apply the Helly-Bray Theorem [138] to obtain

$$
\begin{aligned}
\operatorname*{klim}_{n_0,n_1,p\to\infty} E[\epsilon_0] &= \operatorname*{klim}_{n_0,n_1,p\to\infty} E\left[\Phi\left(-\frac{\hat{G}_0}{\sqrt{\hat{D}_0}}\right)\right] = E\left[\Phi\left(\operatorname*{pklim}_{n_0,n_1,p\to\infty} \frac{-\hat{G}_0}{\sqrt{\hat{D}_0}}\right)\right] = \Phi\left(\frac{-G_0}{\sqrt{D}}\right) \\
\operatorname*{klim}_{n_0,n_1,p\to\infty} E[\epsilon_1] &= \operatorname*{klim}_{n_0,n_1,p\to\infty} E\left[\Phi\left(\frac{\hat{G}_1}{\sqrt{\hat{D}_1}}\right)\right] = E\left[\Phi\left(\operatorname*{pklim}_{n_0,n_1,p\to\infty} \frac{\hat{G}_1}{\sqrt{\hat{D}_1}}\right)\right] = \Phi\left(\frac{G_1}{\sqrt{D}}\right)
\end{aligned}
\tag{C.12}
$$

Proof of Theorem 17

Using the linear transformation introduced in the proof of Theorem 16, we first transform the data to normal distributions with $\Sigma = I$ and $\mu_1 = -\mu_0 = (\frac{\delta}{2}, 0, 0, \ldots, 0)^T$.

We prove that $\mathrm{Var}(\hat{G}_0^r) \xrightarrow{K} 0$ and $\mathrm{Var}(\hat{D}_0^r) \xrightarrow{K} 0$. Notice that the random vector $(X_1, \bar{X}_0, \bar{X}_1)$ has a multivariate normal distribution with mean vector (μ_0, μ_0, μ_1) and covariance matrix

$$
\begin{pmatrix}
I & \frac{I}{n_0} & 0 \\
\frac{I}{n_0} & \frac{I}{n_0} & 0 \\
0 & 0 & \frac{I}{n_1}
\end{pmatrix}
\tag{C.13}
$$

Using properties of the multivariate normal distribution [4], we conclude that

$$
X_1 \mid \bar{X}_0, \bar{X}_1 \sim N\left(\bar{X}_0, \left(1 - \frac{1}{n_0}\right) I\right)
\tag{C.14}
$$

From this it follows easily that

$$
\left(X_1 - \frac{\bar{X}_0 + \bar{X}_1}{2}\right)^T \left(\bar{X}_0 - \bar{X}_1\right) | \bar{X}_0, \bar{X}_1 \sim N\left(\frac{\hat{\delta}^2}{2}, \left(1 - \frac{1}{n_0}\right) \hat{\delta}^2\right)
\tag{C.15}
$$

in which $\hat{\delta}^2 = \left(\bar{X}_0 - \bar{X}_1\right)^T \left(\bar{X}_0 - \bar{X}_1\right)$. Hence, to show that $\mathrm{Var}(\hat{G}_0^r)\xrightarrow{K}0$ and $\mathrm{Var}(\hat{D}_0^r)\xrightarrow{K}0$, all we need to do is to show that $\mathrm{Var}(\hat{\delta}^2)\xrightarrow{K}0$. As we proved (C.8) using (C.7), it follows that

$$\mathrm{Var}(\hat{\delta}^2) = 4\delta^2\left(\frac{1}{n_0}+\frac{1}{n_1}\right) + 2p\left(\frac{1}{n_0}+\frac{1}{n_1}\right)^2 \xrightarrow{K}0 \tag{C.16}$$

as desired. By a simple application of Chebyshev's inequality, it follows that

$$\begin{aligned}\underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ \hat{G}_0^r &= \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\hat{G}_0^r] = \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[W(\bar{X}_0,\bar{X}_1,X_1)]\\ &= \underset{n_0,n_1,p\to\infty}{\mathrm{klim}}\left[\frac{\delta^2}{2}+\frac{p}{2}\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\right] = \frac{1}{2}(\delta^2+\lambda_0+\lambda_1)\triangleq G\end{aligned} \tag{C.17}$$

$$\begin{aligned}\underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ \hat{D}_0^r &= \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\hat{D}_0^r] = \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\mathrm{Var}(W(\bar{X}_0,\bar{X}_1,X_1 \mid \bar{X}_0,\bar{X}_1)]\\ &= \underset{n_0,n_1,p\to\infty}{\mathrm{klim}}\left[\left(1-\frac{1}{n_0}\right)\left(\delta^2+p\left(\frac{1}{n_0}+\frac{1}{n_1}\right)\right)\right] = \delta^2+\lambda_0+\lambda_1\triangleq D\end{aligned} \tag{C.18}$$

An analogous argument shows that$\mathrm{Var}(\hat{G}_1^r)\xrightarrow{K}0$ and $\mathrm{Var}(\hat{D}_1^r)\xrightarrow{K}0$ and

$$\begin{aligned}\underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ \hat{G}_1^r &= \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\hat{G}_1^r] = -\frac{1}{2}(\delta^2+\lambda_0+\lambda_1) = -G,\\ \underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ \hat{D}_1^r &= \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\hat{D}_1^r] = \delta^2+\lambda_0+\lambda_1 = D\end{aligned} \tag{C.19}$$

The rest of the proof proceeds much as in the case of the proof of Theorem 16. By using the Continuous Mapping Theorem [133], it follows that

$$\underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ \epsilon_0^r = \underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ \Phi\left(-\frac{\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}\right) = \Phi\left(\underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ -\frac{\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}\right) = \Phi\left(-\frac{G}{\sqrt{D}}\right) \tag{C.20}$$

$$\underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ \epsilon_1^r = \underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ \Phi\left(\frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}\right) = \Phi\left(\underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ \frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}\right) = \Phi\left(-\frac{G}{\sqrt{D}}\right) \tag{C.21}$$

Boundedness and continuity of Φ allows one to apply the Helly-Bray Theorem [138] to obtain

$$\operatorname*{klim}_{n_0,n_1,p\to\infty} E[\hat{\epsilon}_0^r] = \operatorname*{klim}_{n_0,n_1,p\to\infty} E[\epsilon_0^r] = \operatorname*{klim}_{n_0,n_1,p\to\infty} E\left[\Phi\left(\frac{-\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}\right)\right] = E\left[\Phi\left(\operatorname*{pklim}_{n_0,n_1,p\to\infty}\frac{-\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}\right)\right] = \Phi\left(\frac{-G}{\sqrt{D}}\right) \tag{C.22}$$

$$\operatorname*{klim}_{n_0,n_1,p\to\infty} E[\hat{\epsilon}_1^r] = \operatorname*{klim}_{n_0,n_1,p\to\infty} E[\epsilon_1^r] = \operatorname*{klim}_{n_0,n_1,p\to\infty} E\left[\Phi\left(\frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}\right)\right] = E\left[\Phi\left(\operatorname*{pklim}_{n_0,n_1,p\to\infty}\frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}\right)\right] = \Phi\left(\frac{-G}{\sqrt{D}}\right) \tag{C.23}$$

From this it also follows that

$$\begin{aligned} \operatorname*{klim}_{n_0,n_1,p\to\infty} E[\hat{\epsilon}^r] &= \operatorname*{klim}_{n_0,n_1,p\to\infty} (\hat{\alpha}_0 E[\hat{\epsilon}_0^r] + \hat{\alpha}_1 E[\hat{\epsilon}_1^r]) \\ &= \frac{\lambda_1}{\lambda_0+\lambda_1}\Phi\left(\frac{-G}{\sqrt{D}}\right) + \frac{\lambda_0}{\lambda_0+\lambda_1}\Phi\left(\frac{-G}{\sqrt{D}}\right) = \Phi\left(\frac{-G}{\sqrt{D}}\right) \end{aligned} \tag{C.24}$$

Proof of Theorem 19

Using the linear transformation in the proof of Theorem 16, we transform the data to normal distributions with $\Sigma = I$ and $\mu_1 = -\mu_0 = (\frac{\delta}{2}, 0, 0, \ldots, 0)^T$. In the proof of Theorem 17, it was shown that $\mathrm{Var}(\hat{G}_i^r) \xrightarrow{K} 0$ and $\mathrm{Var}(\hat{D}_i^r) \xrightarrow{K} 0$, for $i = 0, 1$, from which we have:

$$\begin{aligned} &\operatorname*{pklim}_{n_0,n_1,p\to\infty} \hat{G}_0^r = \frac{1}{2}(\delta^2 + \lambda_0 + \lambda_1) = G, \qquad \operatorname*{pklim}_{n_0,n_1,p\to\infty} \hat{G}_0^r = -\frac{1}{2}(\delta^2 + \lambda_0 + \lambda_1) = -G \\ &\operatorname*{pklim}_{n_0,n_1,p\to\infty} \hat{D}_0^r = \operatorname*{pklim}_{n_0,n_1,p\to\infty} \hat{D}_1^r = \delta^2 + \lambda_0 + \lambda_1 = D \end{aligned} \tag{C.25}$$

We now prove that $\mathrm{Var}(\hat{H}_0^r) \xrightarrow{K} 0$. Similarly to the proof of Theorem 17 and the way (C.14) was obtained, it is possible to show that

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \mid \bar{X}_0, \bar{X}_1 \sim N\left(\begin{bmatrix} \bar{X}_0 \\ \bar{X}_0 \end{bmatrix}, \begin{bmatrix} \left(1 - \frac{1}{n_0}\right) I & -\frac{1}{n_0} I \\ -\frac{1}{n_0} I & \left(1 - \frac{1}{n_0}\right) I \end{bmatrix}\right) \tag{C.26}$$

It follows that

$$\hat{H}_0^r = \mathrm{Cov}(W(\bar{X}_0, \bar{X}_1, X_1), W(\bar{X}_0, \bar{X}_1, X_2) \mid \bar{X}_0, \bar{X}_1) = -\frac{1}{n_0}\hat{\delta}^2 \tag{C.27}$$

where $\hat{\delta}^2$ was defined in the proof of Theorem 17. It was shown there that $\mathrm{Var}(\hat{\delta}^2) \xrightarrow{K} 0$. Therefore, $\mathrm{Var}(\hat{H}_0^r) \xrightarrow{K} 0$, as desired. Application of the Chebyshev's inequality yields

$$\begin{aligned} \underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ \hat{H}_0^r &= \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\hat{H}_0^r] \\ &= \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\mathrm{Cov}(W(\bar{X}_0, \bar{X}_1, X_1), W(\bar{X}_0, \bar{X}_1, X_2) \mid \bar{X}_0, \bar{X}_1)] \qquad \text{(C.28)} \\ &= \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} -\frac{1}{n_0}E[\hat{\delta}^2] = \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} \left[-\frac{\delta^2}{n_0} - \frac{p}{2}\left(\frac{1}{n_0^2} + \frac{1}{n_0 n_1}\right)\right] = 0 \end{aligned}$$

An analogous argument shows that $\mathrm{Var}(\hat{H}_1^r) \xrightarrow{K} 0$ and $\underset{n_0,n_1,p\to\infty}{\mathrm{pklim}}\ \hat{H}_1^r = \underset{n_0,n_1,p\to\infty}{\mathrm{klim}} E[\hat{H}_1^r] = 0$. The rest of the proof proceeds as in the case of the proofs of Theorem 16 and 17.

VITA

Amin Zollanvari received his B.S. and M.S. degrees in electrical engineering from Shiraz University, Iran, in 2003 and 2006. He graduated with his Ph.D. in electrical engineering from Texas A&M University in December 2010. His research interests include systems biology, statistical pattern recognition, and small-sample error estimation.

Amin Zollanvari may be reached at 214 Zachry Engineering Center, TAMU 3128, College Station, TX, 77843-3128. His email is zollanvari@gmail.com.

The typist for this thesis was Amin Zollanvari.