

Analytic Study of Performance of Error Estimators for Linear Discriminant Analysis

Amin Zollanvari, *Member, IEEE*, Ulisses M. Braga-Neto, *Member, IEEE*,
and Edward R. Dougherty, *Senior Member, IEEE*

Abstract

We derive double asymptotic 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 in the multivariate Gaussian model under the assumption of a common known covariance matrix and a fixed Mahalanobis distance as dimensionality approaches infinity. Sample sizes for the two classes need not be the same; they are only assumed to reach a fixed, but arbitrary, asymptotic ratio with the dimensionality. From the asymptotic moment representations, we directly obtain double asymptotic expressions for the bias, variance, and RMS of the error estimators. The asymptotic expressions presented here generally provide good small sample approximations, as demonstrated via numerical experiments. The applicability of the theoretical results is illustrated by finding the minimum sample size to bound the RMS in gene-expression classification.

Index Terms

Linear discriminant analysis, Error estimation, Resubstitution, Leave-one-out, RMS, Double asymptotics, Genomic signal processing

A. Zollanvari is with Children's Hospital Informatics Program at Harvard-MIT Division of Health Science, Brigham and Women's Hospital, and Harvard Medical School, Boston, Massachusetts, USA (e-mail: amin.zollanvari@childrens.harvard.edu)

U. M. Braga-Neto is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA (e-mail: ulisses@ece.tamu.edu)

E. R. Dougherty is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843 USA. He is also with Translational Genomics Research Institute (TGEN), Phoenix, AZ 85004 and with the Department of Bioinformatics and Computational Biology, University of Texas M. D. Anderson Cancer Center, Houston TX 77030 USA. (e-mail: edward@ece.tamu.edu; phone: 979-862-8896; fax: 979-845-6259)

I. INTRODUCTION

Linear discriminant analysis (LDA) has a long history in pattern recognition. It originated with an idea from R. Fisher (the “Fisher discriminant”) [1], [2], and was afterwards developed by A. Wald [3] and T.W. Anderson [4]. LDA is written down in terms of the “Anderson statistic” W . There is a large amount of work on obtaining exact, asymptotic or empirical expressions for the distribution of this statistic, which leads to the determination of the expected true error of misclassification by LDA and its variants. For comprehensive reviews of these results, the reader is referred to [5] and [6]. The LDA classification rule is consistent for Gaussian classes with the same covariance matrix; however, the true error rate of LDA depends on the actual parameters of the underlying Gaussian distributions of classes. Since these are generally unknown in practice, the error has to be estimated from the data. Epistemologically, model validity depends on the distance between the true error and the estimated error. In this regard, performance of error estimators is widely based on the bias, variance and root-mean square (RMS) error, the latter combining bias and variance in a single criterion.

Historically, there has been significant study of the moments of error estimators for different classification rules [7]–[15]. Most of this work has concentrated on the first two moments of resubstitution for LDA in the Gaussian model. In a recent paper, we have found the joint distribution between the true error and both the resubstitution and leave-one-out estimators for LDA in the Gaussian model with common known covariance matrix [16]. Here, we obtain, for the first time, analytic expressions for first and second moments of leave-one-out and, more importantly from the perspective of model validity, the second mixed moment between the true error and both resubstitution and leave-one-out for LDA in the same model, which facilitates analytical expression of the RMS. We derive asymptotically exact expressions by employing the Raudys-Kolmogorov double asymptotic approach, where both dimensionality and sample size approach infinity, while their linear ratio approaches a constant [17], [18]; this is justified by the fact that a linear ratio between dimensionality and sample size characterize the performance of linear classifiers both in Vapnik-Chervonenkis theory [35] and in empirical studies [45]. It is also assumed that the Mahalanobis distance between the classes, which is related in one-to-one fashion to the Bayes error (classification difficulty), is kept constant as dimensionality approach infinity, as one wants to examine error estimation performance as a function of fixed, but arbitrary, classification difficulty. Such an approach has been used before to obtain asymptotically-exact expressions for the first moment of actual and estimated (resubstitution) errors for LDA [18]–[21], whereas here we propose for the first time an extension of this approach to obtain asymptotically-exact expressions for the second moments

of actual and estimated (resubstitution and leave-one-out) errors as well as their cross-moments. The resulting approximations obtained are shown to be accurate in small-sample situations by means of numerical experiments. The approach we follow is to obtain Gaussian approximations based on the double asymptotic method. These expressions are simple to compute and accurate even in small sample cases. We will refer to the double asymptotic distribution of the linear discriminant used for classification of the sample point \mathbf{x} , denoted by $W(\mathbf{x})$, as a “first-order” analysis, since this is applied to obtain exact asymptotic expressions for the first-order moment of the actual and estimated errors. A novel aspect of this paper is that we also examine the double asymptotic joint distribution of the pair of random variables $(W(\mathbf{x}), W(\mathbf{x}'))$, which we term a “second-order” analysis, as those results are applied to obtain exact asymptotic expressions for the second-order and cross-moments of the actual and estimated errors.

In recent years, biomarker design in high-throughput genomics has confronted the pattern recognition community with very small samples from which one wishes to design classifiers and estimate errors on the same data. In such cases, asymptotic results regarding the convergence of error estimators for large sample sizes are not relevant. We require performance analysis based on sample size. In particular, we would like performance (RMS) bounds dependent on sample size that would allow us to determine a minimal sample size to achieve a desired level of error-estimation accuracy, which translates into a measure of the validity of the classifier model [22]. Indeed, our motivation for undertaking these kinds of studies has been to statistically characterize model validity in genomic classification, which is a major aspect of genomic signal processing [23]. Linear classification is the obvious first place to start, owing to its long history in pattern recognition and its suitability for small-sample classification. The latter is perhaps the reason that LDA-based classification and recognition systems have found a broad range of applications in many disciplines such as speech recognition [24], face recognition [25], texture classification [26], cancer classification [27], [28] and many more [29].

This paper is organized as follows. Section II reviews LDA classification and error estimation, while Section III discusses first and second-order performance criteria for error estimation. From this point on, the paper focusses its attention of resubstitution and leave-one-out error estimation. Section IV investigates what we call the “first-order” double-asymptotic approximations (i.e., double-asymptotic approximations to first-order moments of actual error and error estimator distributions), highlighting the classical results known from S. Raudys’ work, as well as proposing additional approximations and providing proofs of the asymptotic exactness of all approximations. Section V presents the novel “second-order” analysis of double-asymptotic approximations (i.e., double-asymptotic approximations to second-order moments of actual error and error estimator distributions), which allows one to derive the second-

order performance criteria of variance and RMS of error estimators; all proposed approximations are proved to be asymptotically exact, and are demonstrated with a simple numerical example. Section VI discusses the double asymptotic limit of bias, variance and RMS. Section VII presents analytical RMS bounds that follow from the previous theory, sample size calculations, and a numerical example involving gene-expression classification. Section VIII concludes the paper. Proofs are in the Appendix.

II. LINEAR DISCRIMINANT ANALYSIS AND ERROR ESTIMATION

Consider a set of $n = n_0 + n_1$ independent and identically-distributed training samples in R^p , where $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n_0}$ come from population Π_0 and $\mathbf{x}_{n_0+1}, \mathbf{x}_{n_0+2}, \dots, \mathbf{x}_{n_0+n_1}$ come from population Π_1 . Population Π_i is assumed to follow a multivariate Gaussian distribution $N(\mu_i, \Sigma)$, for $i = 0, 1$. *Linear Discriminant Analysis* (LDA) employs Anderson's W statistic, given by

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

where $\bar{\mathbf{x}}_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} \mathbf{x}_i$ and $\bar{\mathbf{x}}_1 = \frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} \mathbf{x}_i$ are the sample means for each class. The designed LDA classifier is given by

$$\psi(\mathbf{x}) = \begin{cases} 1, & \text{if } W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \leq 0 \\ 0, & \text{if } W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) > 0 \end{cases}, \quad (2)$$

that is, the sign of W determines the classification of the sample point \mathbf{x} . Here, following for example [30]–[32], 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 plugged in as an estimator of Σ in the expressions that are derived in what follows, which can and often does produce useful results, unless the sample size is very small.

Given the training data S_n (and thus $\bar{\mathbf{x}}_0$ and $\bar{\mathbf{x}}_1$), the classification error is given by

$$\epsilon = P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \leq 0, \mathbf{x} \in \Pi_0 \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1) + P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) > 0, \mathbf{x} \in \Pi_1 \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1) = \alpha_0 \epsilon_0 + \alpha_1 \epsilon_1 \quad (3)$$

where $\alpha_i = P(\mathbf{x} \in \Pi_i)$ is the a-priori mixing probability for population Π_i , and ϵ_i is the error rate specific to population Π_i , with

$$\epsilon_0 = P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \leq 0 \mid \mathbf{x} \in \Pi_0, \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1) \quad \text{and} \quad \epsilon_1 = P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) > 0 \mid \mathbf{x} \in \Pi_1, \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1). \quad (4)$$

The *resubstitution error estimator* [33], is given by

$$\hat{\epsilon}_r = \frac{1}{n} \left[\sum_{i=1}^{n_0} I_{\{W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_i) \leq 0\}} + \sum_{i=n_0+1}^{n_0+n_1} I_{\{W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_i) > 0\}} \right] = \hat{\alpha}_0 \hat{\epsilon}_0^r + \hat{\alpha}_1 \hat{\epsilon}_1^r, \quad (5)$$

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

$$\hat{\epsilon}_0^r = \frac{1}{n_0} \sum_{i=1}^{n_0} I_{\{W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_i) \leq 0\}} \quad \text{and} \quad \hat{\epsilon}_1^r = \frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} I_{\{W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_i) > 0\}}. \quad (6)$$

The *leave-one-out error estimator* [34] for the LDA classification rule is given by

$$\hat{\epsilon}^l = \frac{1}{n} \left[\sum_{i=1}^{n_0} I_{\{W^{(i)}(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_i) \leq 0\}} + \sum_{i=n_0+1}^{n_0+n_1} I_{\{W^{(i)}(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_i) > 0\}} \right] = \hat{\alpha}_0 \hat{\epsilon}_0^l + \hat{\alpha}_1 \hat{\epsilon}_1^l \quad (7)$$

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

$$\hat{\epsilon}_0^l = \frac{1}{n_0} \sum_{i=1}^{n_0} I_{\{W^{(i)}(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_i) \leq 0\}} \quad \text{and} \quad \hat{\epsilon}_1^l = \frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} I_{\{W^{(i)}(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_i) > 0\}}. \quad (8)$$

III. CRITERIA FOR 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, deviation variance, and RMS, given by

$$\begin{aligned} \text{Bias}[\hat{\epsilon}] &= E[\hat{\epsilon}] - E[\epsilon], \\ \text{Var}_d[\hat{\epsilon}] &= \text{Var}(\hat{\epsilon} - \epsilon) = \text{Var}(\epsilon) + \text{Var}(\hat{\epsilon}) - 2\text{Cov}(\epsilon, \hat{\epsilon}), \\ \text{RMS}[\hat{\epsilon}] &= \sqrt{E[(\epsilon - \hat{\epsilon})^2]} = \sqrt{E[\epsilon^2] + E[\hat{\epsilon}^2] - 2E[\epsilon\hat{\epsilon}]}, \end{aligned} \quad (9)$$

respectively. The bias and the deviation variance measure the average centrality and dispersion of the error estimator in relation to the actual error, respectively. 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 (see [35, Chapters 23,24] and references therein). The RMS combines these two complementary criteria into a single metric: $\text{RMS}[\hat{\epsilon}] = \sqrt{\text{Bias}[\hat{\epsilon}]^2 + \text{Var}_d[\hat{\epsilon}]}$.

The bias, variance, and RMS can be obtained from 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 $W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x})$. These hold in general, not just for the Gaussian model. We will write all equations for the resubstitution estimator; the corresponding equations for the leave-one-out estimator can be obtained by replacing $W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_i)$ by $W^{(i)}(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_i)$ throughout.

From (3) and (4), the first moment of the actual error is given by

$$E[\epsilon] = \alpha_0 E[\epsilon_0] + \alpha_1 E[\epsilon_1] = \alpha_0 P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \leq 0 \mid \mathbf{x} \in \Pi_0) + \alpha_1 P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) > 0 \mid \mathbf{x} \in \Pi_1) \quad (10)$$

For the second moments of the actual error, it follows immediately from Theorem 24.2 in [35] that

$$E[\epsilon^2] = E[(\alpha_0\epsilon_0 + \alpha_1\epsilon^1)^2] = \alpha_0^2 E[\epsilon_0\epsilon_0] + 2\alpha_0\alpha_1 E[\epsilon_0\epsilon_1] + \alpha_1^2 E[\epsilon^1\epsilon_1] \quad (11)$$

and $E[\epsilon_0\epsilon_0] = P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \leq 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}') \leq 0 \mid \mathbf{x}, \mathbf{x}' \in \Pi_0)$, with similar expressions for the other terms in (11), namely $E[\epsilon_0\epsilon_1]$ and $E[\epsilon^1\epsilon_1]$. In all,

$$\begin{aligned} E[\epsilon^2] &= \alpha_0^2 P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \leq 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}') \leq 0 \mid \mathbf{x}, \mathbf{x}' \in \Pi_0) \\ &\quad + 2\alpha_0\alpha_1 P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \leq 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}') > 0 \mid \mathbf{x} \in \Pi_0, \mathbf{x}' \in \Pi_1) \\ &\quad + \alpha_1^2 P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) > 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}') > 0 \mid \mathbf{x}, \mathbf{x}' \in \Pi_1) \end{aligned} \quad (12)$$

From (5) and (6), for resubstitution we have

$$E[\hat{\epsilon}^r] = \hat{\alpha}_0 E[\hat{\epsilon}_0^r] + \hat{\alpha}_1 E[\hat{\epsilon}_1^r] = \hat{\alpha}_0 P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \leq 0) + \hat{\alpha}_1 P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1}) > 0) \quad (13)$$

The corresponding equation for leave-one-out is obtained by replacing $W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1)$ and $W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1})$ by $W^{(1)}(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1)$ and $W^{(n_0+1)}(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1})$, respectively. Note that $W^{(1)}(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1)$ is distributed as $W'(\bar{\mathbf{x}}'_0, \bar{\mathbf{x}}_1, \mathbf{x})$ conditioned on $\mathbf{x} \in \Pi_0$, where W' and $\bar{\mathbf{x}}'$ are the usual W and $\bar{\mathbf{x}}$ 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 $W^{(n_0+1)}(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1})$. By virtue of (10), this leads to $E[\hat{\epsilon}^l] = E[\epsilon_{n-1}]$, if $\hat{\alpha}_i = \alpha_i$, for $i = 0, 1$.

Using (5) and (6), we can extend these results to resubstitution. The second moment is given by

$$\begin{aligned} E[(\hat{\epsilon}^r)^2] &= \frac{\hat{\alpha}_0^2}{n_0} P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \leq 0) + \frac{\hat{\alpha}_1^2}{n_1} P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1}) > 0) \\ &\quad + \hat{\alpha}_0^2 \frac{n_0 - 1}{n_0} P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \leq 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_2) \leq 0) \\ &\quad + \hat{\alpha}_1^2 \frac{n_1 - 1}{n_1} P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1}) > 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+2}) > 0) \\ &\quad + 2\hat{\alpha}_0\hat{\alpha}_1 P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \leq 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1}) > 0) \end{aligned} \quad (14)$$

From (3) and (5), we obtain the mixed moment

$$E[\hat{\epsilon}\hat{\epsilon}^r] = \alpha_0\hat{\alpha}_0 E[\epsilon_0\hat{\epsilon}_0^r] + \alpha_0\hat{\alpha}_1 E[\epsilon_0\hat{\epsilon}_1^r] + \alpha_1\hat{\alpha}_0 E[\epsilon_1\hat{\epsilon}_0^r] + \alpha_1\hat{\alpha}_1 E[\epsilon_1\hat{\epsilon}_1^r] \quad (15)$$

It follows from (4) and (6) that $E[\epsilon^0\hat{\epsilon}_0^r] = P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \leq 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \leq 0 \mid \mathbf{x} \in \Pi_0)$. Similar expressions hold for the other terms in (15), 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,

$$\begin{aligned} E[\hat{\epsilon}\hat{\epsilon}^r] &= \alpha_0\hat{\alpha}_0 P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \leq 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \leq 0 \mid \mathbf{x} \in \Pi_0) \\ &\quad + \alpha_0\hat{\alpha}_1 P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \leq 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1}) > 0 \mid \mathbf{x} \in \Pi_0) \\ &\quad + \alpha_1\hat{\alpha}_0 P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) > 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \leq 0 \mid \mathbf{x} \in \Pi_1) \\ &\quad + \alpha_1\hat{\alpha}_1 P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) > 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1}) > 0 \mid \mathbf{x} \in \Pi_1) \end{aligned} \quad (16)$$

IV. FIRST-ORDER DOUBLE ASYMPTOTIC ANALYSIS

Most asymptotic theory treats the behavior of a statistic as the sample size goes to infinity, for fixed dimensionality and number of parameters. The double asymptotic method, conceived independently by S. Raudys and A. Kolmogorov [17], [36] in 1967, considers the behavior of a statistic as both sample size and dimensionality (or, more generally, number of parameters) increase to infinity in a controlled fashion, where the ratio between sample size and dimensionality converges to a finite constant [17]–[19], [36]–[39]. This “increasing dimension limit” is also known, in the Machine Learning literature, as the “thermodynamic limit” [36], [40]. Via the double asymptotic approach, one may characterize the asymptotic behavior of the classification error as sample size, n , and dimensionality, p , increase, and n is a fixed number of times larger than p . The finite-sample approximations obtained via these asymptotic expressions have been shown to be remarkably accurate in small-sample cases [41], [42].

A. Previous Work

In [18], Raudys proposed an approximation to the expected actual classification error:

$$E[\epsilon_0] = P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \leq 0 \mid \mathbf{x} \in \Pi_0) \approx \Phi \left(-\frac{E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \mathbf{x} \in \Pi_0]}{\sqrt{\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \mathbf{x} \in \Pi_0)}} \right) \quad (17)$$

in which $\Phi(\cdot)$ is the standard normal cumulative function. To obtain the corresponding approximation to $E[\epsilon_1]$, modify the argument of Φ by replacing Π_0 by Π_1 and multiplying by -1 . If $n_0 = n_1 = n$, then $E[\epsilon] = E[\epsilon_0] = E[\epsilon_1]$. Using (17) in this case, Raudys obtained the approximation [43]:

$$E[\epsilon] \approx \Phi \left(-\frac{\delta}{2} \frac{1}{\sqrt{1 + \frac{1}{n} + \frac{2p}{n\delta^2} + \frac{p}{n^2\delta^2}}} \right) \quad (18)$$

where $\delta^2 = (\mu_0 - \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1)$. In [36], Raudys pointed out, without exhibiting an explicit proof, that this approximation is asymptotically exact under the double asymptotic condition $n \rightarrow \infty$, $p \rightarrow \infty$, $n/p \rightarrow \text{constant}$. Under these conditions, the following asymptotically-equivalent approximation results:

$$E[\epsilon] \approx \Phi \left(-\frac{\delta}{2} \frac{1}{\sqrt{1 + \frac{2p}{n\delta^2}}} \right) \quad (19)$$

To obtain the approximation for the expectation of resubstitution, modify (17) by replacing \mathbf{x} by \mathbf{x}_1 :

$$E[\hat{\epsilon}_0^r] = P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \leq 0) \approx \Phi \left(-\frac{E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1)]}{\sqrt{\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1))}} \right) \quad (20)$$

To obtain the corresponding approximation to $E[\hat{\epsilon}_1^r]$, modify the argument of Φ by replacing \mathbf{x}_1 by \mathbf{x}_{n_0+1} and multiply by -1 . If $n_0 = n_1 = n$, then $E[\hat{\epsilon}^r] = E[\hat{\epsilon}_0^r] = E[\hat{\epsilon}_1^r]$, and (20) yields the approximation

$$E[\hat{\epsilon}^r] \approx \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{2p}{n\delta^2}}{\sqrt{1 + \frac{1}{n} + \frac{2p}{n\delta^2}}} \right) \quad (21)$$

This expression is equivalent to the one published by Raudys in [21], [36], under the double asymptotic condition $n \rightarrow \infty$, $p \rightarrow \infty$, $n/p \rightarrow \text{constant}$, namely:

$$E[\hat{\epsilon}^r] \approx \Phi \left(-\frac{\delta}{2} \sqrt{1 + \frac{2p}{n\delta^2}} \right) \quad (22)$$

We will prove in the following subsections that all the approximations discussed above are asymptotically exact, as $n_0 \rightarrow \infty$, $n_1 \rightarrow \infty$, $p \rightarrow \infty$, $p/n_0 \rightarrow \lambda_0$, $p/n_1 \rightarrow \lambda_1$. Serdobolskii calls these ‘‘Kolmogorov asymptotic conditions’’ in [17]; but since the first paper to employ this approach is due to S. Raudys [43], we prefer to call these ‘‘Raudys-Kolmogorov asymptotic conditions.’’ In what follows, we will denote convergence in probability under Raudys-Kolmogorov asymptotic conditions by ‘‘ $\underset{n_0, n_1, p \rightarrow \infty}{\text{pklim}}$ ’’. Similarly, ‘‘ $\underset{n_0, n_1, p \rightarrow \infty}{\text{klim}}$ ’’ and ‘‘ $\overset{K}{\rightarrow}$ ’’ will denote ordinary convergence under the Raudys-Kolmogorov asymptotic conditions. These limits are referred in the paper as ‘‘Raudys-Kolmogorov limits.’’

While the limiting linear ratio conditions $p/n_0 \rightarrow \lambda_0$ and $p/n_1 \rightarrow \lambda_1$ impose a theoretical constraint on the results, they are quite natural from two practical perspectives. First, the ratio of dimension to sample size serves a measure of complexity for the LDA classification rule (in fact, any linear classification rule): in this case, the VC dimension is $p+1$ [35], so that p/n , where $n = n_0 + n_1$, gives the asymptotic relationship between VC dimension and the sample size. In addition, for LDA with known covariance matrix, under the assumption that $n_0 = n_1$, it has long been known that, if the features contribute equally to the Mahalanobis distance, then the optimal number of features is $p_{\text{opt}} = n - 1$, so that choosing the optimal number of features leads to $p/n_k \rightarrow \text{constant}$, for $k = 0, 1$ [44]. A linear relation between the optimal number of features and the sample size has also been observed empirically in more general settings: assuming a blocked covariance matrix with identical off-diagonal correlation coefficient ρ and with the covariance matrix unknown in the classifier design, empirical linear relationships have been observed, for instance, with $\rho = 0.125$, one has $p_{\text{opt}} \approx n/3$, whereas with $\rho = 0.5$, one has $p_{\text{opt}} \approx n/25$ [45]. Therefore, the stronger non-limiting constraints, $p/n_0 = \lambda_0$ and $p/n_1 = \lambda_1$, would still be relevant from a practical perspective, even though here we only require that $p/n_0 \rightarrow \lambda_0$ and $p/n_1 \rightarrow \lambda_1$.

For simplifying the notations, the following functions are defined that will be used throughout the

paper:

$$\begin{aligned} f_0(n_0, n_1, p, \delta^2) &= \sqrt{1 + \frac{1}{n_1} + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right) + \frac{p}{2\delta^2} \left(\frac{1}{n_0^2} + \frac{1}{n_1^2} \right)}, & f_1(n_0, n_1, p, \delta^2) &= f_0(n_1, n_0, p, \delta^2) \\ g_0(n_0, n_1, p, \delta^2) &= \sqrt{1 + \frac{1}{n_1} + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right) + \frac{p}{2\delta^2} \left(\frac{1}{n_1^2} - \frac{1}{n_0^2} \right)}, & g_1(n_0, n_1, p, \delta^2) &= g_0(n_1, n_0, p, \delta^2) \end{aligned} \quad (23)$$

B. Actual Classification Error

Define a family of Gaussian discrimination problems by a set of parameter and sample sizes:

$$(\mu_{p,0}, \mu_{p,1}, \Sigma_p, n_{p,0}, n_{p,1}), \quad p = 1, 2, \dots, \quad (24)$$

where the means and covariance matrix are arbitrary except that the Mahalanobis distance, defined as $\delta = \sqrt{(\mu_{p,0} - \mu_{p,1}) \Sigma_p^{-1} (\mu_{p,0} - \mu_{p,1})}$, is kept constant as dimensionality approaches infinity (with slightly more work, this condition can be relaxed to a varying Mahalanobis distance converging to a constant δ as $p \rightarrow \infty$, as in [19]). Either condition is justified by the fact that one wants to examine error estimation behavior as a function of fixed, but arbitrary, δ , i.e. Bayes error (classification difficulty). For notational ease, we will omit the subscript “ p ” from the parameters and sample sizes in (24).

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 [17]. Let

$$\hat{G}_i = E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x} \in \Pi_i], \quad \hat{D}_i = \text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x} \in \Pi_i) \quad (25)$$

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

$$\epsilon_0 = \Phi\left(-\frac{\hat{G}_0}{\sqrt{\hat{D}_0}}\right), \quad \epsilon_1 = \Phi\left(\frac{\hat{G}_1}{\sqrt{\hat{D}_1}}\right) \quad (26)$$

Theorem 1: Consider the sequence of Gaussian discrimination problems defined by (24). Then

$$\text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_0] = \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_0 = \Phi\left(\frac{-G_0}{\sqrt{D}}\right), \quad \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_1] = \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_1 = \Phi\left(\frac{G_1}{\sqrt{D}}\right) \quad (27)$$

so that

$$\text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon] = \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon = \alpha_0 \Phi\left(-\frac{G_0}{\sqrt{D}}\right) + \alpha_1 \Phi\left(\frac{G_1}{\sqrt{D}}\right) \quad (28)$$

where

$$\begin{aligned} G_0 &= \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{G}_0] = \frac{1}{2}(\delta^2 + \lambda_1 - \lambda_0), \\ G_1 &= \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{G}_1] = -\frac{1}{2}(\delta^2 + \lambda_0 - \lambda_1), \\ D &= \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{D}_0] = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{D}_1] = \delta^2 + \lambda_0 + \lambda_1 \end{aligned} \quad (29)$$

Proof: See Appendix. ■

(28) is equivalent to the specialization of Deev's formula [36] to the case of known covariance matrix.

Theorem 1 suggests the following finite-sample approximation:

$$E[\epsilon_0] \approx \Phi \left(-\frac{E[\hat{G}_0]}{\sqrt{E[\hat{D}_0]}} \right) = \Phi \left(-\frac{E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \mathbf{x} \in \Pi_0]}{\sqrt{E[\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x} \in \Pi_0)]}} \right) \quad (30)$$

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 (30) yields

$$E[\epsilon_0] \approx \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0} \right)}{\sqrt{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} + \frac{1}{n_0} \right)}} \right) \quad (31)$$

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 1. However, in the case $n_0 = n_1 = n$, (31) reduces to (19) and not (18). The reason is that, if one compares (30) to Raudys' formula (17), one observes that the denominators differ by the term:

$$\begin{aligned} & \text{Var}[E(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x} \in \Pi_0)] = \\ & \text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \mathbf{x} \in \Pi_0) - E[\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x} \in \Pi_0)] = \frac{\delta^2}{n_1} + \frac{p}{n_0^2} + \frac{p}{n_1^2} \xrightarrow{K} 0 \end{aligned} \quad (32)$$

Hence, the finite-sample approximations obtained by (17) and (30) differ, but are asymptotically equivalent. By Theorem 1, this also proves that Raudys' approximation (18) is indeed asymptotically exact. For moderate n_0/p and n_1/p , (32) becomes close to zero, and (17) and (30) yield very similar values.

The next expression is the finite-sample approximation obtained with Raudys' formula (17) in the general case $n_0 \neq n_1$, which has not been available before:

$$E[\epsilon_0] \approx \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0} \right)}{f_0(n_0, n_1, p, \delta^2)} \right) \quad (33)$$

which of course reduces to (18) when $n_0 = n_1 = n$. If we remove the terms which tend to zero under Raudys-Kolmogorov asymptotic conditions, then (33) becomes:

$$E[\epsilon_0] \approx \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0} \right)}{\sqrt{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)}} \right) \quad (34)$$

i.e., the same as (31), which reduces to (19) when $n_0 = n_1 = n$. Also notice that (34) corresponds to replacing λ_0 by p/n_0 and λ_1 by p/n_1 in (27), as it should. To obtain the corresponding approximations for $E[\epsilon_1]$, it suffices to exchange n_0 and n_1 in (33) and (34).

C. Resubstitution Error Estimator

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

$$\epsilon_0^r = P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \leq 0 \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1) \quad \text{and} \quad \epsilon_1^r = P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1}) > 0 \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1) \quad (35)$$

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,

$$E[\hat{\epsilon}^r] = \hat{\alpha}_0 E[\epsilon_0^r] + \hat{\alpha}_1 E[\epsilon_1^r]. \quad (36)$$

Let

$$\begin{aligned} \hat{G}_0^r &= E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1], \\ \hat{G}_1^r &= E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1}) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1], \\ \hat{D}_0^r &= \text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1), \\ \hat{D}_1^r &= \text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1}) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1). \end{aligned} \quad (37)$$

Then

$$\epsilon_0^r = \Phi\left(-\frac{\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}\right), \quad \epsilon_1^r = \Phi\left(\frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}\right) \quad (38)$$

Theorem 2: Consider the sequence of Gaussian discrimination problems defined by (24). Then

$$\text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{\epsilon}^r] = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{\epsilon}_0^r] = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{\epsilon}_1^r] = \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_0^r = \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_1^r = \Phi\left(\frac{-G}{\sqrt{D}}\right) \quad (39)$$

where

$$\begin{aligned} G &= \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{G}_0^r] = - \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{G}_1^r] = \frac{1}{2}(\delta^2 + \lambda_0 + \lambda_1) \\ D &= \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{D}_0^r] = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{D}_1^r] = \delta^2 + \lambda_0 + \lambda_1 \end{aligned} \quad (40)$$

Proof: See Appendix. ■

Theorem 2 suggests the following finite-sample approximation:

$$E[\hat{\epsilon}_0^r] \approx \Phi\left(-\frac{E[\hat{G}_0^r]}{\sqrt{E[\hat{D}_0^r]}}\right) = \Phi\left(-\frac{E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1)]}{\sqrt{E[\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1])}}\right) \quad (41)$$

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 \mathbf{x}_1 by \mathbf{x}_{n_0+1} , and multiply the argument of both Φ functions by -1 . Evaluating the expectation in the numerator and denominator of (41) yields

$$E[\hat{\epsilon}_0^r] \approx \Phi\left(-\frac{\delta}{2\sqrt{1 - \frac{1}{n_0}}} \sqrt{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1}\right)}\right) \quad (42)$$

with the corresponding approximation for $E[\hat{\epsilon}_1^r]$ obtained by exchanging n_0 and n_1 . Theorem 2 shows this approximation is asymptotically exact. If $n_0 = n_1 = n$, then (42) reduces to

$$E[\hat{\epsilon}_0^r] \approx \Phi \left(-\frac{\delta}{2\sqrt{1-\frac{1}{n}}} \sqrt{1 + \frac{2p}{n\delta^2}} \right) \quad (43)$$

which is not the same as (21) or (22). Once again, the reason is that, if one compares (41) to Raudys' formula (20), one observes that the denominators differ by the term:

$$\begin{aligned} \text{Var}[E(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1)] &= \text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1)) - E[\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1)] \\ &= \delta^2 \left(\frac{1}{n_0} + \frac{1}{n_1} \right) + \frac{p}{2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)^2 \xrightarrow{K} 0 \end{aligned} \quad (44)$$

Hence, the finite-sample approximations obtained by (20) and (41) differ, but are asymptotically equivalent. Furthermore, both are asymptotically equivalent to (22). Incidentally, this proves that both (21) and (22) are asymptotically exact. For moderate values of n_0 , n_1 , n_0/p , and n_1/p , the term (44) 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 (20) in the general case $n_0 \neq n_1$, which has not been available before:

$$E[\hat{\epsilon}_0^r] \approx \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)}{g_0(n_0, n_1, p, \delta^2)} \right) \quad (45)$$

which of course reduces to (21) 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 (45). If we remove the terms which tend to zero under Raudys-Kolmogorov asymptotic conditions, then (45) and (42) both become:

$$E[\hat{\epsilon}^r] \approx E[\hat{\epsilon}_0^r] \approx E[\hat{\epsilon}_1^r] \approx \Phi \left(-\frac{\delta}{2} \sqrt{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)} \right) \quad (46)$$

which reduces to (22) when $n_0 = n_1 = n$. Also notice that (46) corresponds to replacing λ_0 by p/n_0 and λ_1 by p/n_1 in (39), as it should.

D. Leave-one-out Error Estimator

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

V. 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 $(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}), W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}'))$, 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.

A. Second-Order Approximations

We start by considering the extension of equations (17) and (20) to second moments. Consider the standard bivariate Gaussian distribution function

$$\Phi(a, b; \rho) = \int_{-\infty}^a \int_{-\infty}^b \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}(x^2 + y^2 - 2\rho xy)\right\} dx dy \quad (47)$$

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 Φ :

$$P(x \leq c, y \leq d) = \Phi\left(\frac{c - \mu_x}{\sigma_x}, \frac{d - \mu_y}{\sigma_y}; \rho_{xy}\right) \quad (48)$$

where $\mu_x = E[x]$, $\mu_y = E[y]$, $\sigma_x = \sqrt{\text{Var}(x)}$, $\sigma_y = \sqrt{\text{Var}(y)}$, and ρ_{xy} is the correlation coefficient between x and y . Using (48), we obtain the second-order extension of Raudys' formula (17):

$$E[\epsilon_0^2] = P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \leq 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}') \leq 0 \mid \mathbf{x}, \mathbf{x}' \in \Pi_0) \approx \Phi\left(-\frac{E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \mathbf{x} \in \Pi_0]}{\sqrt{\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \mathbf{x} \in \Pi_0)}}; \frac{\text{Cov}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}), W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}') \mid \mathbf{x}, \mathbf{x}' \in \Pi_0)}{\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \mathbf{x} \in \Pi_0)}\right) \quad (49)$$

In the general case $n_0 \neq n_1$, evaluation of the terms in (49) yields

$$E[\epsilon_0^2] \approx \Phi\left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0}\right)}{f_0(n_0, n_1, p, \delta^2)}; \frac{\frac{1}{n_1} + \frac{p}{2\delta^2} \left(\frac{1}{n_0^2} + \frac{1}{n_1^2}\right)}{f_0^2(n_0, n_1, p, \delta^2)}\right) \quad (50)$$

Equation (50) is the second-order extension of (33). Similarly, it can be shown that

$$E[\epsilon_0\epsilon_1] \approx \Phi\left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0}\right)}{f_0(n_0, n_1, p, \delta^2)}\right) \Phi\left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} - \frac{1}{n_1}\right)}{f_1(n_0, n_1, p, \delta^2)}\right) \quad (51)$$

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 Raudys-Kolmogorov asymptotic conditions the covariance term in (50) becomes zero, and the pair of random variables $(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}), W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}'))$ become independent. This suggests the approximation

$$E[\epsilon_0^2] \approx \left[\Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0} \right)}{\sqrt{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)}} \right) \right]^2 \quad (52)$$

Equation (52) is simply the square of the approximation (34). 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, (49) is modified by replacing \mathbf{x} and \mathbf{x}' by \mathbf{x}_1 and \mathbf{x}_2 , respectively:

$$\begin{aligned} E[(\hat{\epsilon}_0^r)^2] &= P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \leq 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_2) \leq 0) \\ &\approx \Phi \left(-\frac{E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1)]}{\sqrt{\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1))}}; \frac{\text{Cov}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1), W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_2))}{\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1))} \right) \end{aligned} \quad (53)$$

In the general case $n_0 \neq n_1$, (53) gives

$$E[(\hat{\epsilon}_0^r)^2] \approx \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)}{g_0(n_0, n_1, p, \delta^2)}; \frac{\frac{1}{n_1} + \frac{p}{2\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0} \right)}{g_0^2(n_0, n_1, p, \delta^2)} \right) \quad (54)$$

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

$$E[\hat{\epsilon}_0^r \hat{\epsilon}_1^r] \approx \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)}{g_0(n_0, n_1, p, \delta^2)}, -\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)}{g_1(n_0, n_1, p, \delta^2)}; \frac{\frac{1}{n_0} + \frac{1}{n_1} + \frac{p}{2\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_0} \right)^2}{g_0(n_0, n_1, p, \delta^2) g_1(n_0, n_1, p, \delta^2)} \right) \quad (55)$$

Deleting terms tending to 0 under Raudys-Kolmogorov asymptotic conditions in (54) gives the approximation

$$E[(\hat{\epsilon}_0^r)^2] \approx \left[\Phi \left(-\frac{\delta}{2} \sqrt{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)} \right) \right]^2 \quad (56)$$

Equation (56) is simply the square of the approximation (46). 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

$$\begin{aligned} E[\epsilon_0 \hat{\epsilon}_0^r] &= P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \leq 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \leq 0 \mid \mathbf{x} \in \Pi_0) \approx \\ &\Phi \left(\frac{-E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \mathbf{x} \in \Pi_0]}{\sqrt{\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \mathbf{x} \in \Pi_0)}}; \frac{-E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1)]}{\sqrt{\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1))}}; \frac{\text{Cov}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}), W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \mid \mathbf{x} \in \Pi_0)}{\sqrt{\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \mathbf{x} \in \Pi_0)} \sqrt{\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1))}} \right) \end{aligned} \quad (57)$$

In the general case $n_0 \neq n_1$, (57) gives

$$E[\epsilon_0 \hat{\epsilon}_0^r] \approx \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0} \right)}{f_0(n_0, n_1, p, \delta^2)}, -\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)}{g_0(n_0, n_1, p, \delta^2)}; \frac{\frac{1}{n_1} + \frac{p}{2\delta^2} \left(\frac{1}{n_1^2} - \frac{1}{n_0^2} \right)}{f_0(n_0, n_1, p, \delta^2) g_0(n_0, n_1, p, \delta^2)} \right) \quad (58)$$

Similarly, it can be shown that

$$E[\epsilon_0 \hat{\epsilon}_1^r] \approx \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0} \right)}{f_0(n_0, n_1, p, \delta^2)}, -\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)}{g_1(n_0, n_1, p, \delta^2)}; \frac{\frac{1}{n_1} + \frac{p}{2\delta^2} \left(\frac{1}{n_1^2} - \frac{1}{n_0^2} \right)}{f_0(n_0, n_1, p, \delta^2) g_1(n_0, n_1, p, \delta^2)} \right) \quad (59)$$

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 .

Deleting terms tending to 0 under Raudys-Kolmogorov asymptotic conditions in (58) gives the approximation

$$E[\epsilon_0 \hat{\epsilon}_0^r] \approx \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0} \right)}{\sqrt{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)}} \right) \Phi \left(-\frac{\delta}{2} \sqrt{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)} \right) \quad (60)$$

Equation (60) is simply the product of the approximations in (34) and (46). 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 $E[(\hat{\epsilon}_0^l)^2]$, (53) is modified by replacing $W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1)$ by $W^{(1)}(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1)$ and $W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_2)$ by $W^{(2)}(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_2)$. In the general case $n_0 \neq n_1$, this gives

$$E[(\hat{\epsilon}_0^l)^2] \approx \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0-1} \right)}{f_0(n_0-1, n_1, p, \delta^2)}; \frac{\frac{1}{n_1} + \frac{p}{2\delta^2} \left(\frac{1}{n_1^2} + \frac{2}{(n_0-1)^4} - \frac{(n_0-2)^2}{(n_0-1)^4} \right)}{f_0^2(n_0-1, n_1, p, \delta^2)} \right) \quad (61)$$

Exchanging n_0 and n_1 in $E[(\hat{\epsilon}_0^l)^2]$ yields the corresponding approximation for $E[(\hat{\epsilon}_1^l)^2]$. Similarly,

$$E[\hat{\epsilon}_0^l \hat{\epsilon}_1^l] \approx \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0-1} \right)}{f_0(n_0-1, n_1, p, \delta^2)}, -\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} - \frac{1}{n_1-1} \right)}{f_1(n_0, n_1-1, p, \delta^2)}; \frac{\frac{1}{n_0} + \frac{1}{n_1} + \frac{p}{2\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)^2}{f_0(n_0-1, n_1, p, \delta^2) f_1(n_0, n_1-1, p, \delta^2)} \right) \quad (62)$$

Deleting terms tending to 0 under Raudys-Kolmogorov asymptotic conditions in (61) gives the approximation

$$E[(\hat{\epsilon}_0^l)^2] \approx \left[\Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{(n_0-1)} \right)}{\sqrt{1 + \frac{p}{\delta^2} \left(\frac{1}{(n_0-1)} + \frac{1}{n_1} \right)}} \right) \right]^2 \quad (63)$$

Equation (63) is simply the square of the approximation (34), 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 $W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1)$ by $W^{(1)}(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1)$ in (57). 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

$$E[\epsilon_0 \hat{\epsilon}_0^l] \approx \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0} \right)}{f_0(n_0, n_1, p, \delta^2)}, -\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0-1} \right)}{f_0(n_0-1, n_1, p, \delta^2)}; \frac{\frac{1}{n_1} + \frac{p}{2\delta^2} \left(\frac{1}{n_1^2} - \frac{1}{n_0^2} \right)}{f_0(n_0, n_1, p, \delta^2) f_0(n_0-1, n_1, p, \delta^2)} \right) \quad (64)$$

Similarly, it can be shown that

$$E[\epsilon_0 \hat{\epsilon}_1^l] \approx \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0} \right)}{f_0(n_0, n_1, p, \delta^2)}, -\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} - \frac{1}{n_1-1} \right)}{f_1(n_0, n_1-1, p, \delta^2)}; \frac{\frac{1}{n_1} + \frac{p}{2\delta^2} \left(\frac{1}{n_1^2} - \frac{1}{n_0^2} \right)}{f_0(n_0, n_1, p, \delta^2) f_1(n_0, n_1-1, p, \delta^2)} \right) \quad (65)$$

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 .

Deleting terms tending to 0 under Raudys-Kolmogorov asymptotic conditions in (64) gives the approximation

$$E[\epsilon_0 \hat{\epsilon}_0^l] \approx \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0} \right)}{\sqrt{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1} \right)}} \right) \Phi \left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0-1} \right)}{\sqrt{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0-1} + \frac{1}{n_1} \right)}} \right) \quad (66)$$

Equation (60) is simply the product of the approximations in (34) 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 Raudys-Kolmogorov asymptotic conditions.

B. Actual Classification Error

Note that the populations specific errors satisfy

$$\epsilon_0^2 = \left[\Phi \left(-\frac{\hat{G}_0}{\sqrt{\hat{D}_0}} \right) \right]^2, \epsilon_0 \epsilon_1 = \Phi \left(-\frac{\hat{G}_0}{\sqrt{\hat{D}_0}} \right) \Phi \left(\frac{\hat{G}_1}{\sqrt{\hat{D}_1}} \right), \epsilon_1^2 = \left[\Phi \left(\frac{\hat{G}_1}{\sqrt{\hat{D}_1}} \right) \right]^2 \quad (67)$$

where \hat{G}_i and \hat{D}_i were defined in (25). Using the results of Theorem 1, we obtain:

Theorem 3: Consider the sequence of Gaussian discrimination problems defined by (24). Then

$$\begin{aligned} \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_0^2] &= \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_0^2 = \left[\Phi \left(\frac{-G_0}{\sqrt{D}} \right) \right]^2, \\ \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_1^2] &= \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_1^2 = \left[\Phi \left(\frac{G_1}{\sqrt{D}} \right) \right]^2, \\ \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_0 \epsilon_1] &= \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_0 \epsilon_1 = \Phi \left(-\frac{G_0}{\sqrt{D_0}} \right) \Phi \left(\frac{G_1}{\sqrt{D_1}} \right), \end{aligned} \quad (68)$$

so that

$$\lim_{n_0, n_1, p \rightarrow \infty} E[\epsilon^2] = \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon^2 = \left(\lim_{n_0, n_1, p \rightarrow \infty} E[\epsilon] \right)^2 = \left[\alpha_0 \Phi \left(-\frac{G_0}{\sqrt{D}} \right) + \alpha_1 \Phi \left(\frac{G_1}{\sqrt{D}} \right) \right]^2 \quad (69)$$

where G_0 , G_1 and D are the same as in (29).

Theorem 3 suggests the following finite-sample approximation:

$$E[\epsilon_0^2] \approx \left[\Phi \left(-\frac{E[\hat{G}_0]}{\sqrt{E[\hat{D}_0]}} \right) \right]^2 = \left[\Phi \left(-\frac{E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \mathbf{x} \in \Pi_0]}{\sqrt{E[\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x} \in \Pi_0)]}} \right) \right]^2 \quad (70)$$

with similar approximations for $E[\epsilon_0 \epsilon_1]$ and $[\epsilon_1^2]$ derived from (68). These approximations are asymptotically exact, as shown by Theorem 3. Recalling (31), we see that (70) yields (52), showing that both (52) and (50) are asymptotically exact under the Raudys-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 (69) upon replacing λ_0 by p/n_0 and λ_1 by p/n_1 .

C. 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

$$\begin{aligned} \epsilon_{00}^r &= P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \leq 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_2) \leq 0 \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1) \\ \epsilon_{01}^r &= P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \leq 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1}) > 0 \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1) \\ \epsilon_{11}^r &= P(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1}) > 0, W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+2}) > 0 \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1) \end{aligned} \quad (71)$$

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 the representation of $E[(\hat{\epsilon}^r)^2]$ given in (14), it follows that

$$E[(\hat{\epsilon}^r)^2] = \frac{\hat{\alpha}_0^2}{n_0} E[\epsilon_{00}^r] + \frac{\hat{\alpha}_1^2}{n_1} E[\epsilon_{11}^r] + \hat{\alpha}_0^2 \frac{n_0 - 1}{n_0} E[\epsilon_{00}^r] + \hat{\alpha}_1^2 \frac{n_1 - 1}{n_1} E[\epsilon_{11}^r] + 2\hat{\alpha}_0 \hat{\alpha}_1 E[\epsilon_{01}^r] \quad (72)$$

where ϵ_0^r and ϵ_1^r are defined in (35). Let

$$\begin{aligned} \hat{H}_0^r &= \text{Cov}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1), W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_2) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1), \\ \hat{H}_{01}^r &= \text{Cov}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1), W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1}) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1), \\ \hat{H}_1^r &= \text{Cov}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+1}), W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_{n_0+2}) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1), \end{aligned} \quad (73)$$

Then

$$\epsilon_{00}^r = \Phi \left(\frac{-\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}; \frac{\hat{H}_0^r}{\hat{D}_0^r} \right), \quad \epsilon_{11}^r = \Phi \left(\frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}; \frac{\hat{H}_1^r}{\hat{D}_1^r} \right), \quad \epsilon_{01}^r = \Phi \left(\frac{-\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}; \frac{\hat{H}_0^r}{\hat{D}_0^r} \right) - \Phi \left(\frac{-\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}, \frac{-\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}; \frac{\hat{H}_{01}^r}{\hat{D}_{01}^r} \right), \quad (74)$$

where \hat{G}_i^r and \hat{D}_i^r were defined in (37).

Theorem 4: For the sequence of Gaussian discrimination problems defined by (24),

$$\begin{aligned} \text{ Klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_{00}^r] &= \text{ Klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_{01}^r] = \text{ Klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_{11}^r] \\ &= \text{ pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_{00}^r = \text{ pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_{01}^r = \text{ pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_{11}^r = \Phi\left(-\frac{G}{\sqrt{D}}; \frac{H}{D}\right) = \left[\Phi\left(-\frac{G}{\sqrt{D}}\right)\right]^2 \end{aligned} \quad (75)$$

and

$$\text{ Klim}_{n_0, n_1, p \rightarrow \infty} E[(\hat{\epsilon}^r)^2] = \left(\text{ Klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{\epsilon}^r]\right)^2 = \left[\Phi\left(-\frac{G}{\sqrt{D}}\right)\right]^2 \quad (76)$$

where G and D are given in (40) and $H = \text{ Klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{H}_0^r] = \text{ Klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{H}_1^r] = 0$.

Proof: See Appendix. ■

Theorem 4 suggests the following finite-sample approximation:

$$\begin{aligned} E[(\hat{\epsilon}_0^r)^2] &\approx \Phi\left(-\frac{E[\hat{G}_0^r]}{\sqrt{E[\hat{D}_0^r]}}; \frac{E[\hat{H}_0^r]}{E[\hat{D}_0^r]}\right) \\ &= \Phi\left(\frac{-E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1)]}{\sqrt{E[\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) | \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1)]}}; \frac{E[\text{Cov}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1), W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_2) | \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1)]}{E[\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) | \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1)]}\right) \end{aligned} \quad (77)$$

with corresponding approximations to $E[\hat{\epsilon}_0^r \hat{\epsilon}_1^r]$ and $E[(\hat{\epsilon}_1^r)^2]$ being obtained from (74). These approximations are asymptotically exact, as shown by Theorem 4. Eq. (77) yields

$$E[(\hat{\epsilon}_0^r)^2] \approx \Phi\left(-\frac{\delta}{2\sqrt{1 - \frac{1}{n_0}}} \sqrt{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1}\right)}; -\frac{1}{n_0 - 1}\right) \quad (78)$$

If one throws out extra terms that tend to zero under the Raudys-Kolmogorov limit, this reduces to (56), showing that both (56) and (54) are asymptotically exact under the Raudys-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 (76) 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 (16) 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,

$$\begin{aligned} E[\epsilon \hat{\epsilon}^r] &= \alpha_0 \hat{\alpha}_0 E\left[\Phi\left(\frac{-\hat{G}_0}{\sqrt{\hat{D}_0}}\right) \Phi\left(\frac{-\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}\right)\right] + \alpha_0 \hat{\alpha}_1 E\left[\Phi\left(\frac{-\hat{G}_0}{\sqrt{\hat{D}_0}}\right) \Phi\left(\frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}\right)\right] \\ &+ \alpha_1 \hat{\alpha}_0 E\left[\Phi\left(\frac{\hat{G}_1}{\sqrt{\hat{D}_1}}\right) \Phi\left(\frac{-\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}\right)\right] + \alpha_1 \hat{\alpha}_1 E\left[\Phi\left(\frac{\hat{G}_1}{\sqrt{\hat{D}_1}}\right) \Phi\left(\frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}\right)\right] \end{aligned} \quad (79)$$

where where \hat{G}_i and \hat{D}_i were defined in (25), and \hat{G}_i^r and \hat{D}_i^r were defined in (37). Using the results of Theorems 1 and 2, the following result immediately follows.

Theorem 5: For the sequence of Gaussian discrimination problems defined by (24),

$$\begin{aligned} \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_0 \hat{\epsilon}_0^r] &= \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_0 \hat{\epsilon}_1^r] = \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_0 \epsilon_0^r = \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_0 \epsilon_1^r = \Phi\left(\frac{-G_0}{\sqrt{D}}\right) \Phi\left(\frac{-G}{\sqrt{D}}\right) \\ \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_1 \hat{\epsilon}_0^r] &= \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_1 \hat{\epsilon}_1^r] = \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_1 \epsilon_0^r = \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_1 \epsilon_1^r = \Phi\left(\frac{G_1}{\sqrt{D}}\right) \Phi\left(\frac{-G}{\sqrt{D}}\right) \end{aligned} \quad (80)$$

so that

$$\text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon \hat{\epsilon}^r] = \left(\text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon] \right) \left(\text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{\epsilon}^r] \right) = \Phi\left(\frac{-G}{\sqrt{D}}\right) \left[\alpha_0 \Phi\left(\frac{-G_0}{\sqrt{D}}\right) + \alpha_1 \Phi\left(\frac{G_1}{\sqrt{D}}\right) \right] \quad (81)$$

where G_0 , G_1 , G and D are the same as in (29) and (40).

Theorem 5 suggests the following finite-sample approximation:

$$\begin{aligned} E[\epsilon_0 \hat{\epsilon}_0^r] &\approx \Phi\left(-\frac{E[\hat{G}_0]}{\sqrt{E[\hat{D}_0]}}\right) \Phi\left(-\frac{E[\hat{G}_0^r]}{\sqrt{E[\hat{D}_0^r]}}\right) \\ &= \Phi\left(\frac{-E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \mathbf{x} \in \Pi_0]}{\sqrt{E[\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x} \in \Pi_0])}}\right) \Phi\left(\frac{-E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1)]}{\sqrt{E[\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1])}}\right) \end{aligned} \quad (82)$$

with corresponding approximations to $E[\epsilon_0 \hat{\epsilon}_1^r]$, $E[\epsilon_1 \hat{\epsilon}_1^r]$, and $E[\epsilon_0 \hat{\epsilon}_1^r]$ being obtained from (79). By Theorem 5, these approximations are asymptotically exact. Eq. (82) yields

$$E[\epsilon_0 \hat{\epsilon}_0^r] \approx \Phi\left(-\frac{\delta}{2} \frac{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} - \frac{1}{n_0}\right)}{\sqrt{1 + \frac{p}{\delta^2} \left(\frac{1}{n_1} + \frac{1}{n_0}\right)}}\right) \Phi\left(-\frac{\delta}{2\sqrt{1 - \frac{1}{n_0}}} \sqrt{1 + \frac{p}{\delta^2} \left(\frac{1}{n_0} + \frac{1}{n_1}\right)}\right) \quad (83)$$

If one throws out extra terms that tend to zero under the Raudys-Kolmogorov limit, this reduces to (60), showing that both (60) and (58) are asymptotically exact under the Raudys-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 (81) upon replacing λ_0 by p/n_0 and λ_1 by p/n_1 .

D. Leave-one-out Error Estimator

In theorem 1, we showed that $\text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_0] = \Phi\left(-\frac{G_0}{\sqrt{D}}\right)$. It follows that

$$\text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{\epsilon}_{0, n_0}^l] = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_{0, n_0-1}] = \text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{\epsilon}_{0, n}^l = \Phi\left(-\frac{G_0}{\sqrt{D}}\right) \quad (84)$$

A similar fact applies to $\text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{\epsilon}_{1, n_1}^l]$. 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 [46]. Thus, we have the following result.

Theorem 6: For the sequence of Gaussian discrimination problems defined by (24),

$$\begin{aligned} \text{klim}_{n_0, n_1, p \rightarrow \infty} E[(\hat{\epsilon}_0^l)^2] &= \text{pklim}_{n_0, n_1, p \rightarrow \infty} (\hat{\epsilon}_0^l)^2 = \left[\Phi \left(-\frac{G_0}{\sqrt{D}} \right) \right]^2 \\ \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{\epsilon}_0^l \hat{\epsilon}_1^l] &= \text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{\epsilon}_0^l \hat{\epsilon}_1^l = \Phi \left(-\frac{G_0}{\sqrt{D_0}} \right) \Phi \left(\frac{G_1}{\sqrt{D_1}} \right) \\ \text{klim}_{n_0, n_1, p \rightarrow \infty} E[(\hat{\epsilon}_1^l)^2] &= \text{pklim}_{n_0, n_1, p \rightarrow \infty} (\hat{\epsilon}_1^l)^2 = \left[\Phi \left(\frac{G_1}{\sqrt{D}} \right) \right]^2 \end{aligned} \quad (85)$$

so that

$$\text{klim}_{n_0, n_1, p \rightarrow \infty} E[(\hat{\epsilon}^l)^2] = \text{pklim}_{n_0, n_1, p \rightarrow \infty} (\hat{\epsilon}^l)^2 = \left(\text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{\epsilon}^l] \right)^2 = \frac{1}{\lambda_0 + \lambda_1} \left[\lambda_0 \Phi \left(\frac{-G_0}{\sqrt{D}} \right) + \lambda_1 \Phi \left(\frac{G_1}{\sqrt{D}} \right) \right]^2 \quad (86)$$

where G_0 , G_1 and D are the same as in (29).

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 (63).

Figures 1–3 plot bias, variance, and RMS of resubstitution and leave-one-out using the asymptotically-exact approximations derived previously, with finite n and p . These fundamental error estimation performance measures are plotted as a function of total sample size. The balanced case $n_0 = n_1 = n$ is assumed throughout, with the x -axis representing the total sample size $2n$. Three curves are plotted in each case: the label “asym” identifies curves that use “complete” asymptotically-exact approximations, such as eqs. (50), (51), (54), (55), (58), (59), (61), (62), (64), and (65), where terms that tend to zero under Raudys-Kolmogorov conditions are not removed; the label “asym2” identifies curves that use “simplified” asymptotically-exact approximations, such as eqs. (52), (56), (60), (63), and (66), where terms that tend to zero under Raudys-Kolmogorov conditions are removed; finally, the label “MC” identifies curves obtained by Monte-Carlo approximation ($M = 5000$ MC sample sets were used to obtain each point in these curves). 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. These plots show that the “asym1” and “asym2” analytical approximations are very close to each other, indicating that the terms that tend to zero under Raudys-Kolmogorov conditions contribute very little, even at small n and p . They both show substantial agreement with the MC approximation, which indicates that the analytical approximations are accurate in finite-sample settings.

Figure 4 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, as the Mahalanobis distance will change with varying dimension. 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.

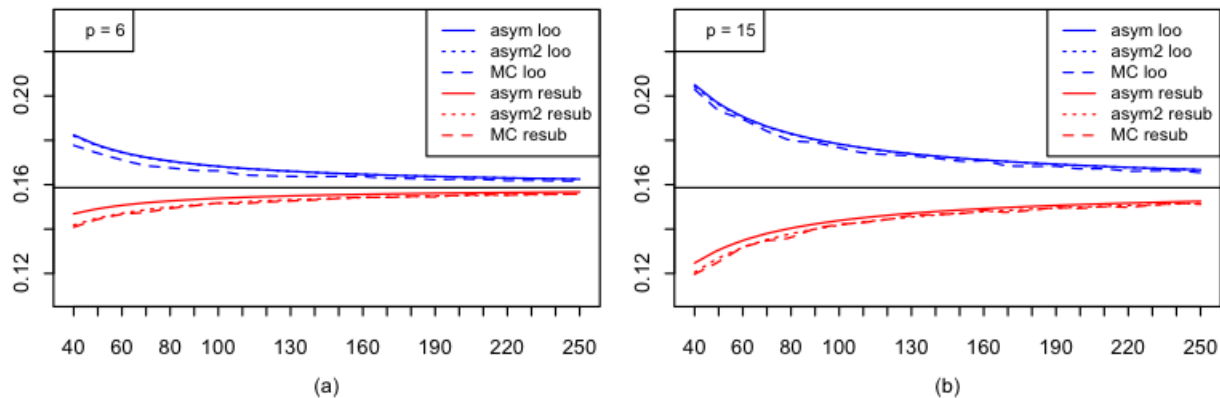


Fig. 1. Comparison of expectation for resubstitution and leave-one-out using the double asymptotic approximation with Monte Carlo estimates as a function of sample size for dimensions $p = 6$ and 15 (Bayes error 0.1586).

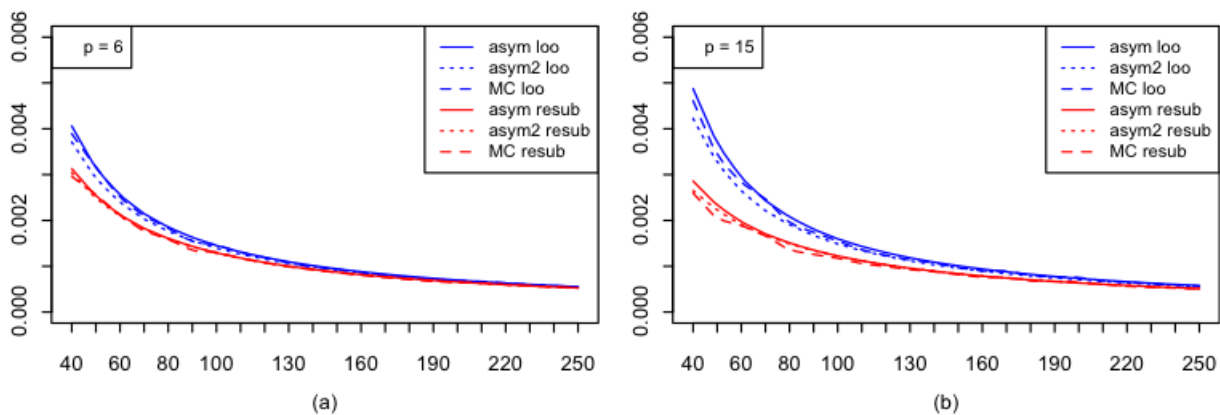


Fig. 2. Comparison of deviation variance for resubstitution and leave-one-out using the double asymptotic approximation with Monte Carlo estimates as a function of sample size for dimensions $p = 6$ and 15 (Bayes error 0.1586).

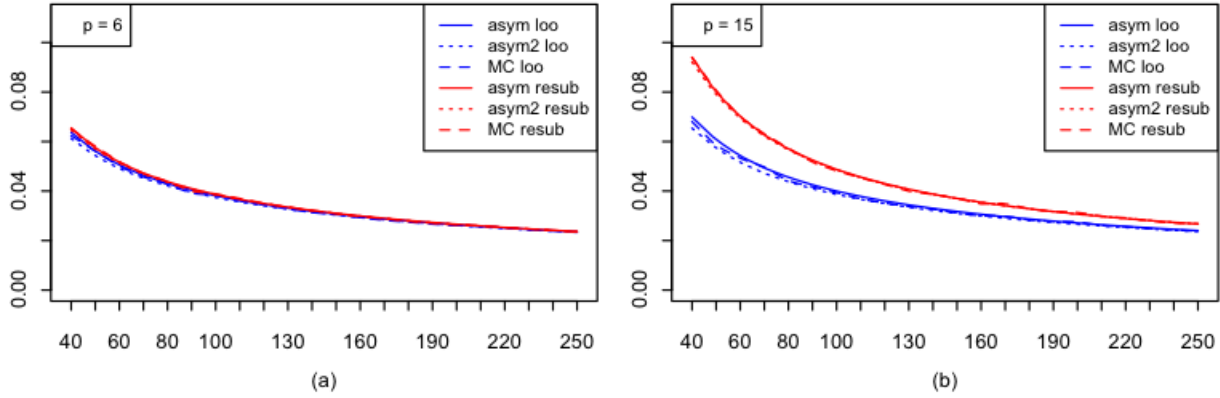


Fig. 3. Comparison of RMS for resubstitution and leave-one-out using the double asymptotic approximation with Monte Carlo estimates as a function of sample size for dimensions $p = 6$ and 15 (Bayes error 0.1586).

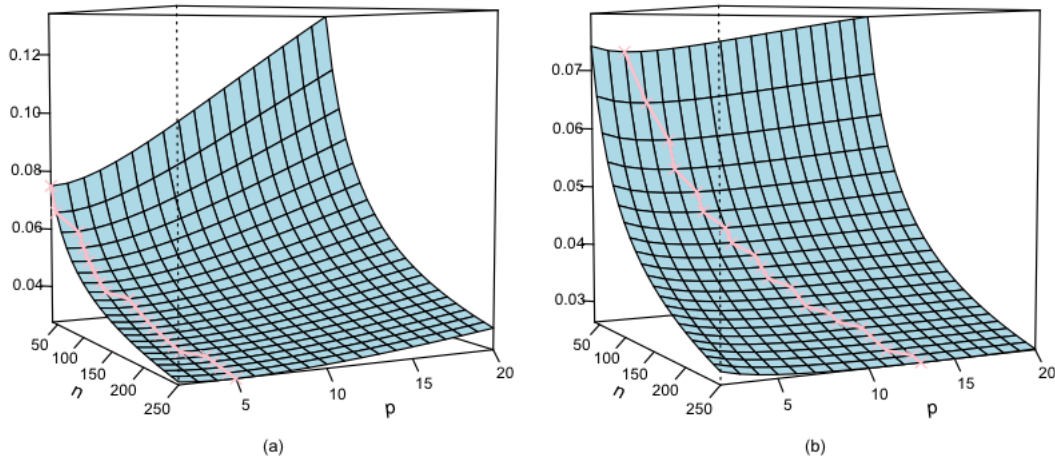


Fig. 4. Demonstration of RMS peaking phenomenon for $\rho = 0.3$: (a) resubstitution, (b) leave-one-out

VI. ASYMPTOTIC PERFORMANCE OF ERROR ESTIMATION

In this section we state the consequences of the Theorems 1–6 to the limiting values of bias, variance, and RMS of resubstitution and leave-one-out error estimators under Raudys-Kolmogorov asymptotic conditions.

From Theorems 1 and 2, we conclude that the asymptotic bias of resubstitution is given by (for the sake of simplicity, we consider here the asymptotically balanced case $\lambda_1 = \lambda_0 = \lambda$):

$$\lim_{n_0, n_1, p \rightarrow \infty} \text{Bias}[\hat{\epsilon}^r] = \Phi\left(-\frac{\delta}{2} \sqrt{1 + \frac{2\lambda}{\delta^2}}\right) - \Phi\left(-\frac{\delta}{2} \frac{1}{\sqrt{1 + \frac{2\lambda}{\delta^2}}}\right) < 0 \quad (87)$$

Therefore, resubstitution has an optimistic asymptotic bias. Recalling that under the Raudys-Kolmogorov limit we have $n_0/p, n_1/p \rightarrow 1/\lambda$, 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 $\text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon] = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{\epsilon}^r] = \frac{1}{2}$. As for the asymptotic bias of leave-one-out, since $E[\hat{\epsilon}_{i, n_i}^l] = E[\epsilon_{i, n_i-1}]$, for $i = 0, 1$, $\text{klim}_{n_0, n_1, p \rightarrow \infty} \text{Bias}[\hat{\epsilon}^l] = 0$. This is true also in the unbalanced case $\lambda_0 \neq \lambda_1$.

A consequence of Theorems 1–6 is that all variances and covariances are asymptotically zero,

$$\text{klim}_{n_0, n_1, p \rightarrow \infty} \text{Var}(\epsilon) = \text{klim}_{n_0, n_1, p \rightarrow \infty} \text{Var}(\hat{\epsilon}^r) = \text{klim}_{n_0, n_1, p \rightarrow \infty} \text{Var}(\hat{\epsilon}^l) = \text{klim}_{n_0, n_1, p \rightarrow \infty} \text{Cov}(\epsilon, \hat{\epsilon}^r) = \text{klim}_{n_0, n_1, p \rightarrow \infty} \text{Cov}(\epsilon, \hat{\epsilon}^l) = 0, \quad (88)$$

which also holds when $\lambda_0 \neq \lambda_1$. Thus, the deviation variances are also asymptotically null:

$$\text{klim}_{n_0, n_1, p \rightarrow \infty} \text{Var}_d[\hat{\epsilon}^r] = \text{klim}_{n_0, n_1, p \rightarrow \infty} \text{Var}_d[\hat{\epsilon}^l] = 0 \quad (89)$$

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

VII. ASYMPTOTIC 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 is 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 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 distribution-free bounds for some classification rules [13], [14], [35], but these bounds tend to be very loose and of limited practical value.

In considering RMS bounds using the RMS expressions developed herein, one must keep in mind that the expressions are asymptotic. In intensive simulation studies, we have observed 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 the Bayes error is between 0.3 and 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 Raudys-Kolmogorov asymptotic conditions are inherently suitable for cases where the dimension is comparable to the sample size.

A. Derivation and Sample Size Calculation

Let the desired bound be $\kappa_{\hat{\epsilon}}(n, p) = \max_{0 \leq \epsilon_{bay} \leq 0.5} \text{RMS}[\hat{\epsilon}]$, where $\hat{\epsilon} = \hat{\epsilon}^r$ or $\hat{\epsilon}^l$, and $n = n_0 = n_1$, for simplicity. From Theorems 1–6, one can obtain a relatively simple analytical expression for $\text{klim}_{n, p \rightarrow \infty} \text{RMS}[\hat{\epsilon}]$, which can be readily verified (by computation of the derivative) to be a decreasing function of the Mahalanobis distance δ , and therefore an increasing function of the Bayes error, $\epsilon_{bay} = \Phi(-\delta/2)$, for fixed λ_0 and λ_1 . Therefore, we have that $\kappa_{\hat{\epsilon}}(n, p) \approx \lim_{\delta \rightarrow 0} \text{klim}_{n, p \rightarrow \infty} \text{RMS}[\hat{\epsilon}]$. Letting $\delta \rightarrow 0$ in our analytical expression for $\lim_{\delta \rightarrow 0} \text{klim}_{n, p \rightarrow \infty} \text{RMS}[\hat{\epsilon}]$ yields therefore an approximate bound for finite samples:

$$\text{RMS}[\hat{\epsilon}^r] \leq \kappa_{\hat{\epsilon}^r}(n, p) \approx \sqrt{\frac{1}{4} + \left(\frac{1}{2n} - 1\right) \left(\Phi\left(-\sqrt{\frac{p}{2n}}\right) - \left[\Phi\left(-\sqrt{\frac{p}{2n}}\right) \right]^2 \right)} \quad (90)$$

$$\text{RMS}[\hat{\epsilon}^l] \leq \kappa_{\hat{\epsilon}^l}(n, p) \approx \sqrt{\frac{1}{8n} + \Phi\left(-\sqrt{\frac{p}{8n^3}}, -\sqrt{\frac{p}{8n^3}}; \frac{1}{n}\right) - \frac{1}{8}} \quad (91)$$

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

It can be seen from (90) and (91) 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 (91) are functions of $\sqrt{p/n^3}$, whereas the corresponding terms in (90) are functions of $\sqrt{p/n}$. This difference in sensitivity to dimension between resubstitution and leave-one-out is not specific to the bound $\kappa_{\hat{\epsilon}}(n, p)$ but holds for the RMS in the whole range of δ^2 . This fact can be seen in the shape of the surfaces in Fig 4.

To find the necessary number of samples to insure a given RMS, one can find the minimum n to satisfy (90) and (91). Table I shows the minimum number of sample points needed for resubstitution and leave-one-out to achieve a given value of $\kappa_{\hat{\epsilon}}(n, p)$. 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 (90) and (91), 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 $\kappa_{\hat{\epsilon}}(n, p)$, which has been obtained under the assumption of a known covariance matrix. This has been accomplished by using the required sample sizes in Table I in a Monte-Carlo estimation of $\kappa_{\hat{\epsilon}}(n, p)$ when the covariance matrix is estimated from the data ($M = 10000$ MC sample sets were used). The results are shown in Table I by the values in parentheses. Comparing these values with the given values of $\kappa_{\hat{\epsilon}}(n, p)$ on the left-hand side of the table reveals that (90) and (91), and the sample sizes determined therefrom, can be reliably used in practice. Table I shows 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. 4. 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 I we see that the accuracy of the results improves as the dimension increases.

TABLE I

MINIMUM SAMPLE SIZE, n , ($n_0 = n_1 = n$) FOR DESIRED $\kappa_{\hat{\epsilon}}(n, p)$. THE VALUES IN PARENTHESES ARE THE MONTE-CARLO ESTIMATES OF $\kappa_{\hat{\epsilon}}(n, p)$ WHEN COVARIANCE MATRIX IS ESTIMATED FROM DATA ($M = 10000$ MC SAMPLE SETS).

$\kappa_{\hat{\epsilon}}(n, p)$	resub					loo	
	$p=2$	$p=3$	$p=4$	$p=6$	$p=10$	$p=3$	$p=10$
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)

B. Comparison with Distribution-Free Bounds

It is illuminating to see how much tighter the RMS bounds based upon an exact (or asymptotic) representation of the RMS can be as compared to using a distribution-free approach, which has been the classical way of attacking the problem. We cannot make this comparison directly because heretofore no distribution-free RMS bounds for LDA have been published, to the best of the authors' knowledge. Consequently, to get some sort of comparison of the magnitudes, we consider the following previously known distribution-free bound for the histogram classification rule and leave-one-out [35]:

$$\text{RMS}[\hat{\epsilon}^l] \leq \sqrt{\frac{1 + 6e^{-1}}{2n} + \frac{6}{\sqrt{\pi(2n-1)}}} \quad (92)$$

The number of sample points determined from (92) to have $\text{RMS} \leq 0.08$ is approximately $n = 140000$; this magnitude is typical of sample size calculations made from ditribution-free bounds. By comparison,

$n = 38$ sample points for $p = 10$ are obtained from Table I for the LDA rule under the Gaussian assumption and leave-one-out.

C. Gene-Expression Classification Example

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 [47]. Discrimination is between good versus bad prognosis. For this experiment, $p = 3$ transcripts are selected. From Table I, to have $\kappa_{\hat{\epsilon}^r}(n, p = 3) \approx 0.1$, we need $n_0 = n_1 = 36$, which also makes $\kappa_{\hat{\epsilon}^l}(n, p = 3) < 0.1$. We selected randomly 36 sample points from each class, and 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 three genes Contig28552_RC, NM_003981 and NM_020188 ($p = 3$), shown to be close to Gaussian by the Shapiro-Wilk test and to have close to equal covariance matrices between classes by Box's M test. 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 5 shows the designed classifier. This example demonstrates how one can use Table I and combine it with the proper assumptions to get to a reliable estimation of the true error.

VIII. CONCLUSION

Using the double asymptotic method of Raudys-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 significantly more difficult. Obtaining corresponding results with unknown covariance matrix is the next logical step in the line of this research.

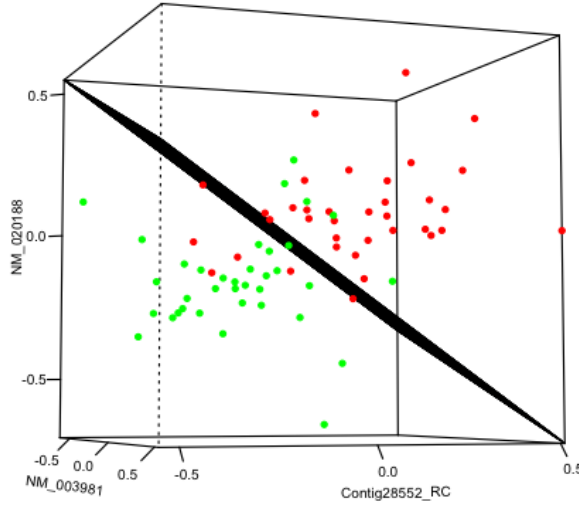


Fig. 5. The designed classifier for good-prognosis (green) vs. bad-prognosis (red) using the minimum number of samples to get a given RMS. The three genes selected are Contig28552_RC, NM_003981 and NM_020188.

IX. ACKNOWLEDGEMENTS

The authors acknowledge the support of the National Science Foundation, through NSF awards CCF-0845407 (Braga-Neto) and CCF-0634794 (Dougherty).

APPENDIX A

PROOF OF THEOREM 1

Since the classification error ϵ is invariant to any linear transformation, we can use the canonical convenient form proposed by [48], with $\Sigma = \mathbf{I}$ and $\mu_1 = -\mu_0 = (\frac{\delta}{2}, 0, \dots, 0)^T$.

We prove that $\text{Var}(\hat{G}_0) \xrightarrow{K} 0$. Let $\mathbf{v}(i)$ denote the i -th component of vector \mathbf{v} . We have

$$\text{Var}(\hat{G}_0) = \text{Var}(E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x} \in \Pi_0]) = \text{Var}\left(\left(-\frac{\delta}{2} - \bar{\mathbf{x}}(1)\right) \bar{\mathbf{a}}(1) - \sum_{i=2}^p \bar{\mathbf{x}}(i) \mathbf{a}(i)\right) \quad (93)$$

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

$$\mu_{\bar{\mathbf{x}}} = (0, \dots, 0), \mu_{\bar{\mathbf{a}}} = (-\delta, 0, \dots, 0), \Sigma_{\bar{\mathbf{x}}} = \frac{1}{4} \left(\frac{1}{n_0} + \frac{1}{n_1} \right) \mathbf{I}_p, \Sigma_{\bar{\mathbf{a}}} = \left(\frac{1}{n_0} + \frac{1}{n_1} \right) \mathbf{I}_p \quad (94)$$

Given the independence of the vector components, and using the results of [49] to find the variance of a product of non central Gaussian vectors, algebraic manipulation leads to:

$$\text{Var}(\hat{G}_0) = \frac{\delta^2}{n_1} + \frac{p}{2} \left(\frac{1}{n_0^2} + \frac{1}{n_1^2} \right) \xrightarrow{K} 0 \quad (95)$$

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

$$\begin{aligned} \text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{G}_0 &= \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{G}_0] = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \mathbf{x} \in \Pi_0] \\ &= \text{klim}_{n_0, n_1, p \rightarrow \infty} \left[\frac{\delta^2}{2} + \frac{p}{2} \left(\frac{1}{n_1} - \frac{1}{n_0} \right) \right] = \frac{1}{2} (\delta^2 + \lambda_1 - \lambda_0) \triangleq G_0 \end{aligned} \quad (96)$$

An analogous argument shows that $\text{Var}(\hat{G}_1) \xrightarrow{K} 0$ and

$$\text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{G}_1 = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{G}_1] = \frac{1}{2} (\delta^2 + \lambda_0 - \lambda_1) \triangleq G_1 \quad (97)$$

Now we prove that $\text{Var}(\hat{D}_0) \xrightarrow{K} 0$. We have

$$\hat{D}_0 = \text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x} \in \Pi_0) = \bar{\mathbf{a}}^T \boldsymbol{\Sigma}_{\mathbf{x}} \bar{\mathbf{a}} = \bar{\mathbf{a}}^T \bar{\mathbf{a}} = \hat{\delta}^2 \quad (98)$$

since $\boldsymbol{\Sigma}_{\mathbf{x}} = \mathbf{I}_p$, where $\bar{\mathbf{a}}$ is defined as before and $\hat{\delta}^2 = (\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)^T (\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)$. Notice that

$$\frac{\hat{\delta}^2}{\frac{1}{n_0} + \frac{1}{n_1}} \sim \chi_1^2 \left(\frac{\delta^2}{\frac{1}{n_0} + \frac{1}{n_1}} \right) + \chi_{p-1}^2 \quad (99)$$

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

$$\text{Var}(\hat{D}_0) = \left(\frac{1}{n_0} + \frac{1}{n_1} \right)^2 \left[\text{Var} \left(\chi_1^2 \left(\frac{\delta^2}{\frac{1}{n_0} + \frac{1}{n_1}} \right) \right) + \text{Var}(\chi_{p-1}^2) \right] = 4\delta^2 \left(\frac{1}{n_0} + \frac{1}{n_1} \right) + 2p \left(\frac{1}{n_0} + \frac{1}{n_1} \right)^2 \xrightarrow{K} 0 \quad (100)$$

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

$$\begin{aligned} \text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{D}_0 &= \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{D}_0] = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\text{Var}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x} \in \Pi_0)] \\ &= \text{klim}_{n_0, n_1, p \rightarrow \infty} \left[\delta^2 + p \left(\frac{1}{n_0} + \frac{1}{n_1} \right) \right] = \delta^2 + \lambda_0 + \lambda_1 \triangleq D \end{aligned} \quad (101)$$

An analogous argument shows that $\text{Var}(\hat{D}_1) \xrightarrow{K} 0$ and $\text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{D}_1 = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{D}_1] = \delta^2 + \lambda_0 + \lambda_1 = D$.

By using the Continuous Mapping Theorem [46], it follows that

$$\begin{aligned} \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_0 &= \text{pklim}_{n_0, n_1, p \rightarrow \infty} \Phi \left(-\frac{\hat{G}_0}{\sqrt{\hat{D}_0}} \right) = \Phi \left(\text{pklim}_{n_0, n_1, p \rightarrow \infty} -\frac{\hat{G}_0}{\sqrt{\hat{D}_0}} \right) = \Phi \left(-\frac{G_0}{\sqrt{D}} \right) \\ \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_1 &= \text{pklim}_{n_0, n_1, p \rightarrow \infty} \Phi \left(\frac{\hat{G}_1}{\sqrt{\hat{D}_1}} \right) = \Phi \left(\text{pklim}_{n_0, n_1, p \rightarrow \infty} \frac{\hat{G}_1}{\sqrt{\hat{D}_1}} \right) = \Phi \left(\frac{G_1}{\sqrt{D}} \right) \end{aligned} \quad (102)$$

Boundedness and continuity of Φ allows one to apply the Helly-Bray Theorem [50] to obtain

$$\begin{aligned} \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_0] &= \text{klim}_{n_0, n_1, p \rightarrow \infty} E \left[\Phi \left(-\frac{\hat{G}_0}{\sqrt{\hat{D}_0}} \right) \right] = E \left[\Phi \left(\text{pklim}_{n_0, n_1, p \rightarrow \infty} -\frac{\hat{G}_0}{\sqrt{\hat{D}_0}} \right) \right] = \Phi \left(\frac{-G_0}{\sqrt{D}} \right) \\ \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_1] &= \text{klim}_{n_0, n_1, p \rightarrow \infty} E \left[\Phi \left(\frac{\hat{G}_1}{\sqrt{\hat{D}_1}} \right) \right] = E \left[\Phi \left(\text{pklim}_{n_0, n_1, p \rightarrow \infty} \frac{\hat{G}_1}{\sqrt{\hat{D}_1}} \right) \right] = \Phi \left(\frac{G_1}{\sqrt{D}} \right) \end{aligned} \quad (103)$$

APPENDIX B

PROOF OF THEOREM 2

Using the linear transformation introduced in the proof of Theorem 1, we first transform the data to normal distributions with $\Sigma = \mathbf{I}$ and $\mu_1 = -\mu_0 = (\frac{\delta}{2}, 0, 0, \dots, 0)^T$.

We prove that $\text{Var}(\hat{G}_0^r) \xrightarrow{K} 0$ and $\text{Var}(\hat{D}_0^r) \xrightarrow{K} 0$. Notice that the random vector $(\mathbf{x}_1^T, \bar{\mathbf{x}}_0^T, \bar{\mathbf{x}}_1^T)$ has a multivariate normal distribution with mean vector $(\mu_0^T, \mu_0^T, \mu_1^T)$ and covariance matrix

$$\begin{pmatrix} \mathbf{I} & \frac{\mathbf{I}}{n_0} & 0 \\ \frac{\mathbf{I}}{n_0} & \frac{\mathbf{I}}{n_0} & 0 \\ 0 & 0 & \frac{\mathbf{I}}{n_1} \end{pmatrix} \quad (104)$$

Using properties of the multivariate normal distribution [51], we conclude that

$$\mathbf{x}_1 \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1 \sim N\left(\bar{\mathbf{x}}_0, \left(1 - \frac{1}{n_0}\right)\mathbf{I}\right) \quad (105)$$

From this it follows easily that

$$\left(\mathbf{x}_1 - \frac{\bar{\mathbf{x}}_0 + \bar{\mathbf{x}}_1}{2}\right)^T (\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1 \sim N\left(\frac{\hat{\delta}^2}{2}, \left(1 - \frac{1}{n_0}\right)\hat{\delta}^2\right) \quad (106)$$

in which $\hat{\delta}^2 = (\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)^T (\bar{\mathbf{x}}_0 - \bar{\mathbf{x}}_1)$. Hence, to show that $\text{Var}(\hat{G}_0^r) \xrightarrow{K} 0$ and $\text{Var}(\hat{D}_0^r) \xrightarrow{K} 0$, all we need to do is to show that $\text{Var}(\hat{\delta}^2) \xrightarrow{K} 0$. As we proved (100) using (99), it follows that

$$\text{Var}(\hat{\delta}^2) = 4\delta^2 \left(\frac{1}{n_0} + \frac{1}{n_1}\right) + 2p \left(\frac{1}{n_0} + \frac{1}{n_1}\right)^2 \xrightarrow{K} 0 \quad (107)$$

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

$$\text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{G}_0^r = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{G}_0^r] = \text{klim}_{n_0, n_1, p \rightarrow \infty} \left[\frac{\delta^2}{2} + \frac{p}{2} \left(\frac{1}{n_0} + \frac{1}{n_1}\right) \right] = \frac{1}{2}(\delta^2 + \lambda_0 + \lambda_1) \triangleq G \quad (108)$$

$$\text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{D}_0^r = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{D}_0^r] = \text{klim}_{n_0, n_1, p \rightarrow \infty} \left[\left(1 - \frac{1}{n_0}\right) \left(\delta^2 + p \left(\frac{1}{n_0} + \frac{1}{n_1}\right)\right) \right] = \delta^2 + \lambda_0 + \lambda_1 \triangleq D \quad (109)$$

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

$$\text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{G}_1^r = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{G}_1^r] = -\frac{1}{2}(\delta^2 + \lambda_0 + \lambda_1) = -G,$$

$$\text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{D}_1^r = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{D}_1^r] = \delta^2 + \lambda_0 + \lambda_1 = D$$

The rest of the proof proceeds much as in the case of the proof of Theorem 1. By using the Continuous Mapping Theorem [46], it follows that

$$\begin{aligned} \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_0^r &= \text{pklim}_{n_0, n_1, p \rightarrow \infty} \Phi\left(-\frac{\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}\right) = \Phi\left(\text{pklim}_{n_0, n_1, p \rightarrow \infty} -\frac{\hat{G}_0^r}{\sqrt{\hat{D}_0^r}}\right) = \Phi\left(-\frac{G}{\sqrt{D}}\right) \\ \text{pklim}_{n_0, n_1, p \rightarrow \infty} \epsilon_1^r &= \text{pklim}_{n_0, n_1, p \rightarrow \infty} \Phi\left(\frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}\right) = \Phi\left(\text{pklim}_{n_0, n_1, p \rightarrow \infty} \frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}}\right) = \Phi\left(-\frac{G}{\sqrt{D}}\right) \end{aligned} \quad (110)$$

Boundedness and continuity of Φ allows one to apply the Helly-Bray Theorem [50] to obtain

$$\begin{aligned} \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{\epsilon}_0^r] &= \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_0^r] = \text{klim}_{n_0, n_1, p \rightarrow \infty} E \left[\Phi \left(\frac{-\hat{G}_0^r}{\sqrt{\hat{D}_0^r}} \right) \right] = E \left[\Phi \left(\text{pklim}_{n_0, n_1, p \rightarrow \infty} \frac{-\hat{G}_0^r}{\sqrt{\hat{D}_0^r}} \right) \right] = \Phi \left(\frac{-G}{\sqrt{D}} \right) \\ \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{\epsilon}_1^r] &= \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\epsilon_1^r] = \text{klim}_{n_0, n_1, p \rightarrow \infty} E \left[\Phi \left(\frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}} \right) \right] = E \left[\Phi \left(\text{pklim}_{n_0, n_1, p \rightarrow \infty} \frac{\hat{G}_1^r}{\sqrt{\hat{D}_1^r}} \right) \right] = \Phi \left(\frac{-G}{\sqrt{D}} \right) \end{aligned} \quad (111)$$

From this it also follows that

$$\text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{\epsilon}^r] = \text{klim}_{n_0, n_1, p \rightarrow \infty} (\hat{\alpha}_0 E[\hat{\epsilon}_0^r] + \hat{\alpha}_1 E[\hat{\epsilon}_1^r]) = \frac{\lambda_1}{\lambda_0 + \lambda_1} \Phi \left(\frac{-G}{\sqrt{D}} \right) + \frac{\lambda_0}{\lambda_0 + \lambda_1} \Phi \left(\frac{-G}{\sqrt{D}} \right) = \Phi \left(\frac{-G}{\sqrt{D}} \right)$$

APPENDIX C

PROOF OF THEOREM 4

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

$$\begin{aligned} \text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{G}_0^r &= \frac{1}{2}(\delta^2 + \lambda_0 + \lambda_1) = G, \\ \text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{G}_1^r &= -\frac{1}{2}(\delta^2 + \lambda_0 + \lambda_1) = -G, \\ \text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{D}_0^r &= \text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{D}_1^r = \delta^2 + \lambda_0 + \lambda_1 = D \end{aligned} \quad (112)$$

We now prove that $\text{Var}(\hat{H}_0^r) \xrightarrow{K} 0$. Similarly to the proof of Theorem 2 and the way (105) was obtained, it is possible to show that

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1 \sim N \left(\begin{bmatrix} \bar{\mathbf{x}}_0 \\ \bar{\mathbf{x}}_0 \end{bmatrix}, \begin{bmatrix} \left(1 - \frac{1}{n_0}\right) \mathbf{I} & -\frac{1}{n_0} \mathbf{I} \\ -\frac{1}{n_0} \mathbf{I} & \left(1 - \frac{1}{n_0}\right) \mathbf{I} \end{bmatrix} \right) \quad (113)$$

It follows that

$$\hat{H}_0^r = \text{Cov}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1), W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_2) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1) = -\frac{1}{n_0} \hat{\delta}^2 \quad (114)$$

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

$$\begin{aligned} \text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{H}_0^r &= \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{H}_0^r] = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\text{Cov}(W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_1), W(\bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1, \mathbf{x}_2) \mid \bar{\mathbf{x}}_0, \bar{\mathbf{x}}_1)] \\ &= \text{klim}_{n_0, n_1, p \rightarrow \infty} -\frac{1}{n_0} E[\hat{\delta}^2] = \text{klim}_{n_0, n_1, p \rightarrow \infty} \left[-\frac{\delta^2}{n_0} - \frac{p}{2} \left(\frac{1}{n_0^2} + \frac{1}{n_0 n_1} \right) \right] = 0 \end{aligned}$$

An analogous argument shows that $\text{Var}(\hat{H}_1^r) \xrightarrow{K} 0$ and $\text{pklim}_{n_0, n_1, p \rightarrow \infty} \hat{H}_1^r = \text{klim}_{n_0, n_1, p \rightarrow \infty} E[\hat{H}_1^r] = 0$. The rest of the proof proceeds as in the case of the proofs of Theorem 1 and 2.

REFERENCES

- [1] R. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugen.*, vol. 7, pp. 179–188, 1936.
- [2] —, "The precision of discriminant function," *Ann. Eugen.*, vol. 10, pp. 422–429, 1940.
- [3] A. Wald, "On a statistical problem arising in the classification of an individual into one of two groups," *Ann. Math. Statist.*, vol. 15, pp. 145–162, 1944.
- [4] T. Anderson, "Classification by multivariate analysis," *Psychometrika*, vol. 16, pp. 31–50, 1951.
- [5] G. T. Toussaint, "Bibliography on estimation of misclassification," *IEEE Transactions on Information Theory*, vol. 20, pp. 472–479, 1974.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] N. Glick, "Additive estimators for probabilities of correct classification," *Pattern recognition*, vol. 10, pp. 211–222, 1978.
- [10] K. Fukunaga and R. R. Hayes, "Estimation of classifier performance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 1087–1101, 1989.
- [11] U. Braga-Neto and E. Dougherty, "Is cross-validation valid for microarray classification?" *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.
- [12] S. Raudys, *Statistical and Neural Classifiers, An Integrated Approach to Design*. London: Springer-Verlag, 2001.
- [13] 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.
- [14] —, "Distribution-free performance bounds for potential function rules," *IEEE Transactions on Information Theory*, vol. 25, pp. 601–604, 1979.
- [15] D. Foley, "Considerations of sample and feature size," *IEEE Transactions on Information Theory*, vol. IT-18, pp. 618–626, 1972.
- [16] A. Zollanvari, U. Braga-Neto, and E. Dougherty, "Joint sampling distribution between actual and estimated classification errors for linear discriminant analysis," *IEEE Transaction on Information Theory*, vol. 56, no. 2, pp. 784–804, 2010.
- [17] V. Serdobolskii, *Multivariate Statistical Analysis: A High-Dimensional Approach*. Springer, 2000.
- [18] S. Raudys, "On the amount of a priori information in designing the classification algorithm," *Technical Cybernetics*, vol. 4, pp. 168–174, 1972, in Russian.
- [19] A. Deev, "Representation of statistics of discriminant analysis and asymptotic expansion when space dimensions are comparable with sample size," *Dokl. Akad. Nauk SSSR*, vol. 195, pp. 759–762, 1970, in Russian.
- [20] —, "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.
- [21] S. Raudys, "Comparison of the estimates of the probability of misclassification," in *Proc. International Joint Conference on Pattern Recognition*, 1978, pp. 280–282.
- [22] 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.
- [23] I. Shmulevich and E. R. Dougherty, *Genomic Signal Processing*. Princeton: Princeton University Press, 2007.

- [24] K. Umapathy, S. Krishnan, and R. K. Rao, "Audio signal feature extraction and classification using local discriminant bases," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1236–1246, 2007.
- [25] D. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 18, pp. 891–896, 1996.
- [26] W. Zheng, "Heteroscedastic feature extraction for texture classification," *IEEE Signal processing letters*, vol. 16, pp. 766–769, 2009.
- [27] 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.
- [28] 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.
- [29] T. V. Bandos, L. Bruzzone, and G. Camps-Valls, "Finger-based personal authentication: a comparison of feature-extraction methods based on principal component analysis, most discriminant features and regularised-direct linear discriminant analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, pp. 862–873, 2009.
- [30] S. John, "Errors in discrimination," *Ann. Math. Statist.*, vol. 32, pp. 1125–1144, 1961.
- [31] M. J. Sorum, "Estimating the conditional probability of misclassification," *Technometrics*, vol. 13, pp. 333–343, 1971.
- [32] M. Moran, "On the expectation of errors of allocation associated with a linear discriminant function," *Biometrika*, vol. 62, no. 1, pp. 141–148, 1975.
- [33] C. Smith, "Some examples of discrimination," *Annals of Eugenics*, vol. 18, pp. 272–282, 1947.
- [34] P. Lachenbruch and M. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, pp. 1–11, 1968.
- [35] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer, 1996.
- [36] 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.
- [37] L. D. Meshalkin and V. I. Serdobolskii, "Errors in the classification of multi-variate observations," *Theory of probability and its applications*, vol. 23, pp. 741–750, 1978.
- [38] 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.
- [39] 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.
- [40] H. S. D. Haussler, M. Kearns and N. Tishby, "Rigorous learning curves from statistical mechanics," in *Proceedings of the Seventh Annual ACM Conference on Computer Learning Theory*. Morgan Kaufman, San Mateo, CA, 1994.
- [41] 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.
- [42] V. Pikelis, "Comparison of methods of computing the expected classification errors," *Automat. Remote Control*, vol. 5, no. 7, pp. 59–63, 1976.
- [43] S. Raudys, "On determining training sample size of a linear classifier," *Computer Systems*, vol. 28, pp. 79–87, 1967, in Russian.
- [44] 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.

- [45] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, "Optimal number of features as a function of sample size for various classification rules," *Bioinformatics*, vol. 21, no. 8, pp. 1509–1515, 2005.
- [46] P. Billingsley, *Probability and Measure*, 3rd ed. New York: Wiley, 1995.
- [47] Y. H. M. van de Vijver 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.
- [48] O. J. Dunn, "Some expected values for probabilities of correct classification in discriminant analysis," *Technometrics*, vol. 13, pp. 345–353, 1971.
- [49] R. Kan, "From moments of sum to moments of product," *Journal of Multivariate Analysis*, vol. 99, pp. 542 – 554, 2008.
- [50] P. Sen and J. Singer, *Large Sample Methods in Statistics*. New York: Chapman and Hall, 1993.
- [51] T. W. Anderson, *Introduction to Multivariate Statistical Analysis*. New York: Wiley, 1958.



Amin Zollanvari received B.S. and M.S. degrees in electrical engineering from Shiraz University, Iran, in 2003 and 2006. He received the Ph.D. degree in Electrical and Computer Engineering from Texas A&M University, College Station, TX, in 2010. He is currently a post-doctoral research fellow at Children's Hospital Informatics Program at Harvard-MIT Division of Health Science, Brigham and Women's Hospital, and Harvard Medical School, Boston, MA. His research interests include pattern recognition, bioinformatics, and ontology engineering.



Ulisses M. Braga-Neto received the Ph.D. degree in electrical and computer engineering from The Johns Hopkins University, Baltimore, MD, in 2002. He held a post-doctoral fellowship in the section of clinical cancer genetics at the University of Texas M.D. Anderson Cancer Center, Houston, from 2002 to 2004. In 2004, he joined the Virology and Experimental Therapy Laboratory at the Oswaldo Cruz Foundation (FIOCRUZ), in Recife, Brazil. Since January 2007, he has been an assistant professor and member of the Genomic Signal Processing Laboratory at the Department of Electrical and Computer Engineering of Texas A&M University, College Station, TX. He received an NSF CAREER Award in 2008. His research interests include small-sample error estimation, statistical pattern recognition, and genomic signal processing, with applications in the study of cancer and infectious diseases.



Edward R. Dougherty is a Professor in the Department of Electrical and Computer Engineering at Texas A&M University in College Station, TX, where he holds the Robert M. Kennedy 26 Chair in Electrical Engineering and is Director of the Genomic Signal Processing Laboratory. He is also co-Director of the Computational Biology Division of the Translational Genomics Research Institute in Phoenix, AZ, and is an Adjunct Professor in the Department of Bioinformatics and Computational Biology at the University of Texas M. D. Anderson Cancer Center in Houston, TX. He holds a Ph.D. in mathematics from Rutgers University and an M.S. in Computer Science from Stevens Institute of Technology, and has been awarded the Doctor Honoris Causa by the Tampere University of Technology in Finland. He is a fellow of SPIE, has received the SPIE Presidents Award, and served as the editor of the SPIE/IS&T Journal of Electronic Imaging.