

Exact Representation of the Second-order Moments for Resubstitution and Leave-one-out Error Estimation for Linear Discriminant Analysis in the Univariate Heteroskedastic Gaussian Model

Amin Zollanvari^{a,b,c}, Ulisses Braga-Neto^d,
Edward Dougherty^{d,e,f,*}

^a*Harvard-MIT Division of Health Science and Technology, Boston, MA, 02115*

^b*Brigham and Women's Hospital, Boston, MA, 02115*

^c*Harvard Medical School, Boston, MA, 02115*

^d*Department of Electrical and Computer Engineering, Texas A&M University,
College Station, TX 77843*

^e*Translational Genomics Research Institute (TGEN), Phoenix, AZ 85004*

^f*Department of Bioinformatics and Computational Biology, University of Texas
M. D. Anderson Cancer Center, Houston TX 77030*

Abstract

This paper provides exact analytical expressions for the bias, variance, and RMS for the resubstitution and leave-one-out error estimators in the case of linear discriminant analysis (LDA) in the univariate heteroskedastic Gaussian model. Neither the variances nor the sample sizes for the two classes need be the same. The generality of heteroskedasticity (unequal variances) is a fundamental feature of the work presented in this paper, which distinguishes it from past work. The expected resubstitution and leave-one-out errors are represented by probabilities involving bivariate Gaussian distributions. Their second moments and cross-moments with the actual error are represented by 4-variate Gaussian distributions. From these, the bias, deviation variance, and RMS for resubstitution and leave-one-out as estimators of the actual error can be computed. The RMS expressions are applied to the determination of sample size and apply to biomarker classification.

Key words: Linear Discriminant Analysis, Error Estimation, Resubstitution, Leave-one-out, RMS, Heteroskedasticity, Genomics.

1 Introduction

Epistemologically, the most important aspect of a classifier is its error, since the error quantifies its predictive capacity and therein lays its scientific validity. In practice, the error must be estimated from data. When samples are large, error estimation is not problematic because the data can be split into training and testing data, the classifier designed on the training data, and its error estimated on the test data. When the error is estimated on independent test data, the RMS between the true error and the estimated error possesses the distribution-free bound $1/2\sqrt{m}$, where m is the number of points in the test set. With small samples, the data cannot be split, the error must be estimated on the same data as training, and estimation accuracy is much more problematic. Here we consider two classical training-data-based error estimators, resubstitution and leave-one-out cross-validation.

Full probabilistic characterization between the true error and an estimated error is given by the joint distribution between the true and estimated errors. Partial information is contained in their mixed moments, in particular, their second mixed moment. Marginal information regarding an error estimator is contained in its marginal moments, in particular, its mean and variance. Since we are interested in estimator accuracy and wish to use the RMS to measure that accuracy, we desire knowledge of the second-order moments, marginal and mixed.

Historically, there has been considerable study of the moments of error estimators for linear discriminant analysis (LDA) in the Gaussian model, mainly on the first two moments of resubstitution [1–9]. In addition, in the 1970s, the expected value of resubstitution was obtained for multinomial discrimination [10, 11]. More recently, error estimation in high-throughput biological classification, where sample sizes are typically quite small, has motivated the desire for distributional knowledge concerning error estimators, both their full joint distribution and their second-order moments (and therefore the RMS and correlation). For multinomial discrimination, exact representations of the second-order moments, both marginal and mixed, for the true error and the resubstitution and leave-one-out estimators have been found [12, 13]. For LDA in the Gaussian model with common known covariance matrix, for both resubstitution and leave-one-out, we have found the marginal distributions for the error estimators [14] and obtained the joint distributions between the true error and both error estimators [15]. In this paper, we obtain exact, nonasymptotic, analytic expressions for the first and second moments, marginal and mixed, for leave-one-out and resubstitution in the univariate heteroskedastic (unequal-variance) Gaussian model, thereby arriving at analytic representation of the RMS. This follows an historical path in pattern recognition: exact error representations are found in the univariate model, with the multivariate model restricted to asymptotic or approximate representations. Note, however, that the generality of heteroskedasticity is a fundamental feature of the present paper, which distinguishes it from past work.

* Corresponding author.

Email address: `edward@ece.tamu.edu` (Edward Dougherty).

Although the discriminatory power of univariate classifiers is limited, their significance becomes apparent when we recognize that most of the common tests for diagnosis and prognosis of cancer are univariate. For instance, PSA for prostate cancer [16], AFP for liver cancer [17], CA 125 for ovarian cancer [18], and CA 19.9 for colorectal cancer [19] are major protein markers. In addition to these protein biomarkers, there are genomic markers such as BRCA1 for breast cancer [20], BRCA2 [21] for male breast cancer, and APC for pancreatic cancer [22].

The paper is organized as follows: Section 2 reviews LDA classification and error estimation, Section 3 discusses performance criteria for error estimation, and Sections 4, 5, and 6 discuss the actual error, the resubstitution error estimate and the leave-one-out error estimate, respectively. Section 7 discusses performance bounds based on the RMS and provides an application to genomic classification. Proofs are in the Appendix.

2 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 X_1, X_2, \dots, X_{n_0} come from population Π_0 and $X_{n_0+1}, X_{n_0+2}, \dots, X_{n_0+n_1}$ come from population Π_1 . Population Π_i is assumed to follow a univariate Gaussian distribution $N(\mu_i, \sigma_i^2)$, for $i = 0, 1$. In general, *Linear Discriminant Analysis* (LDA) utilizes the Anderson W statistic

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

where $\bar{X}_0 = \frac{1}{n_0} \sum_{i=1}^{n_0} X_i$ and $\bar{X}_1 = \frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} X_i$ are the sample means for each class and $\hat{\Sigma}$ is the pooled estimate of the covariance matrix, which is assumed to be common in the LDA discriminant. The designed LDA classifier is given by

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

that is, the sign of W determines the classification of the sample point X . In the univariate model, (1) reduces to

$$W(X) = \frac{1}{\hat{\sigma}^2} (X - \bar{X}) (\bar{X}_0 - \bar{X}_1). \quad (3)$$

Since $\hat{\sigma}^2$ is positive and only the sign of W matters, we can further reduce this to

$$W(X) = (X - \bar{X}) (\bar{X}_0 - \bar{X}_1) \quad (4)$$

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

$$\epsilon = P(W(\bar{X}_0, \bar{X}_1, X) \leq 0, X \in \Pi_0 | \bar{X}_0, \bar{X}_1) + P(W(\bar{X}_0, \bar{X}_1, X) > 0, X \in \Pi_1 | \bar{X}_0, \bar{X}_1) = \alpha_0 \epsilon^0 + \alpha_1 \epsilon^1 \quad (5)$$

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

$$\epsilon^0 = P(W(\bar{X}_0, \bar{X}_1, X) \leq 0 | X \in \Pi_0, \bar{X}_0, \bar{X}_1) \quad \text{and} \quad \epsilon^1 = P(W(\bar{X}_0, \bar{X}_1, X) > 0 | X \in \Pi_1, \bar{X}_0, \bar{X}_1). \quad (6)$$

The *resubstitution error estimator* [23], is given by

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

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

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

The *leave-one-out error estimator* [24] for the LDA classification rule is given by

$$\hat{\epsilon}^l = \frac{1}{n} \left[\sum_{i=1}^{n_0} I_{\{W^{(i)}(\bar{X}_0, \bar{X}_1, X_i) \leq 0\}} + \sum_{i=n_0+1}^{n_0+n_1} I_{\{W^{(i)}(\bar{X}_0, \bar{X}_1, X_i) > 0\}} \right] = \hat{\alpha}_0 \hat{\epsilon}_0^l + \hat{\alpha}_1 \hat{\epsilon}_1^l \quad (9)$$

where $W^{(i)}$ is the discriminant obtained when sample 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

$$\hat{\epsilon}_0^l = \frac{1}{n_0} \sum_{i=1}^{n_0} I_{\{W^{(i)}(\bar{X}_0, \bar{X}_1, X_i) \leq 0\}} \quad \text{and} \quad \hat{\epsilon}_1^l = \frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} I_{\{W^{(i)}(\bar{X}_0, \bar{X}_1, X_i) > 0\}}. \quad (10)$$

3 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

$$\begin{aligned} \text{Bias}[\hat{\epsilon}] &= E[\hat{\epsilon}] - E[\epsilon], \\ \text{Var}_d[\hat{\epsilon}] &= \text{Var}(\hat{\epsilon} - \epsilon) = \text{Var}(\epsilon) + \text{Var}(\hat{\epsilon}) - 2\text{Cov}(\epsilon, \hat{\epsilon}), \\ \text{RMS}[\hat{\epsilon}] &= \sqrt{E[(\epsilon - \hat{\epsilon})^2]} = \sqrt{E[\epsilon^2] + E[\hat{\epsilon}^2] - 2E[\epsilon\hat{\epsilon}]}, \end{aligned} \quad (11)$$

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 leave-one-out error estimator is nearly unbiased, but generally has large variance. 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 $W(\bar{X}_0, \bar{X}_1, X)$. These expressions 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 $W(\bar{X}_0, \bar{X}_1, X_i)$ by $W^{(i)}(\bar{X}_0, \bar{X}_1, X_i)$ throughout.

The first moment of the actual error is given by

$$E[\epsilon] = \alpha_0 P(W(\bar{X}_0, \bar{X}_1, X) \leq 0 | X \in \Pi_0) + \alpha_1 P(W(\bar{X}_0, \bar{X}_1, X) > 0 | X \in \Pi_1) \quad (12)$$

To obtain the second moment of actual error we have:

$$\begin{aligned} E[\epsilon^2] &= E[(\alpha_0 \epsilon^0 + \alpha_1 \epsilon^1)^2] \\ &= \alpha_0^2 E[\epsilon^0 \epsilon^0] + 2\alpha_0 \alpha_1 E[\epsilon^0 \epsilon^1] + \alpha_1^2 E[\epsilon^1 \epsilon^1] \end{aligned} \quad (13)$$

It follows from (6) that

$$\begin{aligned} E[\epsilon^0 \epsilon^0] &= E\left[P(W(\bar{X}_0, \bar{X}_1, X) \leq 0 | X \in \Pi_0, \bar{X}_0, \bar{X}_1) \right. \\ &\quad \times \left. P(W(\bar{X}_0, \bar{X}_1, X') \leq 0 | X' \in \Pi_0, \bar{X}_0, \bar{X}_1)\right] \\ &= E\left[P\left(W(\bar{X}_0, \bar{X}_1, X) \leq 0, W(\bar{X}_0, \bar{X}_1, X') \leq 0 \mid X, X' \in \Pi_0, \bar{X}_0, \bar{X}_1\right)\right] = \\ &P\left(W(\bar{X}_0, \bar{X}_1, X) \leq 0, W(\bar{X}_0, \bar{X}_1, X') \leq 0 \mid X, X' \in \Pi_0\right) \end{aligned} \quad (14)$$

Similar expressions obtain for the other terms in (13), namely $E[\epsilon^0 \epsilon^1]$ and $E[\epsilon^1 \epsilon^1]$. In all,

$$\begin{aligned} E[\epsilon^2] &= \alpha_0^2 P\left(W(\bar{X}_0, \bar{X}_1, X) \leq 0, W(\bar{X}_0, \bar{X}_1, X') \leq 0 \mid X, X' \in \Pi_0\right) \\ &+ 2\alpha_0 \alpha_1 P\left(W(\bar{X}_0, \bar{X}_1, X) \leq 0, W(\bar{X}_0, \bar{X}_1, X') > 0 \mid X \in \Pi_0, X' \in \Pi_1\right) + \\ &\alpha_1^2 P\left(W(\bar{X}_0, \bar{X}_1, X) > 0, W(\bar{X}_0, \bar{X}_1, X') > 0 \mid X, X' \in \Pi_1\right) \end{aligned} \quad (15)$$

Similarly, the first and second moments for resubstitution are

$$\begin{aligned} E[\hat{\epsilon}^r] &= \hat{\alpha}_0 E[\hat{\epsilon}_0^r] + \hat{\alpha}_1 E[\hat{\epsilon}_1^r] \\ &= \hat{\alpha}_0 P(W(\bar{X}_0, \bar{X}_1, X_1) \leq 0) + \hat{\alpha}_1 P(W(\bar{X}_0, \bar{X}_1, X_{n_0+1}) > 0) \end{aligned} \quad (16)$$

and

$$\begin{aligned}
E[(\hat{\epsilon}^r)^2] &= \hat{\alpha}_0^2 E[(\hat{\epsilon}_0^r)^2] + 2\hat{\alpha}_0\hat{\alpha}_1 E[\hat{\epsilon}_0^r \hat{\epsilon}_1^r] + \hat{\alpha}_1^2 E[(\hat{\epsilon}_1^r)^2] \\
&= \frac{\hat{\alpha}_0^2}{n_0^2} E \left[\sum_{i=1}^{n_0} \sum_{j=1}^{n_0} I_{\{W(\bar{X}_0, \bar{X}_1, X_i) \leq 0, W(\bar{X}_0, \bar{X}_1, X_j) \leq 0\}} \right] \\
&+ 2 \frac{\hat{\alpha}_0\hat{\alpha}_1}{n_0 n_1} E \left[\sum_{i=1}^{n_0} \sum_{j=n_0+1}^{n_0+n_1} I_{\{W(\bar{X}_0, \bar{X}_1, X_i) \leq 0, W(\bar{X}_0, \bar{X}_1, X_j) > 0\}} \right] \\
&+ \frac{\hat{\alpha}_1^2}{n_1^2} E \left[\sum_{i=n_0+1}^{n_0+n_1} \sum_{j=n_0+1}^{n_0+n_1} I_{\{W(\bar{X}_0, \bar{X}_1, X_i) > 0, W(\bar{X}_0, \bar{X}_1, X_j) > 0\}} \right] \\
&= \frac{\hat{\alpha}_0^2}{n_0} P(W(\bar{X}_0, \bar{X}_1, X_1) \leq 0) + \frac{\hat{\alpha}_1^2}{n_1} P(W(\bar{X}_0, \bar{X}_1, X_{n_0+1}) > 0) \\
&+ \hat{\alpha}_0^2 \frac{n_0 - 1}{n_0} P(W(\bar{X}_0, \bar{X}_1, X_1) \leq 0, W(\bar{X}_0, \bar{X}_1, X_2) \leq 0) \\
&+ \hat{\alpha}_1^2 \frac{n_1 - 1}{n_1} P(W(\bar{X}_0, \bar{X}_1, X_{n_0+1}) > 0, W(\bar{X}_0, \bar{X}_1, X_{n_0+2}) > 0) \\
&+ 2\hat{\alpha}_0\hat{\alpha}_1 P(W(\bar{X}_0, \bar{X}_1, X_1) \leq 0, W(\bar{X}_0, \bar{X}_1, X_{n_0+1}) > 0)
\end{aligned} \tag{17}$$

respectively. The mixed moment is given by

$$\begin{aligned}
E[\hat{\epsilon}\hat{\epsilon}^r] &= \alpha_0\hat{\alpha}_0 P(W(\bar{X}_0, \bar{X}_1, X) \leq 0, W(\bar{X}_0, \bar{X}_1, X_1) \leq 0 \mid X \in \Pi_0) \\
&+ \alpha_0\hat{\alpha}_1 P(W(\bar{X}_0, \bar{X}_1, X) \leq 0, W(\bar{X}_0, \bar{X}_1, X_{n_0+1}) > 0 \mid X \in \Pi_0) \\
&+ \alpha_1\hat{\alpha}_0 P(W(\bar{X}_0, \bar{X}_1, X) > 0, W(\bar{X}_0, \bar{X}_1, X_1) \leq 0 \mid X \in \Pi_1) \\
&+ \alpha_1\hat{\alpha}_1 P(W(\bar{X}_0, \bar{X}_1, X) > 0, W(\bar{X}_0, \bar{X}_1, X_{n_0+1}) > 0 \mid X \in \Pi_1)
\end{aligned} \tag{18}$$

4 Actual Classification Error

Starting from the expressions in the previous section, we derive the exact expressions for the bias, variance, and RMS of the resubstitution and leave-one-out for LDA in the univariate Gaussian model. In analogy to [25] and [15], the basic method used in these proofs consists in writing out the W statistic in an appropriate matrix form. We refer to the following sampling procedure as the *univariate heteroskedastic Gaussian sampling model*: $X_i \sim N(\mu_0, \sigma_0^2)$ for $i = 1, \dots, n_0$ and $X_i \sim N(\mu_1, \sigma_1^2)$ for $i = n_0 + 1, \dots, n_0 + n_1$ compose a set of $n = n_0 + n_1$ i.i.d. observations used to derive the classifier in (2).

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

Theorem 1 *Under the univariate heteroskedastic Gaussian sampling model,*

$$E[\epsilon] = \alpha_0 [P(Z^I < 0) + P(Z^I \geq 0)] + \alpha_1 [P(Z^{II} < 0) + P(Z^{II} \geq 0)] \tag{19}$$

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

$$\begin{aligned} \mu_{Z^I} &= \begin{bmatrix} \frac{\mu}{2} \\ -\mu \end{bmatrix}, \quad \Sigma_{Z^I} = \begin{pmatrix} (1 + \frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\ \mu_{Z^{II}} &= \begin{bmatrix} \frac{-\mu}{2} \\ \mu \end{bmatrix}, \quad \Sigma_{Z^{II}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \end{aligned} \quad (20)$$

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

Proof. See the Appendix.

Theorem 2 Under the univariate heteroskedastic Gaussian sampling model,

$$\begin{aligned} E[\epsilon^2] &= \alpha_0\alpha_0 [P(Z^I < 0) + P(Z^I \geq 0)] \\ &+ 2\alpha_0\alpha_1 [P(Z^{II} < 0) + P(Z^{II} \geq 0)] \\ &+ \alpha_1\alpha_1 [P(Z^{III} < 0) + P(Z^{III} \geq 0)] \end{aligned} \quad (21)$$

where Z^j for $j = I, \dots, III$, are 3-variate Gaussian random vectors with means and covariance matrices as follows:

$$\mu_{Z^I} = \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ \frac{\mu}{2} \end{bmatrix}, \quad \Sigma_{Z^I} = \begin{pmatrix} (1 + \frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \cdot & (1 + \frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} \end{pmatrix} \quad (22)$$

$$\mu_{Z^{II}} = \mu_{Z^I}, \quad \Sigma_{Z^{II}} = \begin{pmatrix} (1 + \frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ \cdot & \cdot & \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4n_1})\sigma_1^2 \end{pmatrix} \quad (23)$$

where $\mu = \mu_0 - \mu_1$, and $\mu_{Z^{III}}$ and $\Sigma_{Z^{III}}$ are obtained from μ_{Z^I} and Σ_{Z^I} , respectively, by exchanging n_0 and n_1 , and σ_0 and σ_1 .

Proof. See the Appendix.

5 Resubstitution Error Estimator

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

Theorem 3 *Under the univariate heteroskedastic Gaussian sampling model,*

$$E[\hat{\epsilon}^r] = \hat{\alpha}_0 [P(Z^I < 0) + P(Z^I \geq 0)] + \hat{\alpha}_1 [P(Z^{II} < 0) + P(Z^{II} \geq 0)] \quad (24)$$

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

$$\begin{aligned} \mu_{Z^I} &= \begin{bmatrix} \frac{\mu}{2} \\ -\mu \end{bmatrix}, & \Sigma_{Z^I} &= \begin{pmatrix} (1 - \frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\ \mu_{Z^{II}} &= \begin{bmatrix} -\frac{\mu}{2} \\ \mu \end{bmatrix}, & \Sigma_{Z^{II}} &= \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1 - \frac{3}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \end{aligned} \quad (25)$$

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

Proof. Similar to Theorem 1.

Theorem 4 *Under the univariate heteroskedastic Gaussian sampling model,*

$$\begin{aligned} E[(\hat{\epsilon}^r)^2] &= \frac{\hat{\alpha}_0^2}{n_0} [P(Z^I < 0) + P(Z^I \geq 0)] + \frac{\hat{\alpha}_1^2}{n_1} [P(Z^{II} < 0) + P(Z^{II} \geq 0)] \\ &+ \hat{\alpha}_0^2 \frac{n_0 - 1}{n_0} [P(Z^{III} < 0) + P(Z^{III} \geq 0)] \\ &+ \hat{\alpha}_1^2 \frac{n_1 - 1}{n_1} [P(Z^{IV} < 0) + P(Z^{IV} \geq 0)] \\ &+ 2\hat{\alpha}_0\hat{\alpha}_1 [P(Z^V < 0) + P(Z^V \geq 0)] \end{aligned} \quad (26)$$

where Z^I and Z^{II} are defined in Theorem 3, and Z^j for $j = III, IV, V$, are 3-variate Gaussian

random vectors with means and covariances matrices as follows:

$$\begin{aligned}
\mu_{Z^{III}} &= \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ \frac{\mu}{2} \\ -\mu \end{bmatrix}, \quad \Sigma_{Z^{III}} = \begin{pmatrix} (1 - \frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{3\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \cdot & (1 - \frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} \end{pmatrix} \\
\mu_{Z^{IV}} &= \mu_{Z^{III}}, \quad \Sigma_{Z^{IV}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1 - \frac{3}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{3\sigma_1^2}{4n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \cdot & \frac{\sigma_0^2}{4n_0} + (1 - \frac{3}{4n_1})\sigma_1^2 \end{pmatrix} \\
\mu_{Z^V} &= \mu_{Z^{III}}, \quad \Sigma_{Z^V} = \begin{pmatrix} (1 - \frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \cdot & \frac{\sigma_0^2}{4n_0} + (1 - \frac{3}{4n_1})\sigma_1^2 \end{pmatrix}
\end{aligned}$$

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

Proof. See the Appendix.

Theorem 5 Under the univariate heteroskedastic Gaussian sampling model,

$$\begin{aligned}
E[\epsilon \hat{\epsilon}^r] &= \alpha_0 \hat{\alpha}_0 \left[P(Z^I < 0) + P(Z^I \geq 0) \right] \\
&+ \alpha_0 \hat{\alpha}_1 \left[P(Z^{II} < 0) + P(Z^{II} \geq 0) \right] \\
&+ \alpha_1 \hat{\alpha}_0 \left[P(Z^{III} < 0) + P(Z^{III} \geq 0) \right] \\
&+ \alpha_1 \hat{\alpha}_1 \left[P(Z^{IV} < 0) + P(Z^{IV} \geq 0) \right]
\end{aligned} \tag{27}$$

where Z^j for $j = I, \dots, IV$, are 3-variate Gaussian random vectors with means and covariances as follows:

$$\begin{aligned}
\mu_{Z^I} &= \begin{bmatrix} \frac{\mu}{2} \\ -\mu \\ \frac{\mu}{2} \end{bmatrix}, \quad \Sigma_{Z^I} = \begin{pmatrix} (1 + \frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \cdot & (1 - \frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} \end{pmatrix} \\
\mu_{Z^{II}} &= \mu_{Z^I}, \quad \Sigma_{Z^{II}} = \begin{pmatrix} (1 + \frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \cdot & \frac{\sigma_0^2}{4n_0} + (1 - \frac{3}{4n_1})\sigma_1^2 \end{pmatrix}
\end{aligned} \tag{28}$$

$$\mu_{Z^{III}} = \begin{bmatrix} -\frac{\mu}{2} \\ \mu \\ -\frac{\mu}{2} \end{bmatrix}, \quad \Sigma_{Z^{III}} = \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \cdot & (1 - \frac{3}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} \end{pmatrix} \quad (29)$$

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

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

Proof. See the Appendix.

6 Leave-one-out Error Estimator

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

Theorem 6 *Under the univariate heteroskedastic Gaussian sampling model,*

$$\begin{aligned} E[(\hat{\epsilon}^l)^2] &= \frac{\hat{\alpha}_0^2}{n_0} [P(Z^I < 0) + P(Z^I \geq 0)] + \frac{\hat{\alpha}_1^2}{n_1} [P(Z^{II} < 0) + P(Z^{II} \geq 0)] \\ &+ \hat{\alpha}_0^2 \frac{n_0 - 1}{n_0} [P(Z_0^{III} < 0) + P(Z_0^{III} \geq 0) + P(Z_1^{III} < 0) + P(Z_1^{III} \geq 0)] \\ &+ \hat{\alpha}_1^2 \frac{n_1 - 1}{n_1} [P(Z_0^{IV} < 0) + P(Z_0^{IV} \geq 0) + P(Z_1^{IV} < 0) + P(Z_1^{IV} \geq 0)] \\ &+ 2\hat{\alpha}_0\hat{\alpha}_1 \frac{1}{n_1} [P(Z_0^V < 0) + P(Z_0^V \geq 0) + P(Z_1^V < 0) + P(Z_1^V \geq 0)] \end{aligned} \quad (30)$$

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

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

$$\begin{aligned}
\Sigma_{Z_0^{III}} &= \begin{pmatrix} (1 + \frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} & \frac{(-3n_0+2)\sigma_0^2}{4(n_0-1)^2} + \frac{\sigma_1^2}{4n_1} & -\frac{n_0\sigma_0^2}{2(n_0-1)^2} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} & -\frac{n_0\sigma_0^2}{2(n_0-1)^2} - \frac{\sigma_1^2}{2n_1} & \frac{(n_0-2)\sigma_0^2}{(n_0-1)^2} + \frac{\sigma_1^2}{n_1} \\ \cdot & \cdot & (1 + \frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \cdot & \cdot & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\
\Sigma_{Z_1^{III}} &= \begin{pmatrix} (1 + \frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} & -\frac{(-3n_0+2)\sigma_0^2}{4(n_0-1)^2} - \frac{\sigma_1^2}{4n_1} & \frac{n_0\sigma_0^2}{2(n_0-1)^2} + \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} & \frac{n_0\sigma_0^2}{2(n_0-1)^2} + \frac{\sigma_1^2}{2n_1} & -\frac{(n_0-2)\sigma_0^2}{(n_0-1)^2} - \frac{\sigma_1^2}{n_1} \\ \cdot & \cdot & (1 + \frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \cdot & \cdot & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\
\Sigma_{Z_0^{IV}} &= \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} & \frac{\sigma_0^2}{4n_0} + \frac{(-3n_1+2)\sigma_1^2}{4(n_1-1)^2} & -\frac{\sigma_0^2}{2n_0} - \frac{n_1\sigma_1^2}{2(n_1-1)^2} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} & -\frac{\sigma_0^2}{2n_0} - \frac{n_1\sigma_1^2}{2(n_1-1)^2} & \frac{\sigma_0^2}{n_0} + \frac{(n_1-2)\sigma_1^2}{(n_1-1)^2} \\ \cdot & \cdot & \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ \cdot & \cdot & \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix} \\
\Sigma_{Z_1^{IV}} &= \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} & -\frac{\sigma_0^2}{4n_0} - \frac{(-3n_1+2)\sigma_1^2}{4(n_1-1)^2} & \frac{\sigma_0^2}{2n_0} + \frac{n_1\sigma_1^2}{2(n_1-1)^2} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} & \frac{\sigma_0^2}{2n_0} + \frac{n_1\sigma_1^2}{2(n_1-1)^2} & -\frac{\sigma_0^2}{n_0} - \frac{(n_1-2)\sigma_1^2}{(n_1-1)^2} \\ \cdot & \cdot & \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ \cdot & \cdot & \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix} \\
\Sigma_{Z_0^V} &= \begin{pmatrix} (1 + \frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ \cdot & \cdot & \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ \cdot & \cdot & \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix} \\
\Sigma_{Z_1^V} &= \begin{pmatrix} (1 + \frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ \cdot & \cdot & \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ \cdot & \cdot & \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix}
\end{aligned} \tag{31}$$

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

Proof. See the Appendix.

Theorem 7 *Under the univariate heteroskedastic Gaussian sampling model,*

$$\begin{aligned}
E[\epsilon^j] &= \alpha_0 \hat{\alpha}_0 \left[P(Z_0^I < 0) + P(Z_0^I \geq 0) + P(Z_1^I < 0) + P(Z_1^I \geq 0) \right] \\
&+ \alpha_0 \hat{\alpha}_1 \left[P(Z_0^{II} < 0) + P(Z_0^{II} \geq 0) + P(Z_1^{II} < 0) + P(Z_1^{II} \geq 0) \right] \\
&+ \alpha_1 \hat{\alpha}_0 \left[P(Z_0^{III} < 0) + P(Z_0^{III} \geq 0) + P(Z_1^{III} < 0) + P(Z_1^{III} \geq 0) \right] \\
&+ \alpha_1 \hat{\alpha}_1 \left[P(Z_0^{IV} < 0) + P(Z_0^{IV} \geq 0) + P(Z_1^{IV} < 0) + P(Z_1^{IV} \geq 0) \right]
\end{aligned} \tag{32}$$

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

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

$$\begin{aligned}
\Sigma_{Z_0^I} &= \begin{pmatrix} (1 + \frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ \cdot & \cdot & (1 + \frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \cdot & \cdot & \frac{\sigma_0^2}{(n_0-1)} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\
\Sigma_{Z_1^I} &= \begin{pmatrix} (1 + \frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ \cdot & \cdot & (1 + \frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \cdot & \cdot & \frac{\sigma_0^2}{(n_0-1)} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\
\Sigma_{Z_0^{II}} &= \begin{pmatrix} (1 + \frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_0} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ \cdot & \cdot & \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ \cdot & \cdot & \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix} \\
\Sigma_{Z_1^{II}} &= \begin{pmatrix} (1 + \frac{1}{4n_0})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ \cdot & \cdot & \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ \cdot & \cdot & \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix}
\end{aligned}$$

$$\begin{aligned}
\Sigma_{Z_0^{III}} &= \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ \cdot & \cdot & (1 + \frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \cdot & \cdot & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\
\Sigma_{Z_1^{III}} &= \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ \cdot & \cdot & (1 + \frac{1}{4(n_0-1)})\sigma_0^2 + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2(n_0-1)} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \cdot & \cdot & \frac{\sigma_0^2}{n_0-1} + \frac{\sigma_1^2}{n_1} \end{pmatrix} \\
\Sigma_{Z_0^{IV}} &= \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{4n_0} + \frac{\sigma_1^2}{4n_1} & \frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & -\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} \\ \cdot & \cdot & \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ \cdot & \cdot & \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix} \\
\Sigma_{Z_1^{IV}} &= \begin{pmatrix} \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4n_1})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{4n_0} - \frac{\sigma_1^2}{4n_1} & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2n_1} \\ \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & -\frac{\sigma_0^2}{2n_0} - \frac{\sigma_1^2}{2n_1} & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} \\ \cdot & \cdot & \frac{\sigma_0^2}{4n_0} + (1 + \frac{1}{4(n_1-1)})\sigma_1^2 & -\frac{\sigma_0^2}{2n_0} + \frac{\sigma_1^2}{2(n_1-1)} \\ \cdot & \cdot & \cdot & \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1-1} \end{pmatrix}
\end{aligned} \tag{33}$$

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

Proof. See the Appendix.

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

7 RMS Bounds

When one designs a classifier and reports an error estimate, there is no way to tell how accurate the estimate is because we do not know the true error of the classifier. Knowledge of estimation accuracy rests with the accuracy of the error estimation rule, which is most commonly judged by the RMS. When reporting an estimate, it would be beneficial to state some bounds on the RMS. In addition, as in any experimental situation, it would be useful to determine ahead of time the minimum sample size necessary to obtain a desired degree

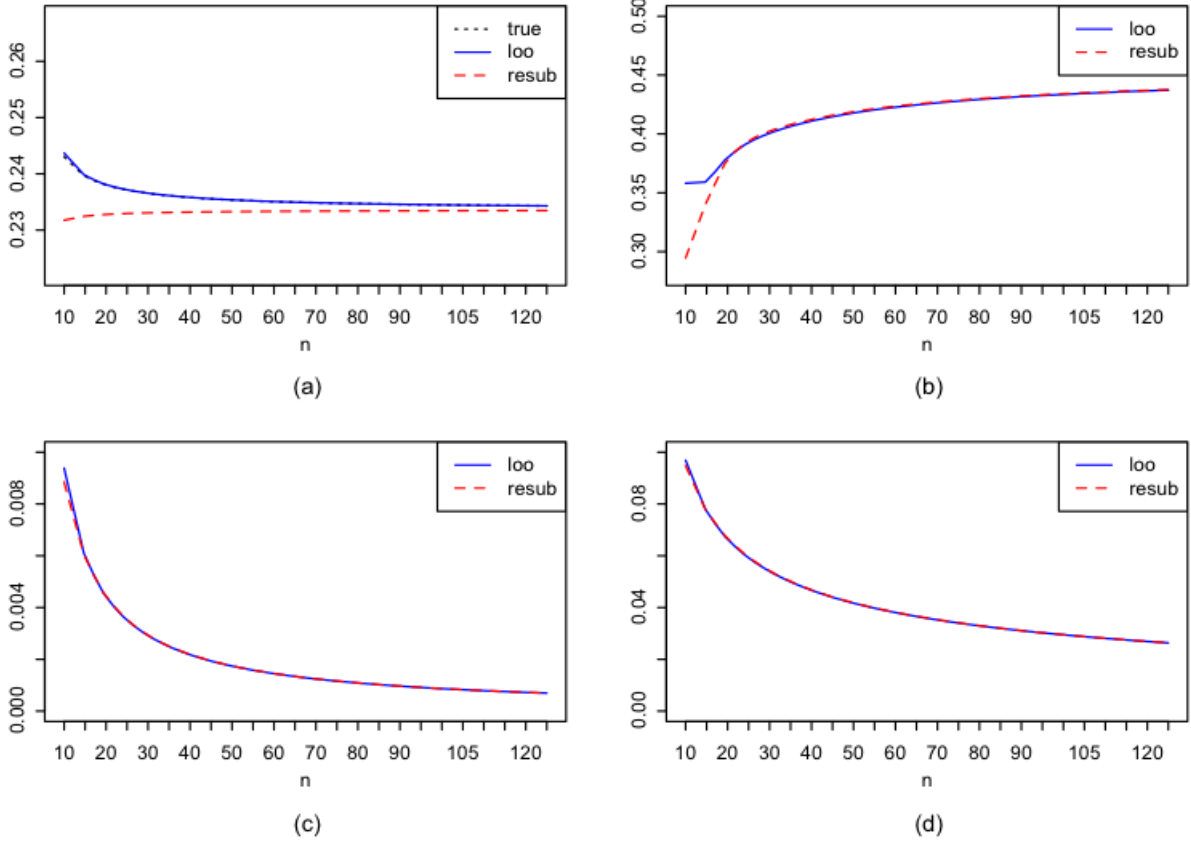


Fig. 1. Performance measures for resubstitution and leave-one-out as a function of sample size for LDA in the univariate model: (a) mean errors and actual error, (b) correlation coefficient with actual error, (c) deviation variance, (d) RMS.

of estimation accuracy. In this vein, some recommendations on sample size requirements have been provided in the literature [26, 27]. 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 [7, 8, 28]; however, these bounds tend to be very loose and therefore of limited practical value. For instance, for resubstitution and the histogram rule we have the bound $\text{RMS}[\hat{\epsilon}^r] \leq \sqrt{\frac{6k}{n_0+n_1}}$, where k is the maximum number of fixed partitions of the feature space [28]. Taking $k = 10$ and $k = 20$ with $n_0 + n_1 = 100$, then the bounds are 0.77 and 1.09, respectively, both being of no practical value.

Now consider leave-one-out, resubstitution and LDA in the model class we have been considering. Consider two equal univariate Gaussian distributions with means $\mu_1 = -\mu_0 = 1$ and $\sigma_0 = \sigma_1 = 1$. Using the RMS expressions obtained in this paper, the RMS versus Bayes error curves are shown in Fig. 2 for different sample sizes and balanced design, $n_0 = n_1 = n$. We see that RMS is an increasing function of the Bayes error, ϵ_{bay} . Letting $\kappa_{\hat{\epsilon}}(n, \tau) = \max_{\epsilon_{bay} \leq \tau} \text{RMS}[\hat{\epsilon}]$ for $n_0 = n_1 = n$ and $\hat{\epsilon} = \hat{\epsilon}^r, \hat{\epsilon}^l$, we have the bounds $\text{RMS}[\hat{\epsilon}^l] \leq \kappa_{\hat{\epsilon}^l}(20, 0.5) = 0.145$ and $\text{RMS}[\hat{\epsilon}^r] \leq \kappa_{\hat{\epsilon}^r}(20, 0.5) = 0.080$ for $n = 20$, and $\text{RMS}[\hat{\epsilon}^l] \leq \kappa_{\hat{\epsilon}^l}(30, 0.5) = 0.127$ and $\text{RMS}[\hat{\epsilon}^r] \leq \kappa_{\hat{\epsilon}^r}(30, 0.5) = 0.065$ for $n = 30$.

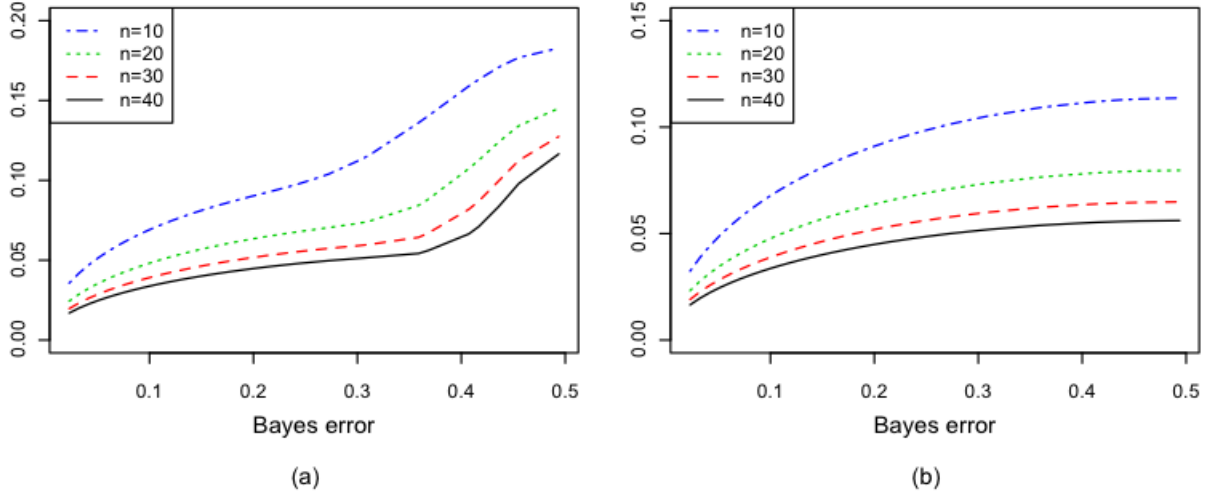


Fig. 2. RMS versus Bayes error in a Gaussian model for (a) leave-one-out, (b) resubstitution.

From a practical perspective, given a desired RMS, the required sample size can be determined. If one desires that the RMS be bounded by η , then one needs only find the minimum value of n so that $\kappa_{\hat{\epsilon}}(n, 0.5) \leq \eta$ where $\hat{\epsilon} = \hat{\epsilon}^r, \hat{\epsilon}^l$. Table 1 shows the required sample size calculated using this scheme for a balanced design ($n_0 = n_1 = n$). Note that the required sample size in Table 1 does not depend on the actual value of the common variance, a peculiar result of the equal-variance model class being considered. In the univariate case, the number of samples needed to achieve a given $\kappa_{\hat{\epsilon}^l}(n, 0.5)$ is much higher than $\kappa_{\hat{\epsilon}^r}(n, 0.5)$, which is evident in Fig. 2, owing to the abrupt increase of $\text{RMS}[\hat{\epsilon}^l]$ for large ϵ_{bay} . While $\text{RMS}[\hat{\epsilon}^l] \approx \text{RMS}[\hat{\epsilon}^r]$ when $\epsilon_{bay} \leq 0.35$, since we do not know the true error, the bound for $\text{RMS}[\hat{\epsilon}^l]$ must take into account the possibility $\epsilon_{bay} > 0.35$. It is instructive (although not directly analogous) to compare the sample sizes determined from Table 1 with those determined from the histogram-rule bound $\text{RMS}[\hat{\epsilon}^r] \leq \sqrt{\frac{6k}{n_0+n_1}}$ to achieve a given RMS, say 0.1. For the latter bound with $k = 10$ and $k = 20$, we need in excess of 3000 and 6000 sample points in each class, respectively, whereas from Table 1 we need 13 sample points in each class for univariate LDA and resubstitution.

Table 1

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

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

7.1 Implementation for gene-expression classification

In this section, we demonstrate the practical use of RMS bounds in the case of classification using gene-expression data from a breast-cancer study that analyzed 295 gene-expression microarrays containing a total of 25760 transcripts on each [29]. Discrimination is between good versus bad prognosis. Here we design of a classifier based on a single gene. Using resubstitution, from Table 1, we need 20 sample points for each class to have $\kappa_{\hat{\epsilon}^r}(n, 0.5) = 0.08$. This bound does not apply to leave-one-out; indeed, $\kappa_{\hat{\epsilon}^l}(20, 0.5) > 0.13$. However, as explained previously, if it happens that $\epsilon_{bay} < 0.35$ then $\text{RMS}[\hat{\epsilon}^l] \approx \text{RMS}[\hat{\epsilon}^r]$, so that $\kappa_{\hat{\epsilon}^l}(20, 0.35) \approx \kappa_{\hat{\epsilon}^r}(20, 0.35) < \kappa_{\hat{\epsilon}^r}(20, 0.5) = 0.08$ also. This example will elucidate this situation because we will have an accurate estimate of the true error. We consider the total of 295 gene-expression profiles for 70 genes from the 295 microarrays as the population and draw a random sample of size 40 with $n_0 = n_1 = 20$. Using the 40 sample points selected, we applied the t-test to find the differentially expressed genes among the 70 genes. Results of the t-test on the sample showed 35 genes to be differentially expressed among the 70 genes. Then the Shapiro-Wilk test (using the R statistical software) was applied on these 35 genes to test the normality of each gene at significance level 0.95. Note that to do so, only the 40 points taken randomly from the whole population were considered, so as to reflect the situation that no additional data are available in practice. The test did not reject the Gaussianity assumption of 26 genes out of the 35 genes previously selected by the t-test. Although not necessary from the theory, the F-test for equality of variances of both classes was performed on these 26 selected genes to test the equality of variances of each gene across the classes. The result of the F-test reduced the number of genes to 13. In sum, these 13 genes are those that show significant different expressions between two classes (by t-test), are close to normal (by Shapiro-Wilk test), and have close to equal variances in the two classes (by F-test). These genes are shown in Table 2. The last column of this table shows the hold out estimate using 190 hold-out points selected from the 255 remaining sample points to reflect the equal prior probability of the classes, as was done for training. With 190 hold-out points, one can expect the hold-out estimate to be very accurate. Comparing the values of hold-out in these examples with those of the estimators themselves, we conclude that both resubstitution and leave-one-out have reasonably estimated the true error. We would certainly have expected this owing to the RMS bound on resubstitution and, as we see the true errors are less than 0.35, so that the Bayes errors must also be less than 0.35, in hindsight we expect this from leave-one-out. In practice, of course, we do not have a population based evaluation of the true error, so that a conservative approach requires taking $\kappa_{\hat{\epsilon}^l}(n, 0.5)$ as the bound.

8 Conclusion

In this paper we have provided exact representation for the main performance criteria, bias, variance, and RMS, for resubstitution and leave-one-out for LDA in a univariate heteroskedastic Gaussian model. The generality of heteroskedasticity is a fundamental feature of the work presented in this paper, which distinguishes it from past work. Since the de-

Table 2

Genes selected using the validity-goodness model selection criterion.

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

rived expressions depend on the actual class variances, in practice, when the variances are not known, one may replace them by the sample variances and obtain approximate results (which converge to the exact results as sample size increases). An obvious, desirable extension of the present work would be to extend these results to the multivariate case when the covariance matrices are known and distinct.

Acknowledgements

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

Appendix

9 Proof of Theorem 1

From (4), it follows that

$$P(W(\bar{X}_0, \bar{X}_1, X) \leq 0 \mid X \in \Pi_0) = P(X - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 > 0) + P(X - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0) \quad (34)$$

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

$$P(W(\bar{X}_0, \bar{X}_1, X) \leq 0 \mid X \in \Pi_0) = P(Z^I < 0) + P(Z^I \geq 0) \quad (35)$$

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

$$A = \begin{pmatrix} 1 & -\frac{1}{2n_0} & -\frac{1}{2n_0} & \cdots & -\frac{1}{2n_0} & -\frac{1}{2n_1} & \cdots & -\frac{1}{2n_1} \\ 0 & -\frac{1}{n_0} & -\frac{1}{n_0} & \cdots & -\frac{1}{n_0} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \end{pmatrix} \quad (36)$$

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

10 Proof of Theorem 2

We expand the first term in (15); the other terms are similar. Using (4), we have

$$\begin{aligned}
& P(W(\bar{X}_0, \bar{X}_1, X) \leq 0, W(\bar{X}_0, \bar{X}_1, X') \leq 0 \mid X, X' \in \Pi_0) = \\
& P(X - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0, X' - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0) + \\
& P(X - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 > 0, X' - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \geq 0) + \\
& P(X - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0, X' - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \geq 0) + \\
& P(X - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \geq 0, X' - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0) = \\
& P(X - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0, X' - \bar{X} \geq 0) + \\
& P(X - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 > 0, X' - \bar{X} < 0)
\end{aligned} \tag{37}$$

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

$$\begin{aligned}
& P(W(\bar{X}_0, \bar{X}_1, X) \leq 0, W(\bar{X}_0, \bar{X}_1, X') \leq 0 \mid X, X' \in \Pi_0) = \\
& P(Z^I < 0) + P(Z^I \geq 0)
\end{aligned} \tag{38}$$

where $Z^I = AY$ in which $Y = [X, X', X_1, \dots, X_{n_0}, X_{n_0+1}, \dots, X_{n_0+n_1}]^T$ and

$$A = \begin{pmatrix} 1 & 0 & -\frac{1}{2n_0} & -\frac{1}{2n_0} & \cdots & -\frac{1}{2n_0} & -\frac{1}{2n_1} & \cdots & -\frac{1}{2n_1} \\ 0 & 0 & -\frac{1}{n_0} & -\frac{1}{n_0} & \cdots & -\frac{1}{n_0} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \\ 0 & 1 & -\frac{1}{2n_0} & -\frac{1}{2n_0} & \cdots & -\frac{1}{2n_0} & -\frac{1}{2n_1} & \cdots & -\frac{1}{2n_1} \end{pmatrix} \tag{39}$$

Therefore, Z^I is a Gaussian random vector with mean $A\mu_Y$ and covariance matrix $A\Sigma_Y A^T$. Plugging in the values of $\mu_Y = [\mu_0 \mathbf{1}_{n_0+2}, \mu_1 \mathbf{1}_{n_1}]^T$ and $\Sigma_Y = \text{diag}(\sigma_0^2 \mathbf{1}_{n_0+2}, \sigma_1^2 \mathbf{1}_{n_1})$ leads to the expression stated in Theorem 2. Evaluating the means and covariance matrices of Z^{II} and Z^{III} stated in the theorem is entirely similar by considering the corresponding terms in (15).

11 Proof of Theorem 3

We expand the first term in (16); the other term is similar. Using (4), we have

$$P(W(\bar{X}_0, \bar{X}_1, X_1) \leq 0) = P(X_1 - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 > 0) + P(X_1 - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0) \tag{40}$$

where $\bar{X} = \frac{\bar{X}_0 + \bar{X}_1}{2}$. Expanding \bar{X}_0 and \bar{X}_1 as $\frac{1}{n_0} \sum_{i=1}^{n_0} X_i$ and $\bar{X}_1 = \frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} X_i$ results in the Gaussian vector Z^I with the mean and covariance matrix stated in Theorem 3.

12 Proof of Theorem 4

We expand the first term in (17); the other terms are similar. Using (4), we have

$$\begin{aligned}
& P(W(\bar{X}_0, \bar{X}_1, X_1) \leq 0, W(\bar{X}_0, \bar{X}_1, X_2) \leq 0) = \\
& P(X_1 - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0, X_2 - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0) + \\
& P(X_1 - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 > 0, X_2 - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \geq 0) + \\
& P(X_1 - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0, X_2 - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \geq 0) + \\
& P(X_1 - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \geq 0, X_2 - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0) = \\
& P(X_1 - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0, X_2 - \bar{X} \geq 0) + \\
& P(X_1 - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 > 0, X_2 - \bar{X} < 0)
\end{aligned} \tag{41}$$

Expanding \bar{X}_0 , \bar{X}_1 , and \bar{X} as $\frac{1}{n_0} \sum_{i=1}^{n_0} X_i$, $\frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} X_i$, and $\frac{\bar{X}_0 + \bar{X}_1}{2}$, respectively, results in the Gaussian vector Z^{III} with the mean and covariance matrix stated in Theorem 4.

13 Proof of Theorem 5

We expand the first term in (18); the other terms are similar. Using (4), we have

$$\begin{aligned}
& P(W(\bar{X}_0, \bar{X}_1, X) \leq 0, W(\bar{X}_0, \bar{X}_1, X_1) \leq 0 \mid X \in \Pi_0) = \\
& P(X - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0, X_1 - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0) + \\
& P(X - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 > 0, X_1 - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \geq 0) + \\
& P(X - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0, X_1 - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \geq 0) + \\
& P(X - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \geq 0, X_1 - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0) = \\
& P(X - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0, X_1 - \bar{X} \geq 0) + \\
& P(X - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 > 0, X_1 - \bar{X} < 0)
\end{aligned} \tag{42}$$

Expanding \bar{X}_0 , \bar{X}_1 , and \bar{X} as $\frac{1}{n_0} \sum_{i=1}^{n_0} X_i$, $\frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} X_i$, and $\frac{\bar{X}_0 + \bar{X}_1}{2}$, respectively, results in the Gaussian vector Z^I with the mean and covariance matrix stated in Theorem 5.

14 Proof of Theorem 6

As described in section 3, the equation corresponding to (17) for the leave-one-out case can be obtained by replacing $W(\bar{X}_0, \bar{X}_1, X_i)$ by $W^{(i)}(\bar{X}_0, \bar{X}_1, X_i)$ throughout. Here, we expand the first term in this equation; the other terms are similar.

$$\begin{aligned}
& P(W^{(1)}(\bar{X}_0, \bar{X}_1, X_1) \leq 0, W^{(2)}(\bar{X}_0, \bar{X}_1, X_2) \leq 0) = \\
& P(X_1 - \bar{X}^{(1)} \geq 0, \bar{X}_0^{(1)} - \bar{X}_1 < 0, X_2 - \bar{X}^{(2)} \geq 0, \bar{X}_0^{(2)} - \bar{X}_1 < 0) + \\
& P(X_1 - \bar{X}^{(1)} < 0, \bar{X}_0^{(1)} - \bar{X}_1 \geq 0, X_2 - \bar{X}^{(2)} < 0, \bar{X}_0^{(2)} - \bar{X}_1 \geq 0) + \quad (43) \\
& P(X_1 - \bar{X}^{(1)} \geq 0, \bar{X}_0^{(1)} - \bar{X}_1 < 0, X_2 - \bar{X}^{(2)} < 0, \bar{X}_0^{(2)} - \bar{X}_1 \geq 0) + \\
& P(X_1 - \bar{X}^{(1)} < 0, \bar{X}_0^{(1)} - \bar{X}_1 \geq 0, X_2 - \bar{X}^{(2)} \geq 0, \bar{X}_0^{(2)} - \bar{X}_1 < 0)
\end{aligned}$$

Expanding $\bar{X}_0^{(j)}$, \bar{X}_1 , and $\bar{X}^{(j)}$ as $\frac{1}{n_0-1} \sum_{i=1}^{n_0} X_i - \frac{X_j}{n_0-1}$, $\frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} X_i$, and $\frac{\bar{X}_0^{(j)} + \bar{X}_1}{2}$, respectively, results in the Gaussian vectors Z_0^{III} and Z_1^{III} with the means and covariance matrices stated in Theorem 6. Evaluating the means and covariance matrices of Z_l^I , Z_l^{II} , Z_l^{IV} , and Z_l^V , $l = 0, 1$ stated in the theorem is entirely similar.

15 Proof of Theorem 7

As described in section 3, the equation corresponding to (18) for the leave-one-out case can be obtained by replacing $W(\bar{X}_0, \bar{X}_1, X_i)$ by $W^{(i)}(\bar{X}_0, \bar{X}_1, X_i)$ throughout. Here, we expand the first term in this equation; the other terms are similar.

$$\begin{aligned}
& P(W(\bar{X}_0, \bar{X}_1, X) \leq 0, W^{(1)}(\bar{X}_0, \bar{X}_1, X_1) \leq 0 \mid X \in \Pi_0) = \\
& P(X - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0, X_1 - \bar{X}^{(1)} \geq 0, \bar{X}_0^{(1)} - \bar{X}_1 < 0 \mid X \in \Pi_0) + \\
& P(X - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \geq 0, X_1 - \bar{X}^{(1)} < 0, \bar{X}_0^{(1)} - \bar{X}_1 \geq 0 \mid X \in \Pi_0) + \quad (44) \\
& P(X - \bar{X} \geq 0, \bar{X}_0 - \bar{X}_1 < 0, X_1 - \bar{X}^{(1)} < 0, \bar{X}_0^{(1)} - \bar{X}_1 \geq 0 \mid X \in \Pi_0) + \\
& P(X - \bar{X} < 0, \bar{X}_0 - \bar{X}_1 \geq 0, X_1 - \bar{X}^{(1)} \geq 0, \bar{X}_0^{(1)} - \bar{X}_1 < 0 \mid X \in \Pi_0)
\end{aligned}$$

Expanding \bar{X}_0 , $\bar{X}_0^{(j)}$, \bar{X}_1 , \bar{X} , and $\bar{X}^{(j)}$ as $\frac{1}{n_0} \sum_{i=1}^{n_0} X_i$, $\frac{1}{n_0-1} \sum_{i=1}^{n_0} X_i - \frac{X_j}{n_0-1}$, $\frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} X_i$, $\frac{\bar{X}_0 + \bar{X}_1}{2}$, and $\frac{\bar{X}_0^{(j)} + \bar{X}_1}{2}$, respectively, results in the Gaussian vectors Z_0^I and Z_1^I with the means and covariance matrices stated in Theorem 7. Evaluating the means and covariance matrices of Z_l^{II} , Z_l^{III} , and Z_l^{IV} , $l = 0, 1$ stated in the theorem is entirely similar.

References

- [1] M. Hills, “Allocation rules and their error rates,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 28, no. 1, pp. 1–31, 1966.
- [2] 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.
- [3] N. Glick, “Additive estimators for probabilities of correct classification,” *Pattern recognition*, vol. 10, pp. 211–222, 1978.
- [4] K. Fukunaga and R. R. Hayes, “Estimation of classifier performance,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 1087–1101, 1989.
- [5] U. Braga-Neto and E. Dougherty, “Is cross-validation valid for microarray classification?” *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.
- [6] S. Raudys, *Statistical and Neural Classifiers, An Integrated Approach to Design*. London: Springer-Verlag, 2001.
- [7] 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.
- [8] —, “Distribution-free performance bounds for potential function rules,” *IEEE Transactions on Information Theory*, vol. 25, pp. 601–604, 1979.
- [9] D. Foley, “Considerations of sample and feature size,” *IEEE Transactions on Information Theory*, vol. IT-18, pp. 618–626, 1972.
- [10] N. Glick, “Sample-based multinomial classification,” *Biometrics*, vol. 29, pp. 241–256, 1973.
- [11] M. Goldstein and E. Wolf, “On the problem of bias in multinomial classification,” *Biometrics*, vol. 33, pp. 325–331, 1977.
- [12] U. M. Braga-Neto and E. R. Dougherty, “Exact correlation between actual and estimated errors in discrete classification,” *Pattern Recognition Letters*, vol. 31, pp. 407–413, 2010.
- [13] —, “Exact performance measures and distributions of error estimators for discrete classifiers,” *Pattern Recognition*, vol. 38, pp. 1799–1814, 2005.
- [14] A. Zollanvari, U. Braga-Neto, and E. Dougherty, “On the sampling distribution of resubstitution and leave-one-out error estimators for linear classifiers,” *Pattern Recognition*, vol. 42, pp. 2705–2723, 2009.
- [15] —, “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.
- [16] F. H. Schröder, J. Hugosson, M. J. Roobol, and et. al., “Screening and prostate-cancer mortality in a randomized european study,” *New Eng. J. Med.*, vol. 360, pp. 1320–1328, 2009.

- [17] C. J. L. Koelinka, P. van Hasseltb, A. van der Ploegc, M. M. van den Heuvel-Eibrinkd, F. A. Wijburge, C. M. A. Bijlevelda, and F. J. van Spronsen, “Tyrosinemia type i treated by ntbc: How does afp predict liver cancer?” *Molecular Genetics and Metabolism*, vol. 89, pp. 310–315, 2006.
- [18] R. C. J. Bast, F. J. Xu, Y. H. Yu, S. Barnhill, Z. Zhang, and G. B. Mills, “Ca 125: the past and the future,” *Int J Biol Markers*, vol. 13, pp. 179–187, 1998.
- [19] X. Filella, R. Molina, J. J. Grau, J. M. PiquŽ, J. C. Garcia-Valdecasas, E. Astudillo, A. Biete, J. M. Bordas, A. Novell, and E. Campo, “Prognostic value of ca 19.9 levels in colorectal cancer,” *Molecular Genetics and Metabolism*, vol. 216, pp. 55–59, 1992.
- [20] T. S. Frank, A. M. Deffenbaugh, J. E. Reid, M. Hulick, B. E. Ward, B. Lingenfelter, K. L. Gumpfer, T. Scholl, S. V. Tavtigian, D. R. Pruss, and G. C. Critchfield, “Clinical characteristics of individuals with germline mutations in brca1 and brca2: Analysis of 10,000 individuals,” *Journal of Clinical Oncology*, vol. 20, pp. 1480–1490, 2002.
- [21] K. Syrjakoski, T. Kuukasjarvi, K. Waltering, K. Haraldsson, A. Auvinen, A. Borg, T. Kainu, O. Kallioniemi, and P. A. Koivisto, “Brca2 mutations in 154 finnish male breast cancer patients,” *Neoplasia*, vol. 6, pp. 541–545, 2004.
- [22] A. Horii, S. Nakatsuru, Y. Miyoshi, S. Ichii, H. Nagase, H. Ando, A. Yanagisawa, E. Tsuchiya, Y. Kato, and Y. Nakamura, “Frequent somatic mutations of the apc gene in human pancreatic cancer,” *Cancer Research*, vol. 52, pp. 6696–6698, 1992.
- [23] C. Smith, “Some examples of discrimination,” *Annals of Eugenics*, vol. 18, pp. 272–282, 1947.
- [24] P. Lachenbruch and M. Mickey, “Estimation of error rates in discriminant analysis,” *Technometrics*, vol. 10, pp. 1–11, 1968.
- [25] A. Zollanvari, U. Braga-Neto, and E. Dougherty, “On the sampling distribution of resubstitution and leave-one-out error estimators for linear classifiers,” *Pattern Recognition*, vol. 42, no. 11, pp. 2705–2723, 2009.
- [26] S. J. Raudys and A. K. Jain, “Small sample size effects in statistical pattern recognition: Recommendations for practitioners,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 252–264, 1991.
- [27] 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.
- [28] L. Devroye, L. Gyorfı, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer, 1996.
- [29] M. van de Vijver, Y. He, 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.