

Joint Sampling Distribution Between Actual and Estimated Classification Errors for Linear Discriminant Analysis¹

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

Abstract

Error estimation must be used to find the accuracy of a designed classifier, an issue that is critical in biomarker discovery for disease diagnosis and prognosis in genomics and proteomics. This paper presents, for what is believed to be the first time, the analytical formulation for the joint sampling distribution of the actual and estimated errors of a classification rule. The analysis presented here concerns the Linear Discriminant Analysis (LDA) classification rule and the resubstitution and leave-one-out error estimators, under a general parametric Gaussian assumption. Exact results are provided in the univariate case, and a simple method is suggested to obtain an accurate approximation in the multivariate case. It is also shown how these results can be applied in the computation of condition bounds and the regression of the actual error, given the observed error estimate. In contrast to asymptotic results, the analysis presented here is applicable to finite training data. In particular, it applies in the small-sample settings commonly found in genomics and proteomics applications. Numerical examples, which include parameters estimated from actual microarray data, illustrate the analysis throughout.

Keywords: Classification, Error Estimation, Linear Discriminant Analysis, Sampling Distribution, Resubstitution, Leave-One-Out, Cross-Validation.

1 Introduction

The main problem today in translational genomics and proteomics is the discovery of biomarker panels for disease diagnosis and prognosis, and this is a problem that depends critically on the design of accurate classifiers. The actual error rate of a classifier, however, can be computed exactly only if the underlying feature-label distribution of the problem is known, which is almost never the case in practice. One must then apply *error estimation* methods to assess the performance of the classifier based on the available data. In addition, with the emergence of high-throughput measurement technologies, medical applications are now often characterized by an extremely large number of measurements made on a small number of samples, which creates significant challenges in the statistical analysis and interpretation of such data, in particular, difficult challenges in the application of error estimation methods.

For example, it is reported in [1] that re-analysis of data from the seven largest published microarray-based studies that have attempted to predict prognosis of cancer patients reveals that five of those seven did not really classify patients better than chance. In another paper [2], the authors reviewed 90 studies, 76% of which were published in journals having impact factor larger than 6, and report that 50% of them are flawed. This state of affairs can be attributed in part to the misuse of error estimation techniques in small-sample situations [3–6].

The resubstitution error estimator [7] and the leave-one-out cross-validation error estimator (variously credited to [8–11]) are the focus of the present paper and have been used extensively in the literature dealing

¹Zollanvari and Braga-Neto (corresponding author) are with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX. Dougherty is with Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX and with Translational Genomics Institute, Phoenix, AZ. This work was supported by the National Science Foundation, through NSF awards CCF-0845407 (Braga-Neto) and CCF-0634794 (Dougherty).

with small-sample biological high-throughput data – for instance, see [12–17], to cite just a few articles. It is noteworthy that some of these cited works have been subsequently criticized for lack of reproducible results due to the improper use of resubstitution and leave-one-out error estimation [2, 18].

Thus, there is a pressing need to study the performance of error estimators in the small-sample settings common in the biomarker discovery applications. Here we point out that all performance metrics of interest for an error estimator, such as bias, variance, RMS, correlation with the actual error, and confidence intervals, can be determined if one has knowledge of the joint sampling distribution between the actual and estimated errors. This observation highlights the importance of deriving exact or approximate expressions for such distributions in the study of error estimation.

Most of the existing analytic results that characterize performance of error estimators are of a large-sample or asymptotic nature. For example, concerning the discrete histogram classifier, it can be shown that the sampling variance of the resubstitution error estimator is $O(1/n)$, where n is the sample size, whereas the variance of the leave-one-out error estimator is $O(1/\sqrt{n})$; thus the latter may tend to zero much more slowly, and in fact does for some distributions [19]. This makes resubstitution preferable in an asymptotic sense. However, as a general rule, asymptotic results are unhelpful in small-sample applications, where the bounds on performance typically become too slack to be meaningful. As an example, consider the comment by D. Hand on asymptotic results by Kittler and Devijver on the variance of so-called average conditional error rate estimators [20]: “Unfortunately, as Kittler and Devijver point out, small-sample performance of these average conditional error rate estimators often does not live up to asymptotic promise” [21].

This may be contrasted to the classical approach to study small-sample performance, proposed at the dawn of modern mathematical statistics, when, in a series of seminal papers, R.A. Fisher derived the sampling distribution of the correlation coefficient, under a parametric Gaussian assumption [22–24]. Fisher’s groundbreaking results meant that all the performance metrics of the correlation coefficient, estimated from a finite population sample, could be determined exactly from the parameters of the population distribution.

This approach was vigorously followed in the early literature on pattern recognition (known as “discriminant analysis” in the statistical community). For classification of continuous variables, there is a wealth of results on the properties of the classification error of the classification rule known as Linear Discriminant Analysis (LDA), which is a simple classification rule, and often more effective in small-sample settings than more complex classification rules [25, 26] — LDA has a long history, having been originally based on an idea by R. Fisher (the “Fisher discriminant”) [27], developed by A. Wald [28], and given the form known today by T.W. Anderson [29]. For example, the exact distribution and expectation of the actual classification error were determined by S. John in the univariate case, and in the multivariate case by assuming that the covariance matrix Σ is known in the formulation of the discriminant [30]. The case when Σ is not known, and the sample covariance matrix S is used in discrimination, is very difficult, and John gives only an asymptotic approximation to the distribution of the actual error. In a publication that appeared in the same year, R. Sitgreaves gives the exact distribution for this case, in terms of an infinite series, when the numbers of samples in each class are equal [31]. Several classical papers have studied the distribution and moments of the actual classification error of LDA under a parametric Gaussian assumption, using exact, approximate, and asymptotic methods [32–39].

Results for estimated classification error are comparatively scarcer in the literature. We briefly summarize these, noting that all pertain to LDA under a parametric Gaussian model. Hills provides an exact formula for the expected resubstitution error estimate in the univariate case, which involves the bivariate Gaussian cumulative distribution [39], and Moran extends this result to the multivariate case when Σ is known in the formulation of the discriminant [40]. Moran’s result can also be seen as a generalization of a similar result given by John in [30] for the expectation of the actual error. McLachlan provides an asymptotic expression for the expected resubstitution error in the multivariate case, for unknown covariance matrix [41], with a

similar result having been provided by Raudys in [42]. In [43], Foley derives an asymptotic expression for the variance of the resubstitution error.

The analytical study of error estimation has waned considerably since the 1970's, and researchers have turned mostly to Monte-Carlo simulation studies. A hint why is provided in a 1978 paper by Jain and Waller, in which they write, "The exact expression for the average probability of error using the W statistic with the restriction that $N_1 = N_2 = N$ as given by Sitgreaves [31] is very complicated and difficult to evaluate even on a computer" [44].

The present paper furthers the analytical study of error estimation by deriving, for what is believed to be the first time, the analytical formulation for the joint sampling distribution of the actual and estimated errors for a classification rule. We consider here the LDA classification rule and the resubstitution and leave-one-out error estimators, under a general parametric Gaussian assumption. This work extends the results in a previous publication by the same authors [45], where the marginal sampling distributions of the resubstitution and leave-one-out error estimators for LDA under a Gaussian model were determined.

This paper is organized as follows. Section 2 provides definitions and introduces the mathematical notation. Section 3 presents the core results of the paper, which are valid for the univariate case, and provide an exact method to determine both the joint distribution and joint density between actual and estimated errors given the model parameters. Section 4 presents a simple method to obtain an approximation of the joint distribution in the multivariate case; numerical examples show the approximation is accurate under certain conditions. Section 5 shows how the results in the previous two sections can be applied in the computation of condition bounds and the regression of the actual error, given the observed error estimate. Finally, Section 6 presents our concluding remarks. Proofs are given in an Appendix section.

2 Mathematical Preliminaries

Consider a set of $n = n_0 + n_1$ i.i.d. samples, where n_0 samples $\{X_1, X_2, \dots, X_{n_0}\}$ come from a population Π_0 , and n_1 samples $\{X_{n_0+1}, X_{n_0+2}, \dots, X_{n_0+n_1}\}$ come from a population Π_1 . *Linear Discriminant Analysis* (LDA) employs Anderson's W discriminant, which is defined as follows:

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

where X comes from the mixture of the populations $p\Pi_0 + (1-p)\Pi_1$, $0 < p < 1$ (in this paper, we assume, for simplicity, that $p = \frac{1}{2}$), and

$$\begin{aligned} \hat{\mu}_0 &= \frac{1}{n_0} \sum_{i=1}^{n_0} X_i \\ \hat{\mu}_1 &= \frac{1}{n_1} \sum_{i=n_0+1}^{n_0+n_1} X_i \end{aligned} \quad (2)$$

are the sample means for each population Π_i , i.e., each *class* i , for $i = 0, 1$. A method to assign a future sample $X = x$ to one of the classes is provided by the designed LDA classifier:

$$\psi(x) = \begin{cases} 1, & \text{if } W(x) < 0 \\ 0, & \text{if } W(x) \geq 0 \end{cases}, \quad (3)$$

that is, the sign of $W(x)$ determines the classification of x . Here we are assuming that the covariance matrix Σ is known; in particular, the W statistic is not a function of the sample covariance matrix $\hat{\Sigma}$.

On the other hand, the *Nearest Mean Classifier* (NMC) avoids having to know or estimate the covariance matrix:

$$\psi(x) = \begin{cases} 1 & \text{if } (x - \hat{\mu})^T(\hat{\mu}_0 - \hat{\mu}_1) < 0 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where

$$\hat{\mu} = \frac{\hat{\mu}_0 + \hat{\mu}_1}{2}. \quad (5)$$

The decision region corresponding to this classifier is a hyperplane passing through $\hat{\mu}$ and perpendicular to the axis joining $\hat{\mu}_1$ and $\hat{\mu}_0$, and x is classified to class 1 if x lies on the same side of the hyperplane as $\hat{\mu}_1$, and to 0 otherwise. This is equivalent to saying that x is assigned to the class of the closest sample mean. It is easy to see that in the univariate case, the LDA classifier reduces necessarily to the NMC classifier.

The resubstitution error estimator [7], also called the apparent error of classifier ψ , is given by

$$\hat{\varepsilon}_r = \frac{1}{n} \left[\sum_{i=1}^{n_0} I_{\{W(X_i) < 0\}} + \sum_{i=n_0+1}^{n_0+n_1} I_{\{W(X_i) \geq 0\}} \right] \quad (6)$$

where I is an indicator variable, defined by $I_{\{u\}} = 1$ if condition u holds, with $I_{\{u\}} = 0$ otherwise. On the other hand, the leave-one-out error estimator [8] for the LDA classification rule is given by

$$\hat{\varepsilon}_l = \frac{1}{n} \left[\sum_{i=1}^{n_0} I_{\{W^{(i)}(X_i) < 0\}} + \sum_{i=n_0+1}^{n_0+n_1} I_{\{W^{(i)}(X_i) \geq 0\}} \right] \quad (7)$$

where $W^{(i)}$ is the discriminant obtained when sample X_i is left out of training.

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

$$P \left(\hat{\varepsilon} = \frac{k}{n_0 + n_1}, \varepsilon < z \right), \quad k = 0, 1, \dots, n_0 + n_1, 0 \leq z \leq 1, \quad (8)$$

where ε is the actual classification error rate, and $\hat{\varepsilon}$ is either the resubstitution estimator $\hat{\varepsilon}_r$ or the leave-one-out estimator $\hat{\varepsilon}_l$, in the case where the classes are Gaussian distributed. By simple summation along the discrete variable, this allows one to easily compute the associated joint (cumulative) distribution functions, if so desired. More importantly, from the expressions for the joint probability in (8), one can compute the exact bias, deviation variance, and RMS of estimation (in terms of the mean, variance and second moment of $\hat{\varepsilon} - \varepsilon$), as well as exact conditional probability $P(\varepsilon < z \mid \hat{\varepsilon})$, which leads to the computation of exact conditional bounds on the actual error, as well as the exact regression $E[\varepsilon \mid \hat{\varepsilon}]$ of the actual on the estimated error, as will be detailed in Section 5.

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

$$p \left(\hat{\varepsilon} = \frac{k}{n_0 + n_1}, \varepsilon = z \right), \quad k = 0, 1, \dots, n_0 + n_1, 0 \leq z \leq 1, \quad (9)$$

where $\hat{\varepsilon}$ is again either the resubstitution estimator $\hat{\varepsilon}_r$ or the leave-one-out estimator $\hat{\varepsilon}_l$, in the case where the classes are Gaussian distributed. Note that, even though we are using the terminology ‘‘density,’’ the quantity in (9) is in fact a combination of density in ε and probability mass function in $\hat{\varepsilon}$.

3 Univariate Case

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

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

$$\psi(x) = \begin{cases} 0, & \text{if } W(x) = (x - \hat{\mu})(\hat{\mu}_0 - \hat{\mu}_1) > 0 \\ 1, & \text{otherwise} \end{cases} \quad (10)$$

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

3.1 Resubstitution

From (10), we see that ψ can be written simply as $\psi(x) = I_{\{x < \hat{\mu}\}}$, if $\hat{\mu}_0 > \hat{\mu}_1$, i.e., sample means are on the same side of the cutpoint $\hat{\mu}$ as the corresponding actual means, and $\psi(x) = I_{\{x > \hat{\mu}\}}$, if $\hat{\mu}_0 < \hat{\mu}_1$, i.e., sample means are on the wrong side of the cutpoint (the case $\hat{\mu}_0 = \hat{\mu}_1$ having probability 0). The first case may be called “direct” classification, while the second case characterizes “reverse” classification.

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

$$\varepsilon^\uparrow(w) = \frac{1}{2} \left[\Phi \left(\frac{w - \mu_0}{\sigma_0} \right) + \Phi \left(\frac{\mu_1 - w}{\sigma_1} \right) \right], \quad (11)$$

and

$$\varepsilon^\downarrow(w) = 1 - \varepsilon^\uparrow(w) = \frac{1}{2} \left[\Phi \left(\frac{\mu_0 - w}{\sigma_0} \right) + \Phi \left(\frac{w - \mu_1}{\sigma_1} \right) \right], \quad (12)$$

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

The actual error for the classifier ψ in (10) is a function of $\hat{\mu}$ and of the “direction” of classification:

$$\varepsilon = \begin{cases} \varepsilon^\uparrow(\hat{\mu}), & \hat{\mu}_0 > \hat{\mu}_1 \text{ (direct classification)} \\ \varepsilon^\downarrow(\hat{\mu}), & \hat{\mu}_0 < \hat{\mu}_1 \text{ (reverse classification)} \end{cases} \quad (13)$$

3.1.1 Equal-Variance Case

In this section, it is assumed that $\sigma_0 = \sigma_1 = \sigma$ (this assumption will be dropped in the next Section). The restriction $\varepsilon < z$ in (8) puts a corresponding restriction on where $\hat{\mu}$ may lie on the real line, which in turn affects the derivation of the joint probability in (8). For direct classification, ε is always under 0.5, while for reverse classification, ε is always above 0.5. In addition, if ε^* denotes the optimal (Bayes) classification error, then

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

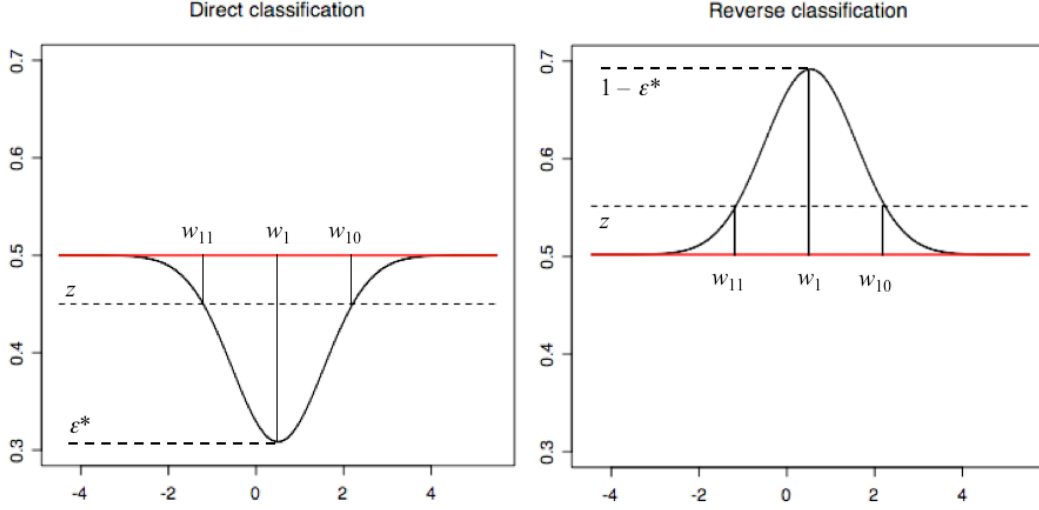


Figure 1: Plots of actual error as a function of $\hat{\mu}$, for $\mu_0 = 1$, $\mu_1 = 0$, and $\sigma_0 = \sigma_1 = 1$. Left: plot of $\varepsilon^\uparrow(w)$, direct classification ($\hat{\mu}_0 > \hat{\mu}_1$). Right: plot of $\varepsilon^\downarrow(w)$, reverse classification ($\hat{\mu}_0 < \hat{\mu}_1$).

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

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

$$[\varepsilon < z] = \begin{cases} \emptyset, & \text{for } z < \varepsilon^* \\ [\hat{\mu} \in (w_{11}, w_{10}), \hat{\mu}_0 > \hat{\mu}_1], & \text{for } \varepsilon^* \leq z \leq 0.5 \\ [\hat{\mu}_0 > \hat{\mu}_1] \cup [\hat{\mu} \notin (w_{11}, w_{10}), \hat{\mu}_0 < \hat{\mu}_1], & \text{for } 0.5 < z \leq 1 - \varepsilon^* \\ \Omega, & \text{for } z > 1 - \varepsilon^* \end{cases} \quad (14)$$

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

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

$$P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \varepsilon < z\right) = \begin{cases} 0, & \text{for } z < \varepsilon^* \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} \in (w_{11}, w_{10}), \hat{\mu}_0 > \hat{\mu}_1\right), & \text{for } \varepsilon^* \leq z \leq 0.5 \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu}_0 > \hat{\mu}_1\right) + \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} \notin (w_{11}, w_{10}), \hat{\mu}_0 < \hat{\mu}_1\right), & \text{for } 0.5 < z \leq 1 - \varepsilon^* \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}\right), & \text{for } z > 1 - \varepsilon^* \end{cases} \quad (15)$$

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

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

$$\begin{aligned}
P(\hat{\varepsilon}_r = 0, \hat{\mu} \in (a, b), \hat{\mu}_0 > \hat{\mu}_1) &= P(Z_1 > \mathbf{0}) \\
P(\hat{\varepsilon}_r = 0, \hat{\mu} \notin (a, b), \hat{\mu}_0 < \hat{\mu}_1) &= P(Z_2 < \mathbf{0}) + P(Z_3 < \mathbf{0}) \\
P(\hat{\varepsilon}_r = 0, \hat{\mu}_0 > \hat{\mu}_1) &= P(Z_4 > \mathbf{0}) \\
P(\hat{\varepsilon}_r = 0) &= P(Z_4 > \mathbf{0}) + P(Z_4 < \mathbf{0})
\end{aligned} \tag{16}$$

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

$$\begin{bmatrix} (\mu_0 - \mu_1)\mathbf{1}_{n_0+n_1+1} \\ (\mu_0 + \mu_1) - 2a \\ -(\mu_0 + \mu_1) + 2b \end{bmatrix} \tag{17}$$

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

$$H_{ij} = \begin{cases} \frac{4n_0-3}{n_0} + \frac{1}{n_1}, & i, j = 1, \dots, n_0, i = j \\ -\frac{3}{n_0} + \frac{1}{n_1}, & i, j = 1, \dots, n_0, i \neq j \\ \frac{1}{n_0} + \frac{4n_1-3}{n_1}, & i, j = n_0 + 1, \dots, n_0 + n_1, i = j \\ \frac{1}{n_0} - \frac{3}{n_1}, & i, j = n_0 + 1, \dots, n_0 + n_1, i \neq j \\ \frac{1}{n_0} - \frac{1}{n_1}, & \begin{cases} i = n_0 + n_1 + 2, j = 1, \dots, n_0 + n_1 + 1 \\ j = n_0 + n_1 + 2, i = 1, \dots, n_0 + n_1 + 1 \end{cases} \\ \frac{1}{n_1} - \frac{1}{n_0}, & \begin{cases} i = n_0 + n_1 + 3, j = 1, \dots, n_0 + n_1 + 1 \\ j = n_0 + n_1 + 3, i = 1, \dots, n_0 + n_1 + 1 \end{cases} \\ -\frac{1}{n_0} - \frac{1}{n_1}, & \begin{cases} i = n_0 + n_1 + 2, j = n_0 + n_1 + 3 \\ i = n_0 + n_1 + 3, j = n_0 + n_1 + 2 \end{cases} \\ \frac{1}{n_0} + \frac{1}{n_1}, & \text{otherwise} \end{cases} \tag{18}$$

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

Proof. See Appendix.

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

$$\begin{aligned}
P([k \text{ errors}]) &= \sum_{l=0}^k P([l \text{ errors in class 0 and } k-l \text{ errors in class 1}]) \\
&= \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} P([X_1, \dots, X_l \text{ in error and } X_{n_0+1}, \dots, X_{n_0+l-k} \text{ in error}])
\end{aligned} \tag{19}$$

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

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

$$\begin{aligned}
P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} \in (a, b), \hat{\mu}_0 > \hat{\mu}_1\right) &= \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} P(E_{l,k-l}^2 Z_1 > \mathbf{0}) \\
P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} \notin (a, b), \hat{\mu}_0 < \hat{\mu}_1\right) &= \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} [P(E_{l,k-l}^1 Z_2 < \mathbf{0}) + P(E_{l,k-l}^1 Z_3 < \mathbf{0})] \\
P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu}_0 > \hat{\mu}_1\right) &= \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} P(E_{l,k-l}^0 Z_4 > \mathbf{0}) \\
P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}\right) &= \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} [P(E_{l,k-l}^0 Z_4 > \mathbf{0}) + P(E_{l,k-l}^0 Z_4 < \mathbf{0})]
\end{aligned} \tag{20}$$

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

Theorem 2, in conjunction with Lemma 1, allows the exact computation of the joint probability in (8) for the resubstitution error estimator. The probabilities of the kind $P(Z > 0)$, where Z is a Gaussian vector, which are needed in the computations above, can be readily computed using an algorithm for integration of multivariate Gaussian densities over rectangular regions, due to Genz and Bretz [46]. This provides an efficient and very accurate method for the exact computation of the joint probability in (8).

3.1.2 Unequal-Variance Case

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

$$\begin{aligned}
w_1 &= \frac{\mu_1 \sigma_0^2 - \mu_0 \sigma_1^2 + \sigma_0 \sigma_1 \sqrt{(\mu_1 - \mu_0)^2 + 2(\sigma_1^2 - \sigma_0^2) \ln \frac{\sigma_1}{\sigma_0}}}{\sigma_0^2 - \sigma_1^2} \\
w_2 &= \frac{\mu_1 \sigma_0^2 - \mu_0 \sigma_1^2 - \sigma_0 \sigma_1 \sqrt{(\mu_1 - \mu_0)^2 + 2(\sigma_1^2 - \sigma_0^2) \ln \frac{\sigma_1}{\sigma_0}}}{\sigma_0^2 - \sigma_1^2},
\end{aligned} \tag{21}$$

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

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

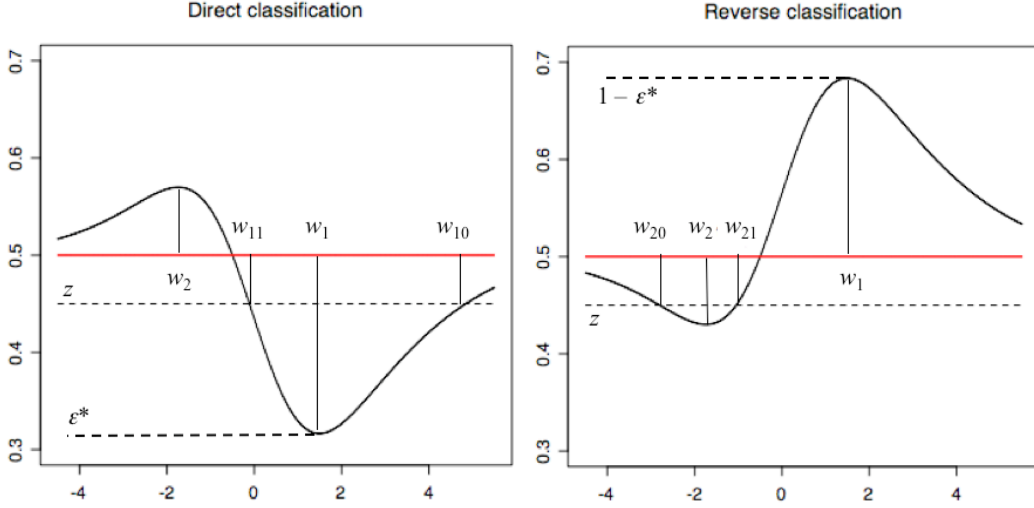


Figure 2: Plots of actual error as a function of $\hat{\mu}$, for $\mu_0 = 1$, $\mu_1 = 0$, $\sigma_0 = 3$, $\sigma_1 = 1$ and $\varepsilon^\downarrow(w_2) < z \leq 0.5$. Left: plot of $\varepsilon^\uparrow(w)$, direct classification ($\hat{\mu}_0 > \hat{\mu}_1$). Right: plot of $\varepsilon^\downarrow(w)$,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

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

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

$$[\varepsilon < z] = \begin{cases} \emptyset, & \text{for } z < \varepsilon^* \\ [\hat{\mu} \in (w_{11}, w_{10}), \hat{\mu}_0 > \hat{\mu}_1], & \text{for } \varepsilon^* \leq z \leq \varepsilon^\downarrow(w_2) \\ [\hat{\mu} \in (w_{11}, w_{10}), \hat{\mu}_0 > \hat{\mu}_1] \cup [\hat{\mu} \in (w_{21}, w_{20}), \hat{\mu}_0 < \hat{\mu}_1], & \text{for } \varepsilon^\downarrow(w_2) < z \leq 0.5 \\ [\hat{\mu} \notin (w_{11}, w_{10}), \hat{\mu}_0 < \hat{\mu}_1] \cup [\hat{\mu} \notin (w_{21}, w_{20}), \hat{\mu}_0 > \hat{\mu}_1], & \text{for } 0.5 < z \leq \varepsilon^\uparrow(w_2) \\ [\hat{\mu}_0 > \hat{\mu}_1] \cup [\hat{\mu} \notin (w_{11}, w_{10}), \hat{\mu}_0 < \hat{\mu}_1], & \text{for } \varepsilon^\uparrow(w_2) < z \leq 1 - \varepsilon^* \\ \Omega, & \text{for } z > 1 - \varepsilon^* \end{cases} \quad (22)$$

where the cutpoints $w_{11} < w_{10}$ and $w_{21} < w_{20}$ can be found easily in each case by numerical inversion of the respective function ε^\uparrow or ε^\downarrow , such that $w_1 \in (w_{11}, w_{10})$ and $w_2 \in (w_{21}, w_{20})$. We have thus established the following Lemma.

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

$$P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \varepsilon < z\right) = \begin{cases} 0, & \text{for } z < \varepsilon^* \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} \in (w_{11}, w_{10}), \hat{\mu}_0 > \hat{\mu}_1\right), & \text{for } \varepsilon^* \leq z \leq \varepsilon^\downarrow(w_2) \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} \in (w_{11}, w_{10}), \hat{\mu}_0 > \hat{\mu}_1\right) + \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} \in (w_{21}, w_{20}), \hat{\mu}_0 < \hat{\mu}_1\right), & \text{for } \varepsilon^\downarrow(w_2) < z \leq 0.5 \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} \notin (w_{11}, w_{10}), \hat{\mu}_0 < \hat{\mu}_1\right) + \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} \notin (w_{21}, w_{20}), \hat{\mu}_0 > \hat{\mu}_1\right), & \text{for } 0.5 < z \leq \varepsilon^\uparrow(w_2) \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu}_0 > \hat{\mu}_1\right) + \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} \notin (w_{11}, w_{10}), \hat{\mu}_0 < \hat{\mu}_1\right), & \text{for } \varepsilon^\uparrow(w_2) < z \leq 1 - \varepsilon^* \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}\right), & \text{for } z > 1 - \varepsilon^* \end{cases} \quad (23)$$

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

Theorem 3. Let $X_i \sim N(\mu_0, \sigma_0^2)$ for $i = 1, \dots, n_0$, and $X_i \sim N(\mu_1, \sigma_1^2)$ for $i = n_0 + 1, \dots, n_0 + n_1$ be i.i.d. observations used to derive the classifier in (10). Then

$$\begin{aligned} P(\hat{\varepsilon}_r = 0, \hat{\mu} \in (a, b), \hat{\mu}_0 > \hat{\mu}_1) &= P(Z_1 > \mathbf{0}) \\ P(\hat{\varepsilon}_r = 0, \hat{\mu} \in (a, b), \hat{\mu}_0 < \hat{\mu}_1) &= P(Z'_1 < \mathbf{0}) \\ P(\hat{\varepsilon}_r = 0, \hat{\mu} \notin (a, b), \hat{\mu}_0 < \hat{\mu}_1) &= P(Z_2 < \mathbf{0}) + P(Z_3 < \mathbf{0}) \\ P(\hat{\varepsilon}_r = 0, \hat{\mu} \notin (a, b), \hat{\mu}_0 > \hat{\mu}_1) &= P(Z'_2 > \mathbf{0}) + P(Z'_3 > \mathbf{0}) \\ P(\hat{\varepsilon}_r = 0, \hat{\mu}_0 > \hat{\mu}_1) &= P(Z_4 > \mathbf{0}) \\ P(\hat{\varepsilon}_r = 0) &= P(Z_4 > \mathbf{0}) + P(Z_4 < \mathbf{0}) \end{aligned} \quad (24)$$

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

$$\mu_{Z_1} = \begin{bmatrix} (\mu_0 - \mu_1)\mathbf{1}_{n_0 + n_1 + 1} \\ (\mu_0 + \mu_1) - 2a \\ -(\mu_0 + \mu_1) + 2b \end{bmatrix} \quad (25)$$

and covariance matrix Σ_{Z_1} given by

$$(\Sigma_{Z_1})_{ij} = \begin{cases} (4n_0 - 3)\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & i, j = 1, \dots, n_0, i = j \\ -3\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & i, j = 1, \dots, n_0, i \neq j \\ \frac{\sigma_0^2}{n_0} + (4n_1 - 3)\frac{\sigma_1^2}{n_1}, & i, j = n_0 + 1, \dots, n_0 + n_1, i = j \\ \frac{\sigma_0^2}{n_0} - 3\frac{\sigma_1^2}{n_1}, & i, j = n_0 + 1, \dots, n_0 + n_1, i \neq j \\ \frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1}, & \begin{cases} i = n_0 + n_1 + 2, j = 1, \dots, n_0 + n_1 + 1 \\ j = n_0 + n_1 + 2, i = 1, \dots, n_0 + n_1 + 1 \end{cases} \\ \frac{\sigma_1^2}{n_1} - \frac{\sigma_0^2}{n_0}, & \begin{cases} i = n_0 + n_1 + 3, j = 1, \dots, n_0 + n_1 + 1 \\ j = n_0 + n_1 + 3, i = 1, \dots, n_0 + n_1 + 1 \end{cases} \\ -\left(\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\right), & \begin{cases} i = n_0 + n_1 + 2, j = n_0 + n_1 + 3 \\ i = n_0 + n_1 + 3, j = n_0 + n_1 + 2 \end{cases} \\ \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & \text{otherwise} \end{cases}, \quad (26)$$

Here Z'_1 is a Gaussian random vector of size $n_0 + n_1 + 3$, obtained from Z_1 by multiplying by -1 the last two components of Z_1 . Furthermore, Z_2 (resp. Z_3) is a Gaussian random vector of size $n_0 + n_1 + 2$, obtained from Z_1 by eliminating component $n_0 + n_1 + 3$ (resp. $n_0 + n_1 + 2$), while Z'_2 (resp. Z'_3) is a Gaussian random vector of size $n_0 + n_1 + 2$, obtained from Z'_1 by eliminating component $n_0 + n_1 + 3$ (resp. $n_0 + n_1 + 2$). Finally, Z_4 is Gaussian random vector of size $n_0 + n_1 + 1$, obtained from Z_1 by eliminating both components $n_0 + n_1 + 2$ and $n_0 + n_1 + 3$.

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

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

$$\begin{aligned} P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} \in (a, b), \hat{\mu}_0 > \hat{\mu}_1\right) &= \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} P(E_{l, k-l}^2 Z_1 > \mathbf{0}) \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} \in (a, b), \hat{\mu}_0 < \hat{\mu}_1\right) &= \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} P(E_{l, k-l}^2 Z'_1 < \mathbf{0}) \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} \notin (a, b), \hat{\mu}_0 < \hat{\mu}_1\right) &= \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} [P(E_{l, k-l}^1 Z_2 < \mathbf{0}) + P(E_{l, k-l}^1 Z_3 < \mathbf{0})] \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} \notin (a, b), \hat{\mu}_0 > \hat{\mu}_1\right) &= \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} [P(E_{l, k-l}^1 Z'_2 > \mathbf{0}) + P(E_{l, k-l}^1 Z'_3 > \mathbf{0})] \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu}_0 > \hat{\mu}_1\right) &= \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} P(E_{l, k-l}^0 Z_4 > \mathbf{0}) \\ P\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}\right) &= \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} [P(E_{l, k-l}^0 Z_4 > \mathbf{0}) + P(E_{l, k-l}^0 Z_4 < \mathbf{0})], \end{aligned} \quad (27)$$

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

Theorem 4, in conjunction with Lemma 2, allows the exact computation of the joint probability in (8) for the resubstitution error estimator. The probabilities of the kind $P(Z > 0)$, where Z is a Gaussian vector, which are needed in the computations above, can be readily computed using the algorithm for integration of multivariate Gaussian densities over rectangular regions due to Genz and Bretz [46]. This provides an efficient and very accurate method for the exact computation of the joint probability in (8) in the resubstitution case.

3.1.3 Joint Density

It is relatively easy to apply a methodology similar to the one in the previous sections to obtain the joint density in (9) for the resubstitution error estimator. Let the value of the gaussian density with mean μ and variance σ^2 at x be denoted by $\varphi(x, \mu, \sigma^2)$, and let $\psi(w) = |\varphi(x, \mu_0, \sigma_0^2) - \varphi(x, \mu_1, \sigma_1^2)|$. The following Lemma can be easily shown.

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

$$p\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \varepsilon = z\right) = \begin{cases} 0, & \text{for } z < \varepsilon^* \\ \frac{1}{\psi(w_{11})} p\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} = w_{11}, \hat{\mu}_0 > \hat{\mu}_1\right) + \\ \frac{1}{\psi(w_{10})} p\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} = w_{10}, \hat{\mu}_0 > \hat{\mu}_1\right), & \text{for } \varepsilon^* \leq z \leq \varepsilon^\downarrow(w_2) \\ \frac{1}{\psi(w_{11})} p\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} = w_{11}, \hat{\mu}_0 > \hat{\mu}_1\right) + \\ \frac{1}{\psi(w_{10})} p\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} = w_{10}, \hat{\mu}_0 > \hat{\mu}_1\right) + \\ \frac{1}{\psi(w_{21})} p\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} = w_{21}, \hat{\mu}_0 < \hat{\mu}_1\right) + \\ \frac{1}{\psi(w_{20})} p\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} = w_{20}, \hat{\mu}_0 < \hat{\mu}_1\right), & \text{for } \varepsilon^\downarrow(w_2) < z \leq 0.5 \\ \frac{1}{\psi(w_{21})} p\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} = w_{21}, \hat{\mu}_0 > \hat{\mu}_1\right) + \\ \frac{1}{\psi(w_{20})} p\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} = w_{20}, \hat{\mu}_0 > \hat{\mu}_1\right) + \\ \frac{1}{\psi(w_{11})} p\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} = w_{11}, \hat{\mu}_0 < \hat{\mu}_1\right) + \\ \frac{1}{\psi(w_{10})} p\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} = w_{10}, \hat{\mu}_0 < \hat{\mu}_1\right), & \text{for } 0.5 < z \leq \varepsilon^\uparrow(w_2) \\ \frac{1}{\psi(w_{11})} p\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} = w_{11}, \hat{\mu}_0 < \hat{\mu}_1\right) + \\ \frac{1}{\psi(w_{10})} p\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} = w_{10}, \hat{\mu}_0 < \hat{\mu}_1\right), & \text{for } \varepsilon^\uparrow(w_2) < z \leq 1 - \varepsilon^* \\ 0 & \text{for } z > 1 - \varepsilon^* \end{cases} \quad (28)$$

Lemma 3 holds for the case of equal variances $\sigma_0 = \sigma_1$, by considering only two regions with $z < 0.5$ and $z > 0.5$ and eliminating all terms that include w_{20} and w_{21} .

The following theorem specifies how to compute the terms on the right hand side of (28).

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

$$\begin{aligned}
p\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} = a, \hat{\mu}_0 > \hat{\mu}_1\right) &= \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} P(E_{l,k-l}^0 Y > \mathbf{0}) \varphi\left(0, \mu_0 + \mu_1 - 2a, \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\right) \\
p\left(\hat{\varepsilon}_r = \frac{k}{n_0 + n_1}, \hat{\mu} = a, \hat{\mu}_0 < \hat{\mu}_1\right) &= \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} P(E_{l,k-l}^0 Y < \mathbf{0}) \varphi\left(0, \mu_0 + \mu_1 - 2a, \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\right)
\end{aligned} \tag{29}$$

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

$$\mu_Y = 2 \frac{n_1 \sigma_0^2 (a - \mu_1) - n_0 \sigma_1^2 (a - \mu_0)}{n_1 \sigma_0^2 + n_0 \sigma_1^2} \mathbf{1}_{n_0 + n_1 + 1} \tag{30}$$

and covariance matrix Σ_Y given by

$$\Sigma_Y = \Sigma_{Y_{11}} - \frac{1}{n_0 n_1} \frac{(n_1 \sigma_0^2 - n_0 \sigma_1^2)^2}{n_1 \sigma_0^2 + n_0 \sigma_1^2} \mathbf{1}_{(n_0 + n_1 + 1) \times (n_0 + n_1 + 1)} \tag{31}$$

where:

$$\Sigma_{Y_{11}} = \begin{cases} (4n_0 - 3) \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & i, j = 1, \dots, n_0, i = j \\ -3 \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & i, j = 1, \dots, n_0, i \neq j \\ \frac{\sigma_0^2}{n_0} + (4n_1 - 3) \frac{\sigma_1^2}{n_1}, & i, j = n_0 + 1, \dots, n_0 + n_1, i = j \\ \frac{\sigma_0^2}{n_0} - 3 \frac{\sigma_1^2}{n_1}, & i, j = n_0 + 1, \dots, n_0 + n_1, i \neq j \\ \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & \text{otherwise} \end{cases} \tag{32}$$

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

Proof. See Appendix.

Theorem 5, in conjunction with Lemma 3, allows the exact computation of the joint density in (9) for the resubstitution error estimator.

3.1.4 Numerical Examples

Figures 3 and 4 display examples of the joint probability in (8) and the corresponding joint density in (9), respectively, for the resubstitution error estimator, computed using the expressions given previously.

3.2 Leave-One-Out

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

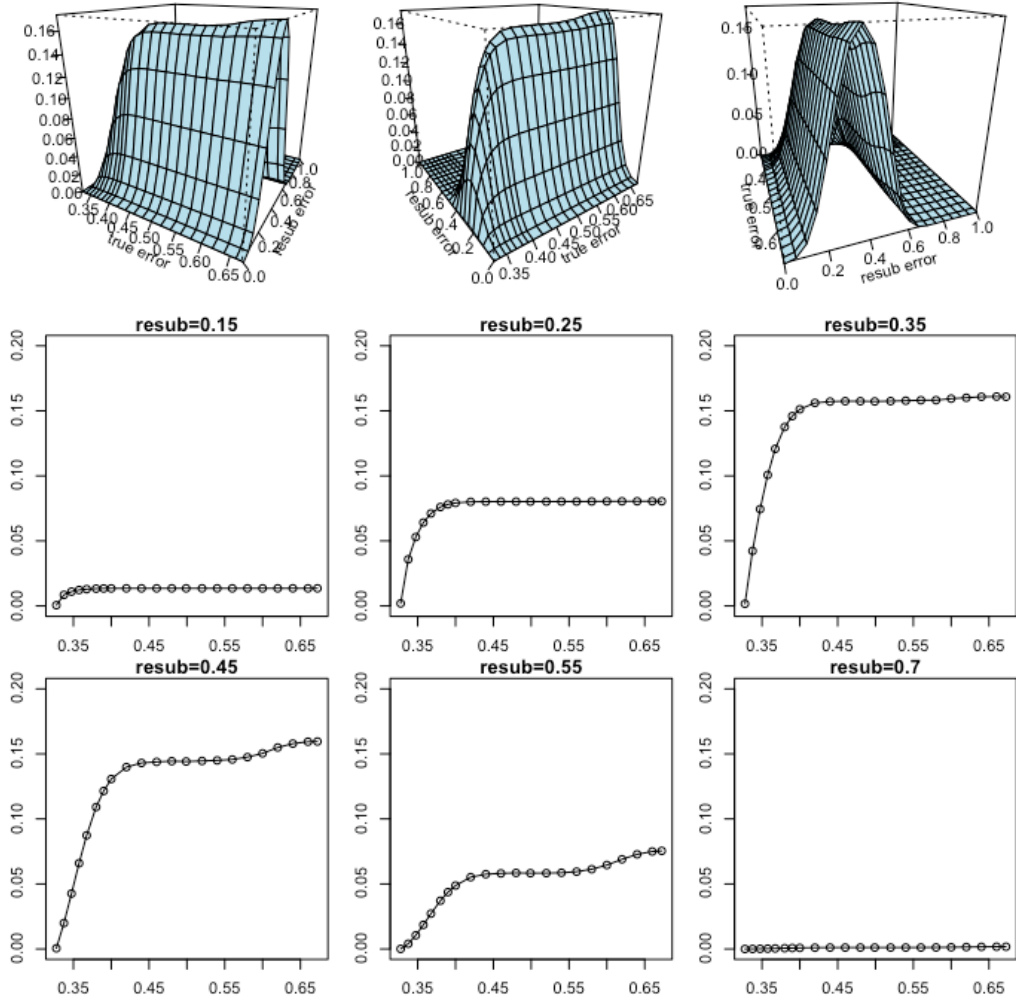


Figure 3: Joint probability in (8) for the resubstitution error estimator: $n_0 = n_1 = 10$, $m_0 = 1$, $m_1 = 0$, $\sigma_0 = 2$, $\sigma_1 = 1$. Bayes error = 0.32742.

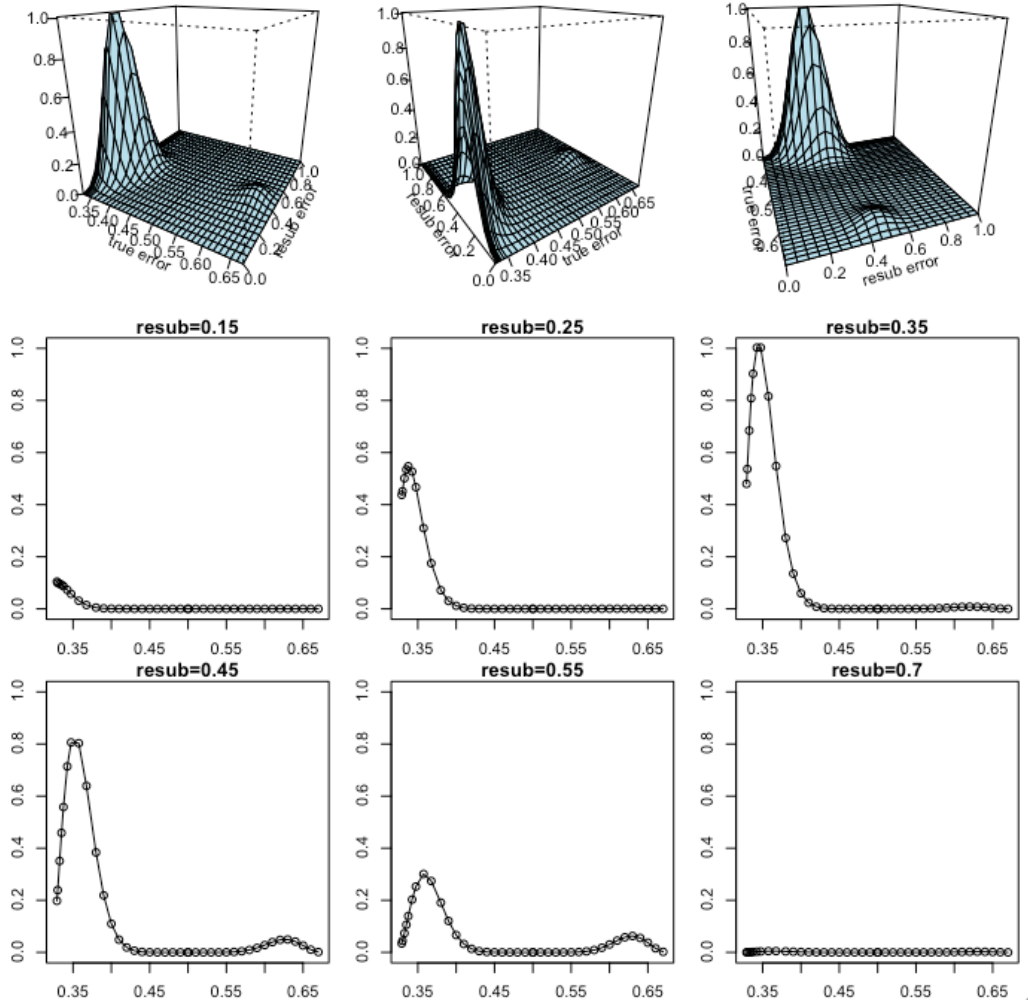


Figure 4: Joint density in (9) for the resubstitution error estimator: $n_0 = n_1 = 10$, $m_0 = 1, m_1 = 0$, $\sigma_0 = 2, \sigma_1 = 1$. Bayes error = 0.32742.

Theorem 6. Let $X_i \sim N(\mu_0, \sigma_0^2)$ for $i = 1, \dots, n_0$, and $X_i \sim N(\mu_1, \sigma_1^2)$ for $i = n_0 + 1, \dots, n_0 + n_1$ be i.i.d. observations used to derive the classifier in (10). Then

$$\begin{aligned}
P(\hat{\epsilon}_l=0, \hat{\mu} \in (a, b), \hat{\mu}_0 > \hat{\mu}_1) &= \sum_{m=0}^{n_0} \sum_{n=0}^{n_1} \binom{n_0}{m} \binom{n_1}{n} P(E_{m,n}^2 Z_1 > \mathbf{0}) \\
P(\hat{\epsilon}_l=0, \hat{\mu} \in (a, b), \hat{\mu}_0 < \hat{\mu}_1) &= \sum_{m=0}^{n_0} \sum_{n=0}^{n_1} \binom{n_0}{m} \binom{n_1}{n} P(E_{m,n}^2 Z_1' < \mathbf{0}) \\
P(\hat{\epsilon}_l=0, \hat{\mu} \notin (a, b), \hat{\mu}_0 < \hat{\mu}_1) &= \sum_{m=0}^{n_0} \sum_{n=0}^{n_1} \binom{n_0}{m} \binom{n_1}{n} \left(P(E_{m,n}^1 Z_2 < \mathbf{0}) + P(E_{m,n}^1 Z_3 < \mathbf{0}) \right) \\
P(\hat{\epsilon}_l=0, \hat{\mu} \notin (a, b), \hat{\mu}_0 > \hat{\mu}_1) &= \sum_{m=0}^{n_0} \sum_{n=0}^{n_1} \binom{n_0}{m} \binom{n_1}{n} \left(P(E_{m,n}^1 Z_2' > \mathbf{0}) + P(E_{m,n}^1 Z_3' > \mathbf{0}) \right) \\
P(\hat{\epsilon}_l=0, \hat{\mu}_0 > \hat{\mu}_1) &= \sum_{m=0}^{n_0} \sum_{n=0}^{n_1} \binom{n_0}{m} \binom{n_1}{n} P(E_{m,n}^0 Z_4 > \mathbf{0}) \\
P(\hat{\epsilon}_l=0) &= \sum_{m=0}^{n_0} \sum_{n=0}^{n_1} \binom{n_0}{m} \binom{n_1}{n} \left(P(E_{m,n}^0 Z_4 > \mathbf{0}) + P(E_{m,n}^0 Z_4 < \mathbf{0}) \right)
\end{aligned} \tag{33}$$

where $E_{m,n}^r$ is a diagonal matrix of size $2(n_0 + n_1) + r + 1$, for $r = 0, 1, 2$, with diagonal elements $(-\mathbf{1}_m, \mathbf{1}_{n_0-m}, -\mathbf{1}_m, \mathbf{1}_{n_0-m}, -\mathbf{1}_n, \mathbf{1}_{n_1-n}, -\mathbf{1}_n, \mathbf{1}_{n_1-n}, \mathbf{1}_{r+1})$. Here Z_1 is a gaussian random vector of size $2(n_0 + n_1) + 3$, with mean μ_{Z_1} given by:

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

and covariance matrix Σ_{Z_1} given by

$$\Sigma_{Z_1} = \begin{bmatrix} C^1 & C^2 & C^4 \\ C^{2T} & C^3 & C^5 \\ C^{4T} & C^{5T} & C^6 \end{bmatrix} \tag{34}$$

where

$$(C^1)_{ij} = \begin{cases} \left(4 - \frac{7}{n_0} + \frac{3}{n_0^2}\right)\sigma_0^2 + \frac{(n_0-1)^2\sigma_1^2}{n_0^2n_1}, & i, j = 1, \dots, n_0, i = j \\ \left(\frac{-3}{n_0} + \frac{2}{n_0^2}\right)\sigma_0^2 + \frac{(n_0-1)^2\sigma_1^2}{n_0^2n_1}, & i, j = 1, \dots, n_0, i \neq j \\ \frac{(n_0-1)\sigma_0^2}{n_0^2} + \frac{(n_0-1)^2\sigma_1^2}{n_0^2n_1}, & i, j = n_0 + 1, \dots, 2n_0, i = j \\ \frac{(n_0-2)\sigma_0^2}{n_0^2} + \frac{(n_0-1)^2\sigma_1^2}{n_0^2n_1}, & i, j = n_0 + 1, \dots, 2n_0, i \neq j \\ \frac{-(n_0-1)\sigma_0^2}{n_0^2} + \frac{(n_0-1)^2\sigma_1^2}{n_0^2n_1}, & \begin{cases} i = n_0 + 1, \dots, 2n_0, j = i - n_0 \\ j = n_0 + 1, \dots, 2n_0, i = j - n_0 \end{cases} \\ \frac{\sigma_0^2}{n_0} + \frac{(n_0-1)^2\sigma_1^2}{n_0^2n_1}, & \text{otherwise} \end{cases} \quad (35)$$

$$C^2 = \left[\frac{(n_1-1)(n_0-1)\sigma_0^2}{n_0^2n_1} + \frac{(n_1-1)(n_0-1)\sigma_1^2}{n_1^2n_0} \right] \mathbf{1}_{2n_0 \times 2n_1} \quad (36)$$

$$(C^3)_{ij} = \begin{cases} \left(4 - \frac{7}{n_1} + \frac{3}{n_1^2}\right)\sigma_1^2 + \frac{(n_1-1)^2\sigma_0^2}{n_0n_1^2}, & i, j = 1, \dots, n_1, i = j \\ \left(\frac{-3}{n_1} + \frac{2}{n_1^2}\right)\sigma_1^2 + \frac{(n_1-1)^2\sigma_0^2}{n_0n_1^2}, & i, j = 1, \dots, n_1, i \neq j \\ \frac{(n_1-1)\sigma_1^2}{n_1^2} + \frac{(n_1-1)^2\sigma_0^2}{n_0n_1^2}, & i, j = n_1 + 1, \dots, 2n_1, i = j \\ \frac{(n_1-2)\sigma_1^2}{n_1^2} + \frac{(n_1-1)^2\sigma_0^2}{n_0n_1^2}, & i, j = n_1 + 1, \dots, 2n_1, i \neq j \\ \frac{-(n_1-1)\sigma_1^2}{n_1^2} + \frac{(n_1-1)^2\sigma_0^2}{n_0n_1^2}, & \begin{cases} i = n_1 + 1, \dots, 2n_1, j = i - n_1 \\ j = n_1 + 1, \dots, 2n_1, i = j - n_1 \end{cases} \\ \frac{\sigma_1^2}{n_1} + \frac{(n_1-1)^2\sigma_0^2}{n_0n_1^2}, & \text{otherwise} \end{cases} \quad (37)$$

$$C^4 = \frac{(n_0-1)}{n_0} \left[\begin{pmatrix} \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} & \frac{\sigma_1^2}{n_1} - \frac{\sigma_0^2}{n_0} \end{pmatrix}_{2n_0 \times 1} \right]_{2n_0 \times 3} \quad (38)$$

$$C^5 = \frac{(n_1-1)}{n_1} \left[\begin{pmatrix} \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1} & \frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1} & \frac{\sigma_1^2}{n_1} - \frac{\sigma_0^2}{n_0} \end{pmatrix}_{2n_1 \times 1} \right]_{2n_1 \times 3} \quad (39)$$

$$C^6 = \begin{pmatrix} \left(\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\right) & \left(\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1}\right) & \left(\frac{\sigma_1^2}{n_1} - \frac{\sigma_0^2}{n_0}\right) \\ \left(\frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1}\right) & \left(\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\right) & -\left(\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\right) \\ \left(\frac{\sigma_1^2}{n_1} - \frac{\sigma_0^2}{n_0}\right) & -\left(\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\right) & \left(\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\right) \end{pmatrix}, \quad (40)$$

whereas Z'_1 is a Gaussian random vector of size $2(n_0+n_1)+3$, obtained from Z_1 by multiplying by -1 the last two components of Z_1 . Furthermore, Z_2 (resp. Z_3) is a Gaussian random vector of size $2(n_0+n_1)+2$, obtained from Z_1 by eliminating component $2(n_0+n_1)+3$ (resp. $2(n_0+n_1)+2$), while Z'_2 (resp. Z'_3) is a Gaussian random vector of size $2(n_0+n_1)+2$, obtained from Z'_1 by eliminating component $2(n_0+n_1)+3$ (resp. $2(n_0+n_1)+2$). Finally, Z_4 is Gaussian random vector of size $2(n_0+n_1)+1$, obtained from Z_1 by eliminating both components $2(n_0+n_1)+2$ and $2(n_0+n_1)+3$.

Proof. See Appendix.

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

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

$$\begin{aligned}
& P\left(\hat{\varepsilon}_l = \frac{k}{n_0+n_1}, \hat{\mu} \in (a, b), \hat{\mu}_0 > \hat{\mu}_1\right) = \\
& \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} \sum_{p=0}^l \sum_{q=0}^{k-l} \binom{l}{p} \binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m} \binom{n_1-(k-l)}{n} P(E_{m,n}^{2,p,q,k,l} Z_2 > \mathbf{0}) \\
& P\left(\hat{\varepsilon}_l = \frac{k}{n_0+n_1}, \hat{\mu} \in (a, b), \hat{\mu}_0 < \hat{\mu}_1\right) = \\
& \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} \sum_{p=0}^l \sum_{q=0}^{k-l} \binom{l}{p} \binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m} \binom{n_1-(k-l)}{n} P(E_{m,n}^{2,p,q,k,l} Z'_1 < \mathbf{0}) \\
& P\left(\hat{\varepsilon}_l = \frac{k}{n_0+n_1}, \hat{\mu} \notin (a, b), \hat{\mu}_0 < \hat{\mu}_1\right) = \\
& \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} \sum_{p=0}^l \sum_{q=0}^{k-l} \binom{l}{p} \binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m} \binom{n_1-(k-l)}{n} [P(E_{m,n}^{1,p,q,k,l} Z_2 < \mathbf{0}) \\
& \qquad \qquad \qquad + P(E_{m,n}^{1,p,q,k,l} Z_3 < \mathbf{0})] \\
& P\left(\hat{\varepsilon}_l = \frac{k}{n_0+n_1}, \hat{\mu} \notin (a, b), \hat{\mu}_0 > \hat{\mu}_1\right) = \\
& \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} \sum_{p=0}^l \sum_{q=0}^{k-l} \binom{l}{p} \binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m} \binom{n_1-(k-l)}{n} [P(E_{m,n}^{1,p,q,k,l} Z'_2 > \mathbf{0}) \\
& \qquad \qquad \qquad + P(E_{m,n}^{1,p,q,k,l} Z'_3 > \mathbf{0})] \\
& P\left(\hat{\varepsilon}_l = \frac{k}{n_0+n_1}, \hat{\mu}_0 > \hat{\mu}_1\right) = \\
& \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} \sum_{p=0}^l \sum_{q=0}^{k-l} \binom{l}{p} \binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m} \binom{n_1-(k-l)}{n} P(E_{m,n}^{0,p,q,k,l} Z_4 > \mathbf{0}) \\
& P\left(\hat{\varepsilon}_l = \frac{k}{n_0+n_1}\right) = \\
& \sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} \sum_{p=0}^l \sum_{q=0}^{k-l} \binom{l}{p} \binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m} \binom{n_1-(k-l)}{n} [P(E_{m,n}^{0,p,q,k,l} Z_4 > \mathbf{0}) \\
& \qquad \qquad \qquad + P(E_{m,n}^{0,p,q,k,l} Z_4 < \mathbf{0})]
\end{aligned} \tag{41}$$

where the vectors Z_i , $i = 1, \dots, 4$, and Z'_i , $i = 1, \dots, 3$, are defined in Theorem 6, and $E_{m,n}^{r,p,q,k,l}$ is a diagonal matrix of size $2(n_0 + n_1) + r + 1$ with diagonal elements given by the component-wise product of the vectors $(-\mathbf{1}_p, \mathbf{1}_{n_0}, -\mathbf{1}_{l-p}, \mathbf{1}_{n_0-l}, -\mathbf{1}_q, \mathbf{1}_{n_1}, -\mathbf{1}_{k-l-q}, -\mathbf{1}_{n_1-k+l}, \mathbf{1}_{r+1})$ and $(-\mathbf{1}_l, \mathbf{1}_m, -\mathbf{1}_{n_0-m}, \mathbf{1}_m, -\mathbf{1}_{n_0-l-m}, -\mathbf{1}_{k-l}, \mathbf{1}_n, -\mathbf{1}_{n_1-n}, \mathbf{1}_n, -\mathbf{1}_{n_1-k+l-n}, \mathbf{1}_{r+1})$.

Theorem 7, in conjunction with Lemma 2, with $\hat{\varepsilon}_r$ replaced by with $\hat{\varepsilon}_l$, allows the exact computation of the joint probability in (8) for the leave-one-out error estimator.

3.2.1 Joint Density

As in the resubstitution case, it is possible to apply a methodology similar to the one in the previous sections to obtain the joint density in (9) for the leave-one-out error estimator. As mentioned previously, Lemma 2 still holds for the case of leave-one-out, by replacing $\hat{\varepsilon}_r$ with $\hat{\varepsilon}_l$, whereas the following result is the counterpart of Theorem 5. The proof of this theorem is similar to the proof of Theorem 5 and is thus omitted.

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

$$\begin{aligned}
p\left(\hat{\varepsilon}_l = \frac{k}{n_0+n_1}, \hat{\mu} = a, \hat{\mu}_0 > \hat{\mu}_1\right) &= \\
\sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} \sum_{p=0}^l \sum_{q=0}^{k-l} \binom{l}{p} \binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m} \binom{n_1-(k-l)}{n} P(E_{m,n}^{0,p,q,k,l} Y > \mathbf{0}) \\
&\quad \times \varphi\left(0, \mu_0 + \mu_1 - 2a, \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\right) \\
p\left(\hat{\varepsilon}_l = \frac{k}{n_0+n_1}, \hat{\mu} = a, \hat{\mu}_0 < \hat{\mu}_1\right) &= \\
\sum_{l=0}^k \binom{n_0}{l} \binom{n_1}{k-l} \sum_{p=0}^l \sum_{q=0}^{k-l} \binom{l}{p} \binom{k-l}{q} \sum_{m=0}^{n_0-l} \sum_{n=0}^{n_1-(k-l)} \binom{n_0-l}{m} \binom{n_1-(k-l)}{n} P(E_{m,n}^{0,p,q,k,l} Y < \mathbf{0}) \\
&\quad \times \varphi\left(0, \mu_0 + \mu_1 - 2a, \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}\right)
\end{aligned} \tag{42}$$

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

$$\mu_Y = 2 \frac{n_1 \sigma_0^2 (a - \mu_1) - n_0 \sigma_1^2 (a - \mu_0)}{n_1 \sigma_0^2 + n_0 \sigma_1^2} \left[\binom{n_0-1}{n_0} \mathbf{1}_{2n_0}^T \binom{n_1-1}{n_1} \mathbf{1}_{2n_1}^T \mathbf{1} \right]_{(2n_0+2n_1+1) \times 1}^T \tag{43}$$

and covariance matrix Σ_Y given by

$$\Sigma_Y = \Sigma_{Y_{11}} - \frac{1}{n_0 n_1} \frac{(n_1 \sigma_0^2 - n_0 \sigma_1^2)^2}{n_1 \sigma_0^2 + n_0 \sigma_1^2} H_{(2n_0+2n_1+1) \times (2n_0+2n_1+1)} \tag{44}$$

where

$$\Sigma_{Y_{11}} = \begin{bmatrix} C^1 & C^2 & \text{ad} \mathbf{1}_{2n_0} \\ C^{2T} & C^3 & \text{cd} \mathbf{1}_{2n_1} \\ \text{ad} \mathbf{1}_{2n_0}^T & \text{cd} \mathbf{1}_{2n_1}^T & \text{d} \end{bmatrix} \tag{45}$$

and

$$H = \begin{bmatrix} a^2 \mathbf{1}_{2n_0 \times 2n_0} & b \mathbf{1}_{2n_0 \times 2n_1} & a \mathbf{1}_{2n_0} \\ b \mathbf{1}_{2n_1 \times 2n_0} & c^2 \mathbf{1}_{2n_1 \times 2n_1} & c \mathbf{1}_{2n_1} \\ a \mathbf{1}_{2n_0}^T & c \mathbf{1}_{2n_1}^T & 1 \end{bmatrix} \tag{46}$$

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

Theorem 8, in conjunction with Lemma 3, with $\hat{\varepsilon}_r$ replaced by with $\hat{\varepsilon}_l$, allows the exact computation of joint density in (9) for the leave-one-out error estimator.

3.2.2 Numerical Examples

Figures 5 and 6 display examples of the joint probability in (8) and the corresponding joint density in (9), respectively, for the leave-one-out error estimator, computed using the expressions given previously. Comparing these figures to Figures 3 and 4, one observes, among other interesting facts, that there is in the present case more probability mass at large values of the error estimator, as expected due to the generally larger variance of leave-one-out with respect to resubstitution.

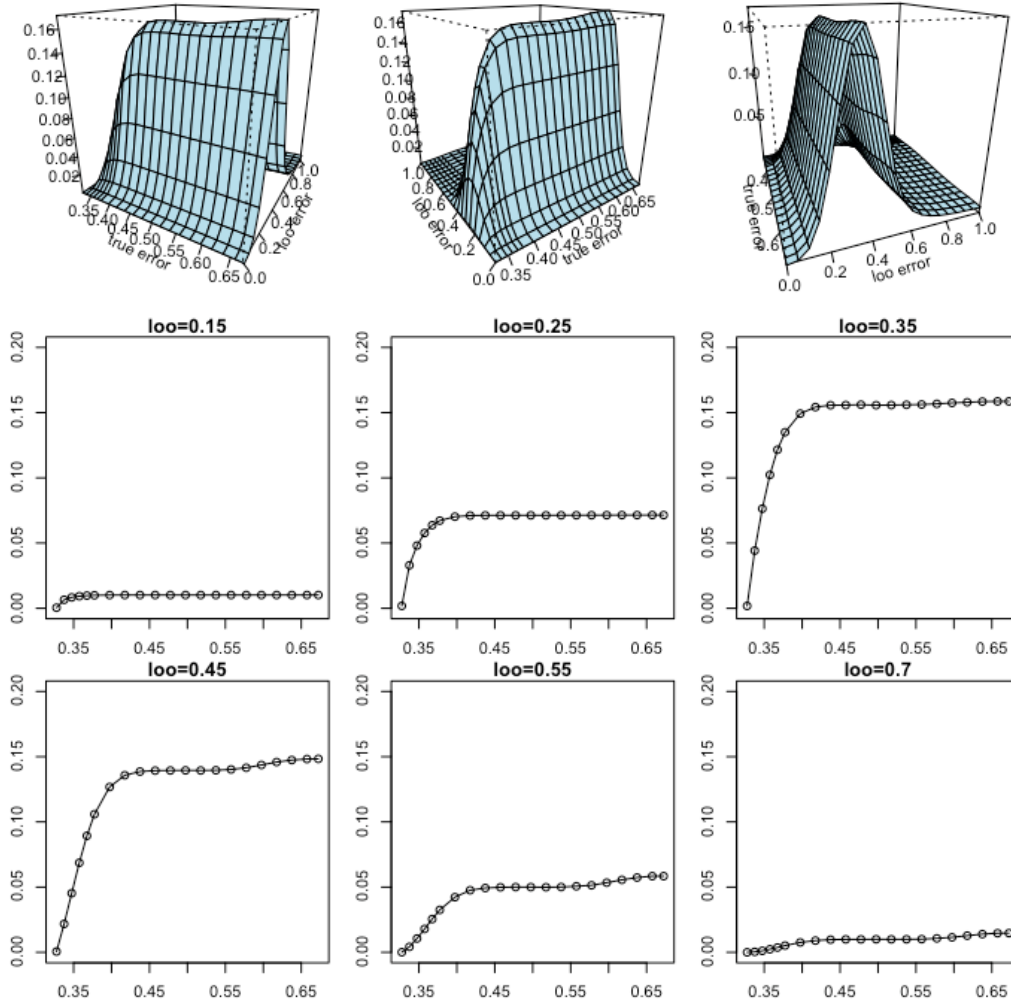


Figure 5: Joint probability in (8) for the leave-one-out error estimator: $n_0 = n_1 = 10$, $m_0 = 1, m_1 = 0$, $\sigma_0 = 2, \sigma_1 = 1$. Bayes error = 0.32742.

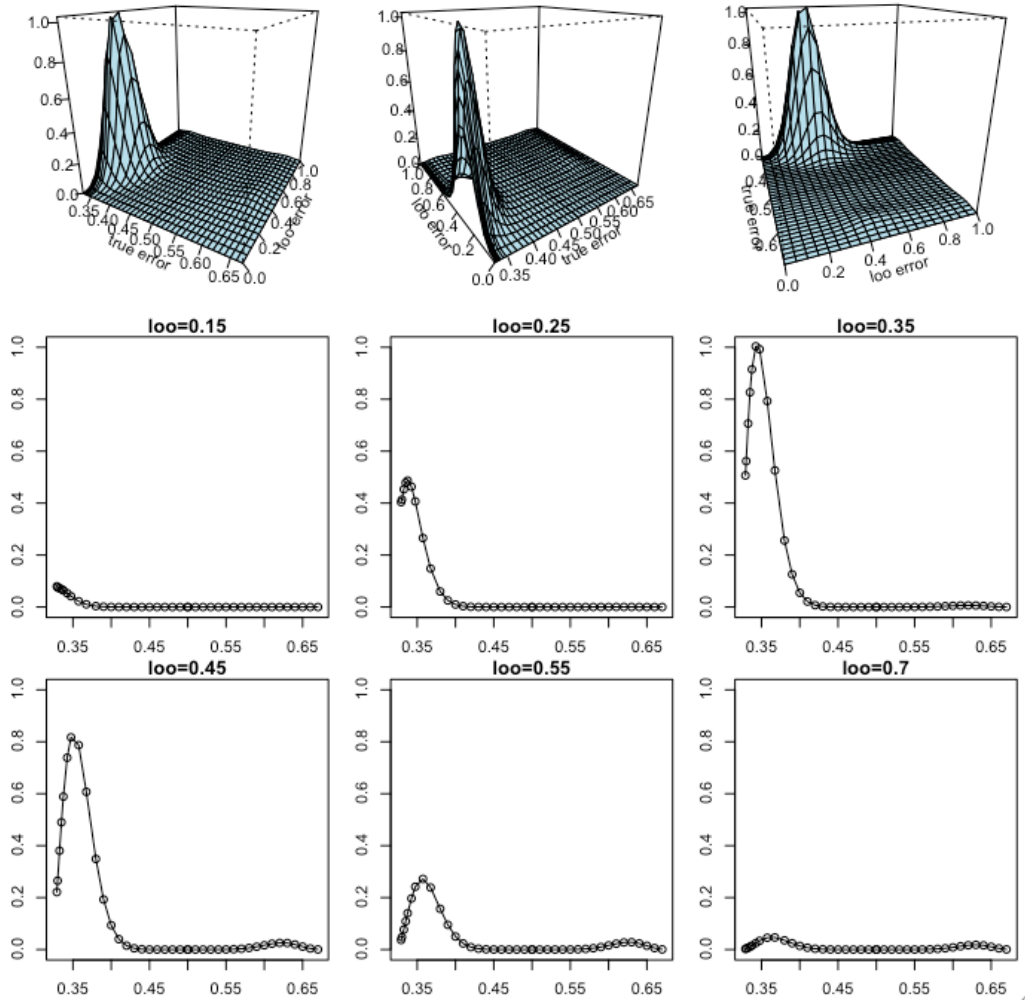


Figure 6: Joint density in (9) for the leave-one-out error estimator: $n_0 = n_1 = 10$, $m_0 = 1, m_1 = 0$, $\sigma_0 = 2, \sigma_1 = 1$. Bayes error = 0.32742.

4 Multivariate Case

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

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

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

for $i = 0, 1$, whereas

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

is the Mahalanobis distance between classes.

4.1 Numerical Examples

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

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

5 Conditional Bounds and Regression for the Actual Error Given the Estimated Error

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

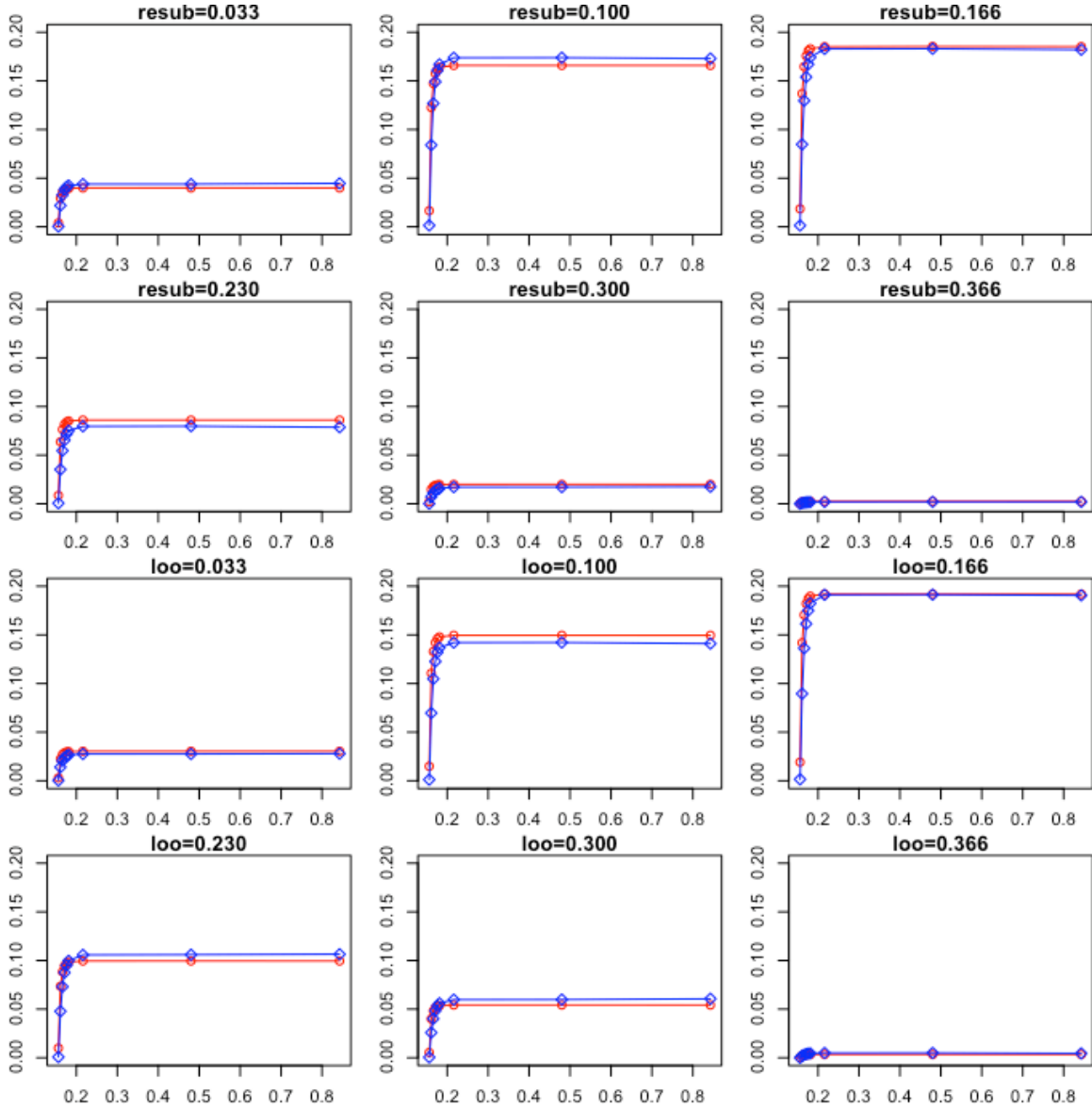


Figure 7: Joint probability in (8) for the resubstitution (top panels) and leave-one-out (bottom panels) in the multivariate case: $n_0 = n_1 = 15$, $m_0 = m_1 = -d\mathbf{1}_{p \times 1}$, $d = 0.75$, $r = 0.1$, $p = 2$. Bayes error = 0.1559. Legend key: proposed approximation (\circ), MC approximation (\diamond).

Given the knowledge of the joint probability in (8), one can write the conditional distribution of the actual error given the estimated error as

$$P\left(\varepsilon < z \mid \hat{\varepsilon} = \frac{k}{n_0 + n_1}\right) = P\left(\hat{\varepsilon} = \frac{k}{n_0 + n_1}, \varepsilon < z\right) / P\left(\hat{\varepsilon} = \frac{k}{n_0 + n_1}\right), \quad k = 0, 1, \dots, n_0 + n_1 \quad (47)$$

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

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

$$P\left(\varepsilon < z_\alpha \mid \hat{\varepsilon} = \frac{k}{n_0 + n_1}\right) = 1 - \alpha. \quad (48)$$

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

As for the regression, note that, from the conditional distribution in (47), one can obtain the conditional expectation of the actual error given the error estimator, via

$$E\left(\varepsilon \mid \hat{\varepsilon}_r = \frac{k}{n_0 + n_1}\right) = \int_0^1 \left(1 - P\left(\varepsilon < z \mid \hat{\varepsilon}_r = \frac{k}{n_0 + n_1}\right)\right) dz, \quad (49)$$

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

Figure 8 illustrates the exact 95% upper conditional bound and regression in the univariate case, using the expressions for the joint probability in (8) obtained previously, whereas Figure 9 provides similar examples in the bivariate case ($p = 2$), using the proposed approximation for the joint probability in (8) developed previously. The total number of sample points is kept to 20 to facilitate computation. In the multivariate case, we have assumed mean vectors of opposite signs $\mu_0 = m_0 = d\mathbf{1}_{p \times 1}$ and $\mu_1 = m_1 = -d\mathbf{1}_{p \times 1}$, and covariance matrix Σ with variance 1 on the diagonal and correlation r for the off-diagonal elements, where $|r| \leq 1$. In all examples, in order to avoid numerical instability issues, the conditional bounds and regression are calculated for only those values of the error estimate that have a significant probability mass, e.g. $P(\hat{\varepsilon} = \frac{k}{n_0 + n_1}) > \tau$, where τ is a small positive number. Note that the aforementioned probability mass function is displayed in the plots to show the concentration of mass of the error estimates.

Figure 10 presents univariate and bivariate examples derived from gene-expression data from a recently-published breast cancer study [47]. Discrimination is between good (class 0) vs. bad (class 1) prognosis. A subset of 30 samples was randomly selected among the total of 295 included in the aforementioned study, with $n_0 = 18$ and $n_1 = 12$ to reflect the proportion between classes observed in the full data set, and corresponding normalized gene expression measurements were extracted for the genes ‘‘LOC51203’’ and ‘‘FGF18.’’ Those are the top genes according to both the t -test and fold change. Univariate and bivariate Shapiro-Wilk tests (using the R statistical software) applied on the full data set, for more sensitivity, did not reject Gaussianity of these genes, either individually or as a pair, over either of the classes at a 95% significance level. Sample means and variances (the pooled covariance matrix was used in the bivariate case) were used as estimates of the unknown true means and variances.

These results confirm the lack of regression for small-sample error estimation observed in the simulation study in [48], as one can see in the figures that both the confidence bounds and the nonlinear regressions are virtually horizontal, except for a slight bit of upward movement at the extreme right, where there is very little error-estimator mass and therefore negligible practical significance. This means that the error estimate provides essentially no information regarding the error as in practically useless, both for predicting the actual error or bounding it with confidence in the small-sample setting for this Gaussian model. As

might be expected, the situation is worse with two features as opposed to one, but there is virtually no regression in either case. We are employing very small sample sizes in these examples (up to a total of 30 sample points), but the number of features is also very small. Consider, by contrast, the much larger numbers of features often used in practice and consider the much more complex classification rules being employed. These results provide strong analytic support for the synthetic results obtained in [48].

6 Conclusion

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

In practice the unknown parameters of the Gaussian distributions, which figure in the expressions, are not known and need to be estimated. Using the usual maximum-likelihood estimates for such parameters and plugging them into the theoretical exact expressions provides a sample-based approximation to the joint distribution, and also sample-based methods to estimate upper conditional bounds on the actual error; this approach was employed in the numerical example based on gene-expression data of Section 5. As the ML estimators are consistent and all expressions are smooth, these sample-based approximations will converge to the actual values as sample size increases without bound.

Appendix

Proof of Theorem 1.

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

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

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

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

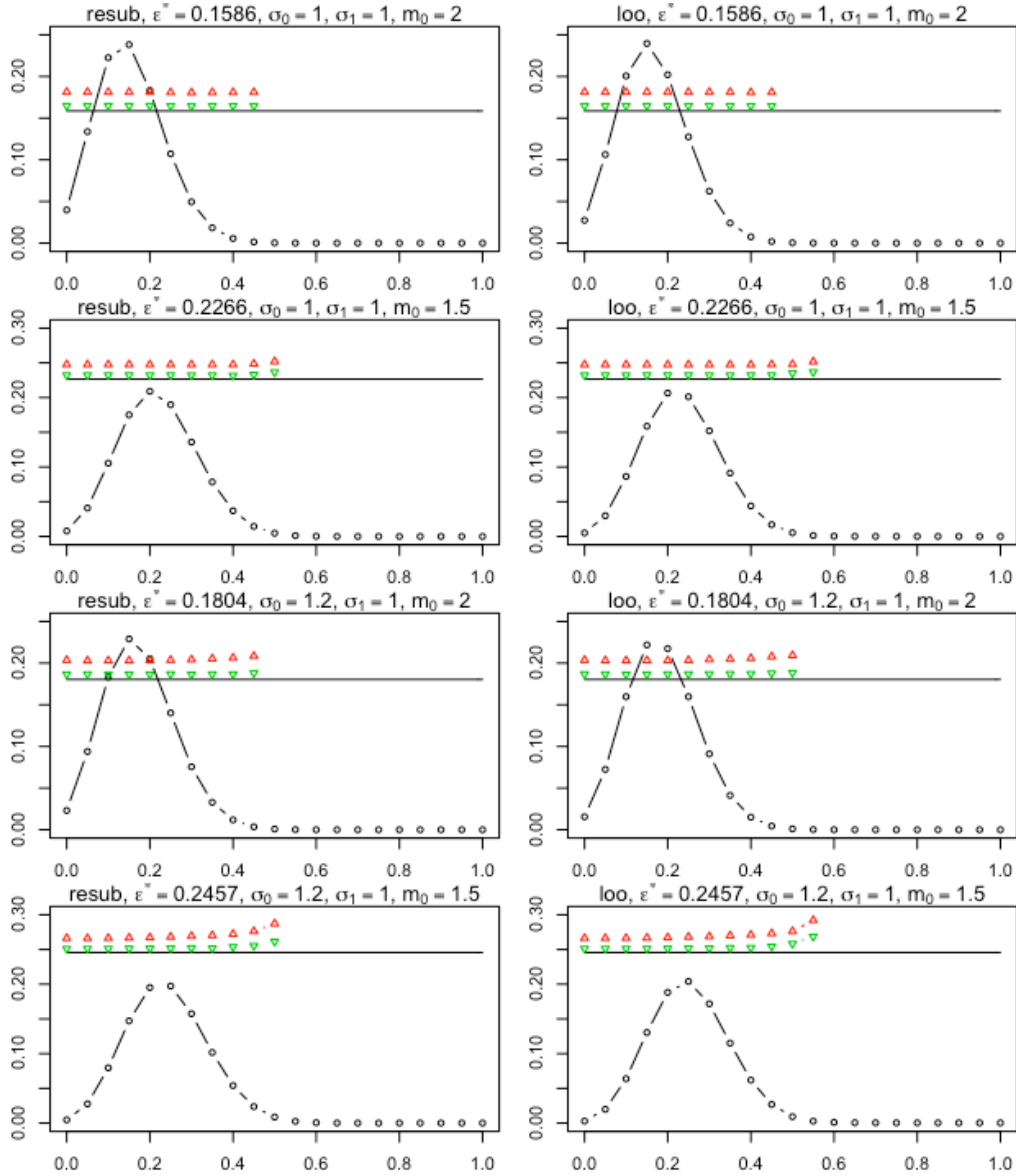


Figure 8: The 95% upper conditional bounds and regression of actual error given the resubstitution and leave-one-out error estimates in the univariate case. In all cases, $n = 20$ and $m_1 = 0$. The horizontal solid line displays the Bayes error. The marginal probability mass function for the error estimators in each case is also plotted for reference. Legend key: 95% upper conditional bound (\triangle), regression (∇), probability mass function (\circ).

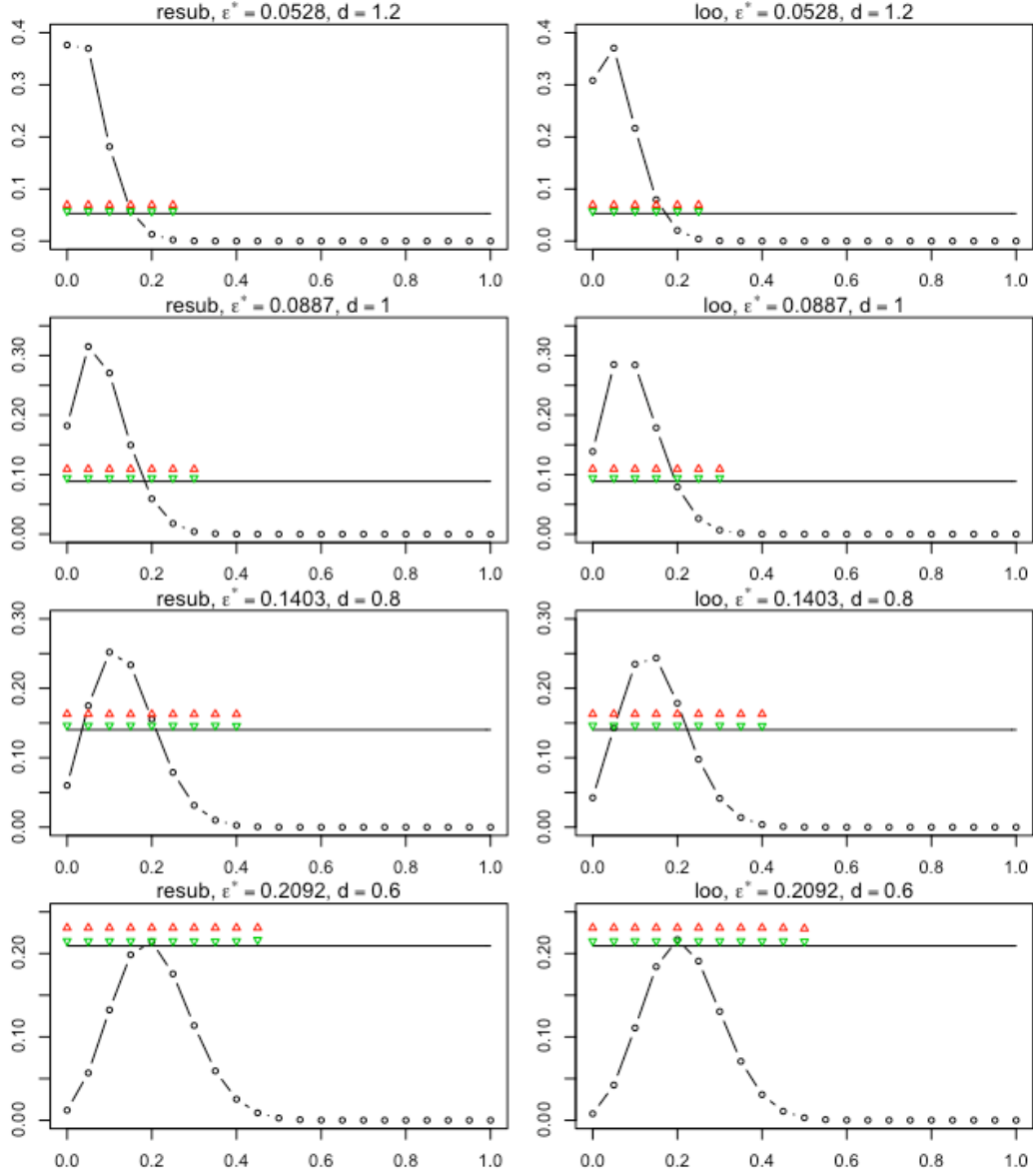


Figure 9: The 95% upper conditional bounds and regression of actual error given the resubstitution and leave-one-out error estimates in the bivariate case: $n = 20$, $m_0 = -m_1 = d\mathbf{1}_p$, $r = 0.1$, $p = 2$. The marginal probability mass function for the error estimators in each case is also plotted for reference. The horizontal solid line displays the Bayes error. Legend key: 95% upper conditional bound (Δ), regression (∇), probability mass function (\circ).

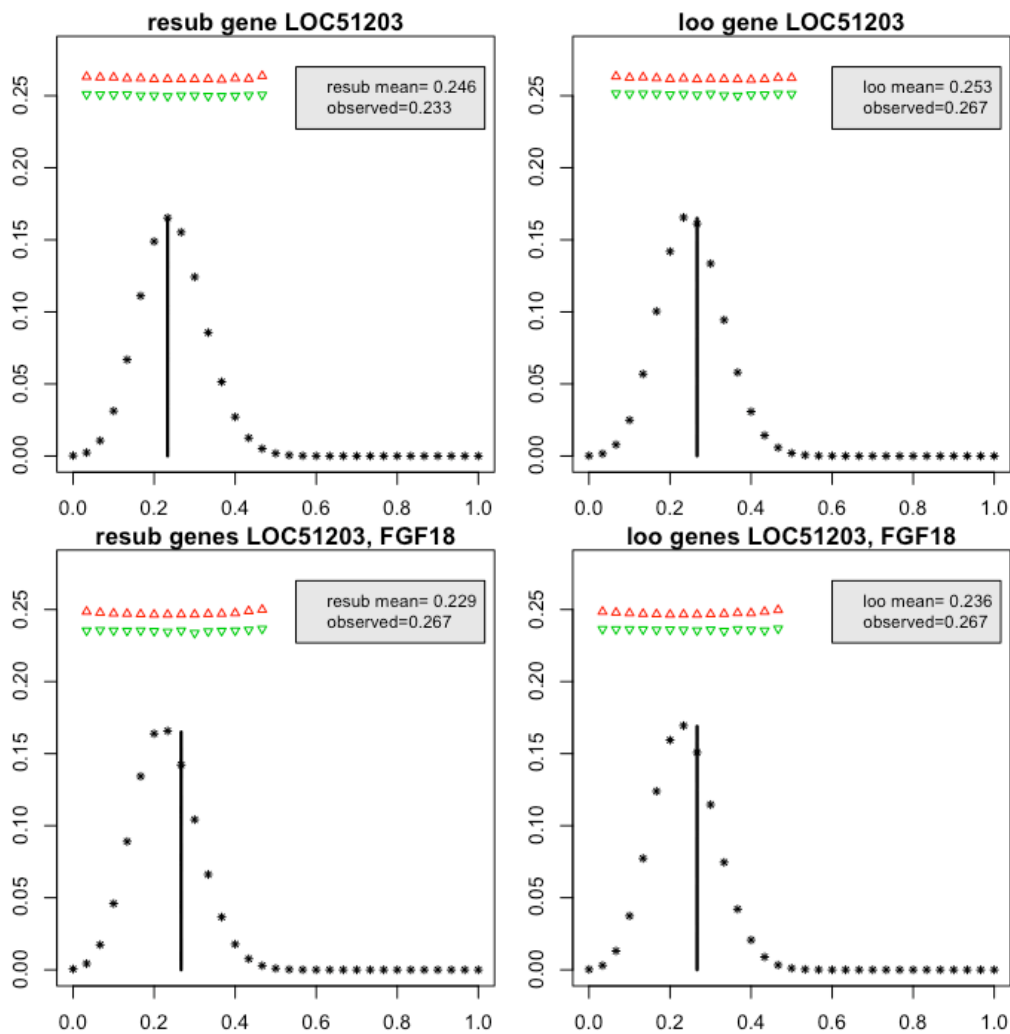


Figure 10: The 95% upper conditional bounds and regression of actual error given the resubstitution and leave-one-out error estimates in the univariate case (top row) and bivariate case (bottom row), for distributional parameters estimated from gene-expression data (see text). The marginal probability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

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

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

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

$$\mathbf{c} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 2a \\ -2b \end{pmatrix}_{(n_0+n_1+3) \times 1} \quad (50)$$

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

$$A = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix}_{(n_0+n_1+3) \times (n_0+n_1)} \quad (51)$$

where

$$A_1 = \begin{pmatrix} \left(2 - \frac{1}{n_0}\right) & -\frac{1}{n_0} & \cdots & -\frac{1}{n_0} & -\frac{1}{n_1} & \cdots & -\frac{1}{n_1} \\ -\frac{1}{n_0} & \left(2 - \frac{1}{n_0}\right) & \cdots & -\frac{1}{n_0} & -\frac{1}{n_1} & \cdots & -\frac{1}{n_1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n_0} & -\frac{1}{n_0} & \cdots & \left(2 - \frac{1}{n_0}\right) & -\frac{1}{n_1} & \cdots & -\frac{1}{n_1} \end{pmatrix}_{n_0 \times (n_0+n_1)} \quad (52)$$

$$A_2 = \begin{pmatrix} \frac{1}{n_0} & \cdots & \frac{1}{n_0} & \left(\frac{1}{n_1} - 2\right) & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \\ \frac{1}{n_0} & \cdots & \frac{1}{n_0} & \frac{1}{n_1} & \left(\frac{1}{n_1} - 2\right) & \cdots & \frac{1}{n_1} \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n_0} & \cdots & \frac{1}{n_0} & \frac{1}{n_1} & \frac{1}{n_1} & \cdots & \left(\frac{1}{n_1} - 2\right) \end{pmatrix}_{n_1 \times (n_0+n_1)} \quad (53)$$

$$A_3 = \begin{pmatrix} \frac{1}{n_0} & \cdots & \frac{1}{n_0} & -\frac{1}{n_1} & \cdots & -\frac{1}{n_1} \\ \frac{1}{n_0} & \cdots & \frac{1}{n_0} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \\ \frac{1}{n_0} & \cdots & \frac{1}{n_0} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \end{pmatrix}_{3 \times (n_0+n_1)} \quad (54)$$

Therefore, Z is a Gaussian random vector, with mean $\mu_Z = A\mu_X - \mathbf{c}$ and covariance $\Sigma_Z = A\Sigma_X A^T$. Substituting the values of $\mu_X = [\mu_0 \mathbf{1}_{n_0}, \mu_1 \mathbf{1}_{n_1}]^T$ and $\Sigma_X = \text{diag}(\mathbf{1}_{n_0}, \mathbf{1}_{n_1})$ results in (17) and (18). \square

Proof of Theorem 5.

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

$$P(\hat{\varepsilon}_r = 0, \hat{\mu} > a, \hat{\mu}_0 > \hat{\mu}_1) = P(Z > \mathbf{0}) \quad (55)$$

where

$$Z = [X_1 - \hat{\mu}, \dots, X_{n_0} - \hat{\mu}_0, \hat{\mu} - X_{n_0+1}, \dots, \hat{\mu} - X_{n_0+n_1}, \hat{\mu}_0 - \hat{\mu}_1, 2(\hat{\mu} - a)] \quad (56)$$

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

$$\mu_Z = \begin{bmatrix} (\mu_0 - \mu_1)\mathbf{1}_{n_0+n_1+1} \\ (\mu_0 + \mu_1) - 2a \end{bmatrix} \quad (57)$$

and covariance matrix Σ_Z given by

$$(\Sigma_Z)_{ij} = \begin{cases} (4n_0 - 3)\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & i, j = 1, \dots, n_0, i = j \\ -3\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & i, j = 1, \dots, n_0, i \neq j \\ \frac{\sigma_0^2}{n_0} + (4n_1 - 3)\frac{\sigma_1^2}{n_1}, & i, j = n_0 + 1, \dots, n_0 + n_1, i = j \\ \frac{\sigma_0^2}{n_0} - 3\frac{\sigma_1^2}{n_1}, & i, j = n_0 + 1, \dots, n_0 + n_1, i \neq j \\ \frac{\sigma_0^2}{n_0} - \frac{\sigma_1^2}{n_1}, & \begin{cases} i = n_0 + n_1 + 2, j = 1, \dots, n_0 + n_1 + 1 \\ j = n_0 + n_1 + 2, i = 1, \dots, n_0 + n_1 + 1 \end{cases} \\ \frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1}, & \text{otherwise} \end{cases} \quad (58)$$

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

$$\begin{aligned} p(\hat{\varepsilon}_r = 0, \hat{\mu} = a, \hat{\mu}_0 > \hat{\mu}_1) &= P(\hat{\varepsilon}_r = 0, \hat{\mu}_0 > \hat{\mu}_1 \mid \hat{\mu} = a) p(\hat{\mu} = a) \\ &= P(Y > \mathbf{0} \mid \hat{\mu} = a) p(\hat{\mu} = a) \\ &= P(Y > \mathbf{0} \mid W = 0) p(\hat{\mu} = a) \end{aligned} \quad (59)$$

Now, it is a well-known fact (e.g. see Theorem 2.5.1 in [49]) that the distribution of vector Y given W is again Gaussian, with mean $\mu_Y - \frac{\mu_W}{\sigma_W^2} \Sigma_{YW}$, and covariance matrix $\Sigma_Y - \frac{1}{\sigma_W^2} \Sigma_{YW} \Sigma_{YW}^T$. In addition, $p(\hat{\mu} = a)$ is a Gaussian density with mean $\frac{\mu_0 + \mu_1}{2}$ and variance $\frac{1}{4}(\frac{\sigma_0^2}{n_0} + \frac{\sigma_1^2}{n_1})$. The computation of $p(\hat{\varepsilon}_r = \frac{k}{n_0+n_1}, \hat{\mu} = a, \hat{\mu}_0 < \hat{\mu}_1)$ is entirely similar. \square

Proof of Theorem 6.

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

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

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

$$\begin{aligned}
& P(W^{(1)}(X_1) \geq 0, \dots, W^{(n_0)}(X_{n_0}) \geq 0, W^{(n_0+1)}(X_{n_0+1}) < 0, \dots, W^{(n_0+n_1)}(X_{n_0+n_1}) < 0, \mathbf{A}) \\
&= P(X_1 - \hat{\mu}^{(1)} \geq 0, \hat{\nu}^{(1)} \geq 0, X_2 - \hat{\mu}^{(2)} \geq 0, \hat{\nu}^{(2)} \geq 0, \dots, X_{n_0} - \hat{\mu}^{(n_0)} \geq 0, \hat{\nu}^{(n_0)} \geq 0, \\
&\quad \hat{\mu}^{(n_0+1)} - X_{n_0+1} \geq 0, \hat{\nu}^{(n_0+1)} \geq 0, \dots, \hat{\mu}^{(n_0+n_1)} - X_{n_0+n_1} \geq 0, \hat{\nu}^{(n_0+n_1)} \geq 0, \mathbf{A}) \\
&+ P(X_1 - \hat{\mu}^{(1)} < 0, \hat{\nu}^{(1)} < 0, X_2 - \hat{\mu}^{(2)} \geq 0, \hat{\nu}^{(2)} \geq 0, \dots, X_{n_0} - \hat{\mu}^{(n_0)} \geq 0, \hat{\nu}^{(n_0)} \geq 0, \\
&\quad \hat{\mu}^{(n_0+1)} - X_{n_0+1} \geq 0, \hat{\nu}^{(n_0+1)} \geq 0, \dots, \hat{\mu}^{(n_0+n_1)} - X_{n_0+n_1} \geq 0, \hat{\nu}^{(n_0+n_1)} \geq 0, \mathbf{A}) \\
&\quad \vdots \\
&+ P(X_1 - \hat{\mu}^{(1)} < 0, \hat{\nu}^{(1)} < 0, X_2 - \hat{\mu}^{(2)} < 0, \hat{\nu}^{(2)} < 0, \dots, X_{n_0} - \hat{\mu}^{(n_0)} < 0, \hat{\nu}^{(n_0)} < 0, \\
&\quad \hat{\mu}^{(n_0+1)} - X_{n_0+1} < 0, \hat{\nu}^{(n_0+1)} < 0, \dots, \hat{\mu}^{(n_0+n_1)} - X_{n_0+n_1} < 0, \hat{\nu}^{(n_0+n_1)} < 0, \mathbf{A})
\end{aligned}$$

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

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

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

$$\mathbf{c} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 2a \\ -2b \end{pmatrix}_{(2n_0+2n_1+3) \times 1} \quad (60)$$

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

$$A = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \end{pmatrix}_{(2n_0+2n_1+3) \times (n_0+n_1)} \quad (61)$$

where

$$A_1 = \begin{pmatrix} 2(1 - \frac{1}{n_0}) & -\frac{1}{n_0} & \dots & -\frac{1}{n_0} & -(\frac{1}{n_1} - \frac{1}{n_0 n_1}) & \dots & -(\frac{1}{n_1} - \frac{1}{n_0 n_1}) \\ -\frac{1}{n_0} & 2(1 - \frac{1}{n_0}) & \dots & -\frac{1}{n_0} & -(\frac{1}{n_1} - \frac{1}{n_0 n_1}) & \dots & -(\frac{1}{n_1} - \frac{1}{n_0 n_1}) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ -\frac{1}{n_0} & -\frac{1}{n_0} & \dots & 2(1 - \frac{1}{n_0}) & -(\frac{1}{n_1} - \frac{1}{n_0 n_1}) & \dots & -(\frac{1}{n_1} - \frac{1}{n_0 n_1}) \end{pmatrix}_{n_0 \times (n_0+n_1)} \quad (62)$$

$$A_2 = \begin{pmatrix} 0 & \frac{1}{n_0} & \cdots & \frac{1}{n_0} & -\left(\frac{1}{n_1} - \frac{1}{n_0 n_1}\right) & -\left(\frac{1}{n_1} - \frac{1}{n_0 n_1}\right) & \cdots & -\left(\frac{1}{n_1} - \frac{1}{n_0 n_1}\right) \\ \frac{1}{n_0} & 0 & \cdots & \frac{1}{n_0} & -\left(\frac{1}{n_1} - \frac{1}{n_0 n_1}\right) & -\left(\frac{1}{n_1} - \frac{1}{n_0 n_1}\right) & \cdots & -\left(\frac{1}{n_1} - \frac{1}{n_0 n_1}\right) \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{1}{n_0} & \cdots & \frac{1}{n_0} & 0 & -\left(\frac{1}{n_1} - \frac{1}{n_0 n_1}\right) & -\left(\frac{1}{n_1} - \frac{1}{n_0 n_1}\right) & \cdots & -\left(\frac{1}{n_1} - \frac{1}{n_0 n_1}\right) \end{pmatrix}_{n_0 \times (n_0 + n_1)} \quad (63)$$

$$A_3 = \begin{pmatrix} \left(\frac{1}{n_0} - \frac{1}{n_0 n_1}\right) & \cdots & \left(\frac{1}{n_0} - \frac{1}{n_0 n_1}\right) & -2\left(1 - \frac{1}{n_1}\right) & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \\ \left(\frac{1}{n_0} - \frac{1}{n_0 n_1}\right) & \cdots & \left(\frac{1}{n_0} - \frac{1}{n_0 n_1}\right) & \frac{1}{n_1} & -2\left(1 - \frac{1}{n_1}\right) & \cdots & \frac{1}{n_1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \left(\frac{1}{n_0} - \frac{1}{n_0 n_1}\right) & \cdots & \left(\frac{1}{n_0} - \frac{1}{n_0 n_1}\right) & \frac{1}{n_1} & \cdots & \frac{1}{n_1} & -2\left(1 - \frac{1}{n_1}\right) \end{pmatrix}_{n_1 \times (n_0 + n_1)} \quad (64)$$

$$A_4 = \begin{pmatrix} \left(\frac{1}{n_0} - \frac{1}{n_0 n_1}\right) & \cdots & \left(\frac{1}{n_0} - \frac{1}{n_0 n_1}\right) & 0 & -\frac{1}{n_1} & -\frac{1}{n_1} & \cdots & -\frac{1}{n_1} \\ \left(\frac{1}{n_0} - \frac{1}{n_0 n_1}\right) & \cdots & \left(\frac{1}{n_0} - \frac{1}{n_0 n_1}\right) & -\frac{1}{n_1} & 0 & -\frac{1}{n_1} & \cdots & -\frac{1}{n_1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \left(\frac{1}{n_0} - \frac{1}{n_0 n_1}\right) & \cdots & \left(\frac{1}{n_0} - \frac{1}{n_0 n_1}\right) & -\frac{1}{n_1} & -\frac{1}{n_1} & \cdots & -\frac{1}{n_1} & 0 \end{pmatrix}_{n_1 \times (n_0 + n_1)} \quad (65)$$

$$A_5 = \begin{pmatrix} \frac{1}{n_0} & \cdots & \frac{1}{n_0} & -\frac{1}{n_1} & \cdots & -\frac{1}{n_1} \\ \frac{1}{n_0} & \cdots & \frac{1}{n_0} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \\ \frac{1}{n_0} & \cdots & \frac{1}{n_0} & -\frac{1}{n_1} & \cdots & -\frac{1}{n_1} \\ \frac{1}{n_0} & \cdots & \frac{1}{n_0} & \frac{1}{n_1} & \cdots & \frac{1}{n_1} \end{pmatrix}_{3 \times (n_0 + n_1)} \quad (66)$$

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

References

- [1] C. H. S. Michiels, S. Koscielny, "Prediction of cancer outcome with microarrays: a multiple random validation strategy," *The Lancet*, vol. 365, pp. 488–492, 2005.
- [2] A. Dupuy and R. Simon, "Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting," *J. National Cancer Institute*, vol. 99, pp. 147–147, 2007.
- [3] U. Braga-Neto and E. Dougherty, "Is cross-validation valid for microarray classification?," *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.
- [4] U. Braga-Neto, "Fads and fallacies in the name of small-sample microarray classification," *IEEE Signal Processing Magazine*, vol. 24, no. 1, pp. 91–99, 2007.

- [5] E. Dougherty, J. Hua, and M. Bittner, "Validation of computational methods in genomics," *Current Genomics*, vol. 8, no. 1, pp. 1–19, 2007.
- [6] U. Braga-Neto, R. Hashimoto, E. Dougherty, D. Nguyen, and R. Carroll, "Is cross-validation better than resubstitution for ranking genes?," *Bioinformatics*, vol. 20, no. 2, pp. 253–258, 2004.
- [7] C. Smith, "Some examples of discrimination," *Annals of Eugenics*, vol. 18, pp. 272–282, 1947.
- [8] P. Lachenbruch and M. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, pp. 1–11, 1968.
- [9] T. Cover, "Learning in pattern recognition," in *Methodologies of Pattern Recognition* (S. Watanabe, ed.), pp. 111–132, New York, NY: Academic Press, 1969.
- [10] G. Toussaint and R. Donaldson, "Algorithms for recognizing contour-traced hand-printed characters," *IEEE Transactions on Computers*, vol. 19, pp. 541–546, 1970.
- [11] M. Stone, "Cross-validators choice and assessment of statistical predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, pp. 111–147, 1974.
- [12] T. Nagahata, M. Onda, M. Emi, H. Nagai, K. Tsumagari, and T. F. et al., "Expression profiling to predict postoperative prognosis for estrogen receptor-negative breast cancers by analysis of 25,344 genes on a cDNA microarray," *Cancer Sci.*, vol. 95, no. 3, pp. 218–225, 2004.
- [13] S. Chiaretti, X. Li, R. Gentleman, A. Vitale, M. Vignetti, and F. M. et al., "Gene expression profile of adult t-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival," *Blood*, vol. 103, no. 7, pp. 2771–2778, 2004.
- [14] F. De Smet, N. Pochet, K. Engelen, T. Van Gorp, P. Van Hummelen, K. Marchal, F. Amant, D. Timmerman, B. De Moor, and I. Vergote, "Predicting the clinical behavior of ovarian cancer from gene expression profiles," *Int. J. Gynecol. Cancer*, vol. 16, no. Suppl 1, pp. 147–151, 2006.
- [15] 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," *Oncol. Rep.*, vol. 19, no. 2, pp. 489–495, 2008.
- [16] M. Shirahata, K. Iwao-Koizumi, S. Saito, N. Ueno, M. Oda, N. Hashimoto, J. Takahashi, and K. Kato, "Gene expression-based molecular diagnostic system for malignant gliomas is superior to histological diagnosis," *Clin. Cancer Res.*, vol. 13, no. 24, pp. 7341–7356, 2007.
- [17] C. Rimkus, J. Friederichs, A. Boulesteix, J. Theisen, J. Mages, K. Becker, H. Nekarda, R. Rosenberg, K. Janssen, and J. Siewert, "Microarray-based prediction of tumor response to neoadjuvant radiochemotherapy of patients with locally advanced rectal cancer," *Clin. Gastroenterol. Hepatol.*, vol. 6, no. 1, pp. 53–61, 2008.
- [18] O. Gevaert, F. De Smet, T. V. Gorp, N. Pochet, K. Engelen, F. Amant, B. De Moor, D. Timmerman, and I. Vergote, "Expression profiling to predict the clinical behaviour of ovarian cancer fails independent evaluation," *BMC Cancer*, vol. 8, p. 18, 2008.

- [19] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer, 1996.
- [20] J. Kittler and P. DeVijver, “Statistical properties of error estimators in performance assessment of recognition systems,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 4, pp. 215–220, 1982.
- [21] D. Hand, “Recent advances in error rate estimation,” *Pattern Recognition Letters*, vol. 4, pp. 335–346, 1986.
- [22] R. Fisher, “Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population,” *Biometrika*, vol. 10, pp. 507–521, 1915.
- [23] R. Fisher, “On the ”probable error” of a coefficient of correlation deduced from a small sample,” *Metron*, vol. 1, pp. 3–32, 1921.
- [24] R. Fisher, “The general sampling distribution of the multiple correlation coefficient,” *Proc. Roy. Soc., Ser. A.*, vol. 121, pp. 654–673, 1928.
- [25] D. Hand, “Classifier technology and the illusion of progress,” *Statistical Science*, vol. 21, pp. 1–14, 2006.
- [26] S. Attoor and E. Dougherty, “Classifier performance as a function of distributional complexity,” *Pattern Recognition*, vol. 37, no. 8, pp. 1629–1640, 2004.
- [27] R. Fisher, “The use of multiple measurements in taxonomic problems,” *Ann. Eugen.*, vol. 7, pp. 179–188, 1936.
- [28] 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.
- [29] T. Anderson, “Classification by multivariate analysis,” *Psychometrika*, vol. 16, pp. 31–50, 1951.
- [30] S. John, “Errors in discrimination,” *Ann. Math. Statist.*, vol. 32, pp. 1125–1144, 1961.
- [31] R. Sitgreaves, “Some results on the distribution of the w -classification,” in *Studies in Item Analysis and Prediction* (H. Solomon, ed.), pp. 241–251, Stanford University Press, 1961.
- [32] A. Bowker, “A representation of hotelling’s t^2 and anderson’s classification statistic w in terms of simple statistics,” in *Studies in Item Analysis and Prediction* (H. Solomon, ed.), pp. 285–292, Stanford University Press, 1961.
- [33] A. Bowker and R. Sitgreaves, “An asymptotic expansion for the distribution function of the w -classification statistic,” in *Studies in Item Analysis and Prediction* (H. Solomon, ed.), pp. 292–310, Stanford University Press, 1961.
- [34] H. Harter, “On the distribution of wald’s classification statistics,” *Ann. Math. Statist.*, vol. 22, pp. 58–67, 1951.
- [35] R. Sitgreaves, “On the distribution of two random matrices used in classification procedures,” *Ann. Math. Statist.*, vol. 23, pp. 263–270, 1951.

- [36] D. Teichrow and R. Sitgreaves, "Computation of an empirical sampling distribution for the w -classification statistic," in *Studies in Item Analysis and Prediction* (H. Solomon, ed.), pp. 285–292, Stanford University Press, 1961.
- [37] M. Okamoto, "An asymptotic expansion for the distribution of the linear discriminant function," *Ann. Math. Statist.*, vol. 34, pp. 1286–1301, 1963. Correction: *Ann. Math. Statist.*, 39:1358–1359, 1968.
- [38] D. Kabe, "Some results on the distribution of two random matrices used in classification procedures," *Ann. Math. Statist.*, vol. 34, pp. 181–185, 1963.
- [39] 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.
- [40] M. Moran, "On the expectation of errors of allocation associated with a linear discriminant function," *Biometrika*, vol. 62, no. 1, pp. 141–148, 1975.
- [41] G. McLachlan, "The bias of the apparent error in discriminant analysis," *Biometrika*, vol. 63, no. 2, pp. 239–244, 1976.
- [42] S. Raudys, "Comparison of the estimates of the probability of misclassification," in *Proc. 4th Int. Conf. Pattern Recognition*, (Kyoto, Japan), pp. 280–282, 1978.
- [43] D. Foley, "Considerations of sample and feature size," *IEEE Transactions on Information Theory*, vol. IT-18, no. 5, pp. 618–626, 1972.
- [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] A. Zollanvari, U. Braga-Neto, and E. Dougherty, "On the distribution of resubstitution and cross-validation error estimators for linear classifiers," 2009. Submitted.
- [46] A. Genz and F. Bretz, "Methods for the computation of multivariate t-probabilities," *J. Stat. Comp. Simul.*, vol. 11, pp. 950–971, 2002.
- [47] M. van de Vijver, Y. He, L. van't Veer, H. Dai, A. Hart, D. Voskuil, G. Schreiber, J. Peterse, C. Roberts, M. Marton, M. Parrish, D. Astma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. Rutgers, S. Friend, and R. Bernards, "A gene-expression signature as a predictor of survival in breast cancer," *The New England Journal of Medicine*, vol. 347, pp. 1999–2009, Dec 2002.
- [48] B. Hanczar, J. Hua, and E. Dougherty, "Decorrelation of the true and estimated classifier errors in high-dimensional settings," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, 2007. Article ID 38473, 12 pages.
- [49] W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: Wiley, 2nd ed., 1984.

AMIN ZOLLANVARI received B.S. and M.S. degrees in electrical engineering from Shiraz University, Iran, in 2003 and 2006. He is currently a Ph.D. student in Genomic Signal Processing Laboratory at the Department of Electrical and Computer Engineering of Texas A&M University, College Station, TX. His current research concerns the analytical study of performance of error estimators in small-sample classification.

ULISSES 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 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 Chair and is Director of the Genomic Signal Processing Laboratory. He is also the Director of the Computational Biology Division of the Translational Genomics Research Institute in Phoenix, AZ. 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 President's Award, and served as the editor of the SPIE/IS&T Journal of Electronic Imaging. At Texas A&M he has received the Association of Former Students Distinguished Achievement Award in Research, been named Fellow of the Texas Engineering Experiment Station, and named Halliburton Professor of the Dwight Look College of Engineering. Prof. Dougherty is author of 14 books, editor of five others, and author of more than 200 journal papers. He has contributed extensively to the statistical design of nonlinear operators for image processing and the consequent application of pattern recognition theory to nonlinear image processing. His research in genomic signal processing is aimed at diagnosis and prognosis based on genetic signatures and using gene regulatory networks to develop therapies based on the disruption or mitigation of aberrant gene function contributing to the pathology of a disease.