Estimation of Classifier Performance

KEINOSUKE FUKUNAGA, FELLOW, IEEE, AND RAYMOND R. HAYES, MEMBER, IEEE

Abstract—**An expression for expected classifier performance is derived and applied to a series of test procedures. For the holdout method, the roles of the independent design and test sets are identified. For the resubstitution and leave-one-out methods, the relationship between dependent design and test sets is investigated. The effect of outlier design samples is studied as a special case of the leave-one-out method. Also, the statistical properties of the bootstrap resampling technique are analyzed. The theoretical conclusions were experimentally verified using artificial data under various design conditions.**

Index Terms—**Bootstrap, classifier performance, holdout method, leave-one-out method, quadratic classifier, resubstitution method.**

I. INTRODUCTION

ESTIMATION of the expected performance of a classifier is an important, yet difficult problem in pattern recognition. In practice, the true distributions are never known and only a finite number of training samples are available. The designer must decide whether this sample size is adequate or not, and also decide how many samples should be used to design the classifier and how many should be used to test it.

A number of testing procedures have been proposed and are widely used. In the holdout method, a number of the original samples are withheld from the design process. This provides an independent test set, but drastically reduces the size of the design set. In the resubstitution method, the classifier is tested on the original design samples. This maintains the size of the design set, but ignores the independence issue generating a dangerously optimistic performance estimate [1]. The leave-one-out method [2] is designed to alleviate these difficulties. It avoids drastically dividing the available sample set into design and test, while maintaining an independence between them. Thus, the procedure utilizes all available samples more efficiently, and produces a conservative error estimate. By using these last two methods simultaneously, we can obtain upper and lower bounds of the true performance of the classifier.

More recently, Efron [3] proposed a resampling procedure, called the bootstrap method, in which artificial samples are generated from the existing samples, and the optimistic bias between the resubstitution error and the classifier error when tested on independent samples is estimated from them.

The analysis of these techniques has been a popular pattern recognition research topic. In [4], Novak presents a method of computing the error of a specific classifier, given the parameters of the test distribution. Raudys and Pikelis [5] give an excellent review of work done in approximating the expected performance in the parametric case and provide explicit expressions for several empirically designed classifiers. However, this work does not address the interaction between the design and test sets or consider testing procedures other than the holdout method. Toussaint [6] catalogs these and other testing methods and gives an overview of some of the early associated work. More recent work is surveyed in Hand [7].

Pattern recognition research has considered various questions concerning the relationship between the limited size of the training set, the number of features, and the estimation of performance criteria. While a number of these works present approximate expressions for the probability of misclassification and guidelines for selecting the size of the design sample set [5], [8]–[11], none of them present general expressions relating these relationships for a family of classifiers. In addition, only a few [1], [2], [6], [11] address the interaction between the design and test sets.

In [12], Fukunaga and Hayes investigated the effect of sample size on a family of functions, and found a manageable expression for the errors of classifiers, including the quadratic and Fisher linear classifiers. Using the expression, they computed the degradation of classifier performance due to a finite design set.

The objective of this paper is to apply the error expression of [12] to the various methods of error estimation mentioned above, and to offer a unified and comprehensive approach to the analysis of classifier performance. In Section II, after the error expression is introduced, it is applied to three cases: 1) a given classifier and a finite test set, 2) given test distributions and a finite design set, and 3) finite and independent design and test sets. For all cases, the expected values and variances of the classifier errors are presented. Although the study of Case 1 does not produce any new results, it is important to confirm that the proposed approach produces the known results, and also to show how these results are modified when the design set becomes finite, as in Cases 2 and 3. In Section III, the error expression of [12] is used to compute the bias between the leave-one-out and resubstitution errors

Manuscript received April 3, 1988; revised November 28, 1988. Recommended for acceptance by A. K. Jain. This work was supported in part by the National Science Foundation under Grants ESC-8513720 and ECS-8720655 and by IBM under the Resident Study Program.

K. Fukunaga is with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

R. R. Hayes was with the School of Electrical Engineering, Purdue University, West Lafayette, IN 47907. He is now with IBM Palo Alto Scientific Center, 1530 Page Mill Road, Palo Alto, CA 94304.

IEEE Log Number 8929955.

for quadratic classifiers. Note that in this case the design and test sample sets are no longer independent. Again, the expected value and variance of the bias are presented. Also, because of its similarity to the analysis of the leave-one-out method, the effect of outliers in design samples on the classification error is discussed. Finally, in Section IV, the theoretical analysis of the bootstrap method is presented for quadratic classifiers. The explicit error expression can be obtained for the optimistic bias of the bootstrap resubstitution error. The expected value of the bias with respect to the bootstrap procedure is shown to be very close to the bias between the conventional leave-one-out and resubstitution errors. The variance of the bootstrap bias also can be computed in a closed form.

Throughout all sections, the theoretical conclusions are experimentally verified. The results of these analyses allow us to delve into the theoretical differences between the methods and account for a series of frequency observed experimental trends.

II. Classification Errors for Finite Samples

In this section, we will discuss the effects of finite test and design samples on classification performance. John [11] provides a similar discussion for the linear classifier. Previous extensions to this work are presented in Raudys and Pikelis [5].

A. Error Expression

For the two-class problem, a classifier can be expressed by

$$h(X) \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 0 \tag{1}$$

where $h(\boldsymbol{X})$ is the discriminant function of an n-dimensional vector $\boldsymbol{X}$, and ω_i indicates the class $i (i = 1, 2)$. The probabilities of errors for this classifier from ω_1 and ω_2 are

$$\begin{aligned}\epsilon_1 &= \int_{h(X)>0} p_1(X)\,dX = \int_S u(h(X))\,p_1(X)\,dX \\ &= \frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}\left[\frac{1}{j\omega}+\pi\delta(\omega)\right]e^{j\omega h(X)}p_1(X)\,d\omega\,dX \\ &= \frac{1}{2}+\frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}\frac{e^{j\omega h(X)}}{j\omega}p_1(X)\,d\omega\,dX \end{aligned}\tag{2}$$

and

$$\begin{aligned}\epsilon_2 &= \int_{h(X)<0} p_2(X)\,dX \\ &= \frac{1}{2}-\frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}\frac{e^{j\omega h(X)}}{j\omega}p_2(X)\,d\omega\,dX \end{aligned}\tag{3}$$

where $p_i(\boldsymbol{X})$ is the density function of class i tested by the classifier, $u(h(\boldsymbol{X})) = 1$ when $h(\boldsymbol{X}) > 0$ and 0 otherwise, and S indicates the entire n-dimensional space. The second line of (2) is obtained using the fact that the Fourier transform of the step function, $u(h(\boldsymbol{X}))$, is $[1/j\omega + \pi\delta(\omega)]$.

The total probability of error is

$$\begin{aligned}\epsilon &= P_1\epsilon_1 + P_2\epsilon_2 \\ &= \frac{1}{2}+\frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}\frac{e^{j\omega h(X)}}{j\omega}\tilde{p}(X)\,d\omega\,dX \end{aligned}\tag{4}$$

where P_i is the *a priori* probability of ω_i and

$$\tilde{p}(X) = P_1p_1(X) - P_2p_2(X). \tag{5}$$

B. Effect of Test Samples

When a finite number of samples are tested by a given classifier, $p_i(\boldsymbol{X})$ of (5) may be replaced by

$$\hat{p}_i(X) = \frac{1}{N_i}\sum_{j=1}^{N_i}\delta(X - X_j^{(i)}) \tag{6}$$

where $\delta(y) = 1$ when $y = 0$ and 0 otherwise, and $\boldsymbol{X}_1^{(i)}, \cdots, \boldsymbol{X}_{N_i}^{(i)}$ are N_i test samples drawn from $p_i(\boldsymbol{X})$. Throughout the paper, boldface indicates randomness.

Thus, the estimate of the error probability is

$$\begin{aligned}\hat{\epsilon} &= \frac{1}{2}+\frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}\frac{e^{j\omega h(X)}}{j\omega}\left[\frac{P_1}{N_1}\sum_{j=1}^{N_1}\delta(X - X_j^{(1)})\right. \\ &\quad \left. -\frac{P_2}{N_2}\sum_{j=1}^{N_2}\delta(X - X_j^{(2)})\right]d\omega\,dX \\ &= \frac{1}{2}+\frac{P_1}{N_1}\sum_{j=1}^{N_1}\alpha_j^{(1)} - \frac{P_2}{N_2}\sum_{j=1}^{N_2}\alpha_j^{(2)} \end{aligned}\tag{7}$$

where

$$\alpha_j^{(i)} = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{e^{j\omega h(X_j^{(i)})}}{j\omega}\,d\omega. \tag{8}$$

The expected value of $\alpha_j^{(i)}$ with respect to $\boldsymbol{X}_j^{(i)}$ (w.r.t. the test samples) is

$$\begin{aligned}\bar{\alpha}_i = E_t\{\alpha_j^{(i)}\} &= \frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}\frac{e^{j\omega h(X)}}{j\omega}p_i(X)\,d\omega\,dX \\ &= \begin{cases}\epsilon_1 - \dfrac{1}{2} & \text{for } i = 1 \\ \dfrac{1}{2} - \epsilon_2 & \text{for } i = 2.\end{cases}\end{aligned}\tag{9}$$

The second line of (9) can be obtained from (2) and (3), respectively. The second-order moments are also computed as

$$\begin{aligned}E_t\{\alpha_j^{(i)2}\} &= E_t\left\{\left[\frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{e^{j\omega h(X)}}{j\omega}\,d\omega\right]^2\right\} \\ &= E_t\left\{\left[\frac{1}{2}\operatorname{sgn}(h(X))\right]^2\right\} \\ &= \frac{1}{4}\end{aligned}\tag{10}$$

$$E_t\{\alpha_j^{(i)}\alpha_k^{(l)}\} = \bar{\alpha}_i\bar{\alpha}_l \qquad k \neq j \tag{11}$$

where sgn (h) equals $+1$ for $h > 0$ and -1 for $h < 0$. Equation (11) is obtained because $\alpha_j^{(i)}$ and $\alpha_k^{(l)}$ are independent due to the independence between $X_j^{(i)}$ and $X_k^{(l)}$.

From (7) and (9)–(11),

$$E_t\{\hat{\epsilon}\} = \frac{1}{2} + P_1\bar{\alpha}_1 - P_2\bar{\alpha}_2$$
$$= \frac{1}{2} + P_1\left(\epsilon_1 - \frac{1}{2}\right) - P_2\left(\frac{1}{2} - \epsilon_2\right) = \epsilon \tag{12}$$

$$\mathrm{Var}_t\{\hat{\epsilon}\} = \frac{P_1^2}{N_1}\mathrm{Var}_t\{\alpha_j^{(1)}\} + \frac{P_2^2}{N_2}\mathrm{Var}_t\{\alpha_j^{(2)}\}$$
$$= \frac{P_1^2}{N_1}\left[\frac{1}{4} - \left(\epsilon_1 - \frac{1}{2}\right)^2\right]$$
$$+ \frac{P_2^2}{N_2}\left[\frac{1}{4} - \left(\frac{1}{2} - \epsilon_2\right)^2\right]$$
$$= P_1^2\frac{\epsilon_1(1-\epsilon_1)}{N_1} + P_2^2\frac{\epsilon_2(1-\epsilon_2)}{N_2}. \tag{13}$$

That is, $\hat{\epsilon}$ is an unbiased estimate, and its variance has the well-known form derived from the binomial distribution [13].

C. Effect of Design Samples

It is more difficult to discuss the effect of using a finite number of design samples. Although we would like to keep the formula as general as possible, in this section a specific family of discriminant functions is investigated to help determine which approximations should be used.

Assume that the discriminant function is a function of two expected vectors, M_1 and M_2, and covariance matrices, Σ_1 and Σ_2. Typical examples are the quadratic classifier and Fisher's linear classifier:

$$h(X) = \frac{1}{2}(X - M_1)^T\Sigma_1^{-1}(X - M_1)$$
$$- \frac{1}{2}(X - M_2)^T\Sigma_2^{-1}(X - M_2)$$
$$+ \frac{1}{2}\ln\frac{|\Sigma_1|}{|\Sigma_2|} \tag{14}$$

$$h(X) = (M_2 - M_1)^T\bar{\Sigma}^{-1}X$$
$$+ \frac{1}{2}(M_1^T\bar{\Sigma}^{-1}M_1 - M_2^T\bar{\Sigma}^{-1}M_2) \tag{15}$$

where $\Sigma = [\Sigma_1 + \Sigma_2]/2$. When only a finite number of design samples are available and M_i and Σ_i are estimated from them,

$$\Delta h(X) = \hat{h}(X) - h(X) = \sum_{k=1}^{\infty}\mathbf{0}^{(k)} \tag{16}$$

where $\hat{h}(X) = h(X, \hat{M}_1, \hat{M}_2, \hat{\Sigma}_1, \hat{\Sigma}_2)$, $h(X) = h(X, M_1, M_2, \Sigma_1, \Sigma_2)$ and $\mathbf{0}^{(k)}$ is the kth order term of the Taylor series expansion in terms of the variations of $\hat{M}_i$ and $\hat{\Sigma}_i$. If the design samples are drawn from Gaussian distributions, and $\hat{M}_i$ and $\hat{\Sigma}_i$ are unbiased estimates (e.g., the sample mean and sample covariance), it is known [12] that

$$E_d\{\mathbf{0}^{(1)}\} = 0, \quad E_d\{\mathbf{0}^{(2)}\} \sim 1/\mathfrak{N},$$
$$E_d\{\mathbf{0}^{(3)}\} = 0, \quad E_d\{\mathbf{0}^{(4)}\} \sim 1/\mathfrak{N}^2 \cdots \tag{17}$$

where E_d indicates the expectation with respect to the design samples, and $\mathfrak{N}$ is the number of design samples (while N indicates the number of test samples). Therefore, from (16) and (17),

$$E_d\{\Delta h(X)\} \sim 1/\mathfrak{N}, \quad E_d\{\Delta h^2(X)\} \sim 1/\mathfrak{N},$$
$$E_d\{\Delta h^3(X)\} \sim 1/\mathfrak{N}^2, \quad E_d\{\Delta h^4(X)\} \sim 1/\mathfrak{N}^2 \cdots. \tag{18}$$

Assuming that $\mathfrak{N}$ is reasonably large, we can eliminate $E\{\Delta h^m(X)\}$ for m larger than 2.

Thus, the error of a random classifier for given test distributions is expressed by (4)

$$\hat{\epsilon} = \frac{1}{2} + \frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}\frac{e^{j\omega\hat{h}(X)}}{j\omega}\tilde{p}(X)\,d\omega\,dX \tag{19}$$

The expected value $\bar{\epsilon}$ with respect to the design samples is

$$\bar{\epsilon} = E_d\{\hat{\epsilon}\} = \frac{1}{2} + \frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}\frac{E_d\{e^{j\omega\hat{h}(X)}\}}{j\omega}$$
$$\cdot\,\tilde{p}(X)\,d\omega\,dX = \epsilon + \overline{\Delta\epsilon} \tag{20}$$

where

$$\overline{\Delta\epsilon} \cong \frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}E_d\left\{\Delta h(X) + \frac{j\omega}{2}\Delta h^2(X)\right\}$$
$$\cdot\,e^{j\omega h(X)}\tilde{p}(X)\,d\omega\,dX. \tag{21}$$

The approximation from (20) to (21) was made by using

$$e^{j\omega\hat{h}(X)} = e^{j\omega h(X)}e^{j\omega\Delta h(X)}$$
$$\cong e^{j\omega h(X)}\left[1 + j\omega\Delta h(X) + \frac{(j\omega)^2}{2}\Delta h^2(X)\right].$$

When $h(X)$ is the Bayes classifier, ϵ must be a minimum. Appendix 1 gives the proof that $\hat{\epsilon}$ of (19) is indeed larger than ϵ of (4).

When two Gaussian distributions are classified by the quadratic or linear classifier whose parameters are estimated from a finite sample set, $\overline{\Delta\epsilon}$ of (21) can be computed. Explicit solutions for the case with $M_1 = 0$, $M_2 = M$ and $\Sigma_1 = \Sigma_2 = I$ are given in [12].

The variance of $\hat{\epsilon}$ may be computed from (19) and (20)

as

$$\begin{aligned}\mathrm{Var}_d\{\hat{\varepsilon}\} &= \frac{1}{4\pi^2}\int_{S_x}\int_{S_y}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{E_d\{e^{j\omega_1\hat{h}(X)}e^{j\omega_2\hat{h}(Y)}\}}{j\omega_1 j\omega_2}\\ &\quad\cdot \tilde{p}(X)\,\tilde{p}(Y)\;d\omega_1 d\omega_2\,dX\,dY-\left(\bar{\epsilon}-\frac{1}{2}\right)^2\\ &\cong \frac{1}{4\pi^2}\int_{S_x}\int_{S_y}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}E_d\{\Delta h(X)\,\Delta h(Y)\}\\ &\quad\cdot e^{j\omega_1 h(X)}e^{j\omega_2 h(Y)}\tilde{p}(X)\,\tilde{p}(Y)\,d\omega_1 d\omega_2\,dX\,dY\\ &= \int_{S_x}\int_{S_y}E_d\{\Delta h(X)\,\Delta h(Y)\}\,\delta(h(X))\\ &\quad\cdot\delta(h(Y))\,\tilde{p}(X)\,\tilde{p}(Y)\,dX\,dY\\ &= \int_{h(X)=0}\int_{h(Y)=0}E_d\{\Delta h(X)\,\Delta h(Y)\}\\ &\quad\cdot\tilde{p}(X)\,\tilde{p}(Y)\,dX\,dY \qquad (22)\end{aligned}$$

where the derivation from the first line to the second line is given in Appendix 2. Equation (22) indicates that the integration is carried out along the classification boundary where $h(X) = 0$. When $h(X)$ is the Bayes classifier, $\tilde{p}(X)$ of (5) must be zero at the boundary. Thus, (22) becomes 0. Since we neglected the higher order terms of $\Delta h(X)$, $\mathrm{Var}_d\{\hat{\varepsilon}\}$ is not zero, but proportional to $1/\mathfrak{N}^2$. When $h(X)$ is not the Bayes classifier, $\tilde{p}(X) \neq 0$ at $h(X) = 0$. Thus, we may observe a variance dominated by a term proportional to $1/\mathfrak{N}$ due to the fact that $E_d\{\Delta h(X)\,\Delta h(Y)\} \sim 1/\mathfrak{N}$.

In order to confirm the above theoretical conclusion, an experiment has been run for the quadratic classifier between two Gaussian distributions which share the same covariance matrix I and differ in the means to give a Bayes error of 10 percent. The dimensionality n was varied from 4 to 64 in powers of 2 and the ratio of the sample size and the dimensionality k ($= \mathfrak{N}/n$) was varied from 3 to 50. $\mathfrak{N}$ ($= nk$) samples were generated from each class according to the given mean and covariance, and $\hat{M}_i$ and $\hat{\Sigma}_i$ were estimated from the generated data using the sample mean and sample covariance. The quadratic classifier was designed by (14). Testing was done by Novak's program which numerically computes the error of any discriminant function with a quadratic form tested on separately specified Gaussian distributions [4]. This procedure was repeated 10 times. The second and third lines of Table I show the average and standard deviation of these experiments. The first line shows the theoretically computed errors from (20) and (21) [12]. Also, Fig. 1 shows the relationship between $1/k$ ($= n/\mathfrak{N}$) and the standard deviation. From these results, we may confirm that the standard deviation is very small and roughly proportional to $1/\mathfrak{N}$. Thus, the variance is proportional to $1/\mathfrak{N}^2$.

An intuitive reason why the standard deviation due to a finite number of design samples is proportional to $1/\mathfrak{N}$ may be observed as follows. When the Bayes classifier is implemented, $\Delta\varepsilon$ is always positive and thus generates a positive bias. As (21) suggests, the bias is proportional to $1/\mathfrak{N}$. Since $\Delta\varepsilon$ varies between 0 and some positive value with an expected value $a/\mathfrak{N}$ (where a is a positive number), we can expect that the standard deviation is also proportional to $1/\mathfrak{N}$.

In addition, it should be noted that design samples affect the variance of the error in a different way from test samples. When a classifier is fixed, the variations of the two test distributions are independent. Thus, $\mathrm{Var}_t\{\hat{\varepsilon}\} = P_1^2\,\mathrm{Var}\{\hat{\varepsilon}_1\} + P_2^2\,\mathrm{Var}\{\hat{\varepsilon}_2\}$ as is seen in (13). On the other hand, when the test distributions are fixed and the classifier varies, $\hat{\varepsilon}_1$ and $\hat{\varepsilon}_2$ are strongly correlated with a correlation coefficient close to -1. That is, when $\hat{\varepsilon}_1$ increases, $\hat{\varepsilon}_2$ decreases and vice versa. Thus, when P_1 =

TABLE I
QUADRATIC CLASSIFIER DEGRADATION FOR I-I (%)

k	n = 4	8	16	32	64
	14.50	16.89	21.15	30.67	48.94
3	16.68	20.41	22.04	26.73	31.31
	3.51	2.35	2.89	1.95	1.33
	12.70	14.14	16.91	22.40	33.36
5	14.03	16.40	17.34	20.81	25.54
	2.11	1.86	0.91	0.57	0.74
	11.35	12.07	13.45	16.20	21.68
10	11.52	12.40	13.66	15.73	19.34
	0.81	0.61	0.70	0.54	0.85
	10.90	11.38	12.30	14.13	17.79
15	10.86	11.84	12.32	14.15	16.58
	0.44	0.61	0.42	0.53	0.42
	10.67	11.03	11.73	13.10	15.84
20	10.77	11.05	11.90	13.93	15.13
	0.21	0.23	0.51	0.22	0.32
	10.45	10.69	11.15	12.07	13.89
30	10.54	10.71	11.14	13.07	13.65
	0.19	0.21	0.20	0.19	0.22
	10.34	10.52	10.86	11.55	12.92
40	10.37	10.57	10.87	11.50	12.75
	0.24	0.13	0.13	0.13	0.18
	10.27	10.41	10.69	11.24	12.34
50	10.25	10.44	10.68	11.25	12.21
	0.13	0.10	0.13	0.09	0.07

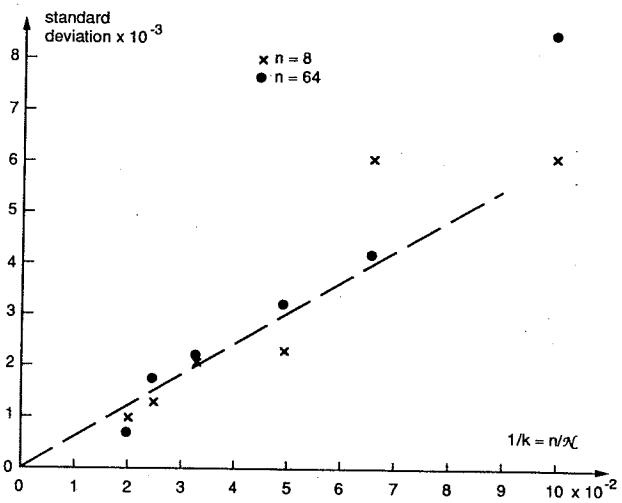


Fig. 1. Quadratic classifier degradation for I-I (standard deviation versus $n/\mathfrak{N}$).

P_2, $\text{Var}_d\{\hat{\varepsilon}\} = (0.5)^2 E_d\{\Delta\varepsilon_1^2\} + (0.5)^2 E_d\{\Delta\varepsilon_2^2\} + 2(0.5)^2 E_d\{\Delta\varepsilon_1\Delta\varepsilon_2\} \cong (0.5)^2 [E_d\{\Delta\varepsilon_1^2\} + E_d\{(-\Delta\varepsilon)^2\} + 2E_d\{\Delta\varepsilon_1(-\Delta\varepsilon_1)\}] = 0$. The covariance of $\hat{\varepsilon}_1$ and $\hat{\varepsilon}_2$ cancels the individual variances of $\hat{\varepsilon}_1$ and $\hat{\varepsilon}_2$.

D. Effect of Independent Design and Test Samples

When both design and test samples sizes are finite, the error is expressed as

$$\hat{\varepsilon} = \frac{1}{2} + \frac{P_1}{N_1}\sum_{j=1}^{N_1}\hat{\alpha}_j^{(1)} - \frac{P_2}{N_2}\sum_{j=1}^{N_2}\hat{\alpha}_j^{(2)} \tag{23}$$

where

$$\hat{\alpha}_j^{(i)} = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{e^{j\omega\hat{h}(X_j^{(i)})}}{j\omega}\,d\omega. \tag{24}$$

That is, the randomness comes from $\hat{h}$ due to the finite design samples as well as from the test samples $X_j^{(i)}$.

The expected value and variance of $\hat{\varepsilon}$ can be computed as follows:

$$\bar{\epsilon} = E\{\hat{\varepsilon}\} = E_t E_d\{\hat{\varepsilon}\} = \tfrac{1}{2} + P_1\bar{\alpha}_1 - P_2\bar{\alpha}_2 \tag{25}$$

where

$$\bar{\alpha}_i = \frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}\frac{E_d\{e^{j\omega\hat{h}(X)}\}}{j\omega}p_i(X)\,d\omega\,dX = \begin{cases}\bar{\epsilon}_1 - \dfrac{1}{2} & \text{for } i = 1\\[2ex] \dfrac{1}{2} - \bar{\epsilon}_2 & \text{for } i = 2.\end{cases} \tag{26}$$

Substituting (26) into (25),

$$\bar{\epsilon} = P_1\bar{\epsilon}_1 + P_2\bar{\epsilon}_2. \tag{27}$$

This average error is the same as the error of (20). That is, the performance degradation due to finite design and test samples is identical to the degradation due to finite design samples alone. Finite test samples do not contribute.

The variance of $\hat{\varepsilon}$ can be obtained from (23) as

$$\begin{aligned}\text{Var}\{\hat{\varepsilon}\} = {} & P_1^2\left[\frac{1}{N_1}\text{Var}\{\hat{\alpha}_j^{(1)}\} + \left(1 - \frac{1}{N_1}\right)\text{Cov}\{\hat{\alpha}_j^{(1)}\hat{\alpha}_k^{(1)}\}\right]\\ & + P_2^2\left[\frac{1}{N_2}\text{Var}\{\hat{\alpha}_j^{(2)}\} + \left(1 - \frac{1}{N_2}\right)\text{Cov}\{\hat{\alpha}_j^{(2)}\hat{\alpha}_k^{(2)}\}\right]\\ & - 2P_1P_2\,\text{Cov}\{\hat{\alpha}_j^{(1)}\hat{\alpha}_k^{(2)}\}\end{aligned} \tag{28}$$

where

$$\begin{aligned}\text{Var}\{\hat{\alpha}_j^{(i)}\} &= E\left\{\left[\frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{e^{j\omega\hat{h}(X)}}{j\omega}\,d\omega\right]^2\right\} - \left(\bar{\epsilon}_i - \frac{1}{2}\right)^2\\ &= \frac{1}{4} - \left(\bar{\epsilon}_i - \frac{1}{2}\right)^2\\ &= \bar{\epsilon}_i(1 - \bar{\epsilon}_i)\end{aligned} \tag{29}$$

$$\begin{aligned}\text{Cov}\{\hat{\alpha}_j^{(i)}\hat{\alpha}_k^{(l)}\} = {} & \frac{1}{4\pi^2}\int_{S_x}\int_{S_y}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdot\frac{E_d\{e^{j\omega_1\hat{h}(X)}e^{j\omega_2\hat{h}(Y)}\}}{j\omega_1 j\omega_2}p_i(X)\,p_l(Y)\\ & \cdot d\omega_1 d\omega_2\,dX\,dY - \bar{\alpha}_i\bar{\alpha}_l.\end{aligned} \tag{30}$$

The second line of (29) can be derived from the first line as is seen in (10). From (30), a portion of (28) can be expressed as

$$\begin{aligned}& P_1^2\,\text{Cov}\{\hat{\alpha}_j^{(1)}\hat{\alpha}_k^{(1)}\} + P_2^2\,\text{Cov}\{\hat{\alpha}_j^{(2)}\hat{\alpha}_k^{(2)}\} - 2P_1P_2\,\text{Cov}\{\hat{\alpha}_j^{(1)}\hat{\alpha}_k^{(2)}\}\\ &= \frac{1}{4\pi^2}\int_{S_x}\int_{S_y}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\frac{E_d\{e^{j\omega_1\hat{h}(X)}e^{j\omega_2\hat{h}(Y)}\}}{j\omega_1 j\omega_2}\cdot\tilde{p}(X)\,\tilde{p}(Y)\;d\omega_1 d\omega_2\,dX\,dY - \left(\bar{\epsilon} - \frac{1}{2}\right)^2\\ &= \text{Var}_d\{\hat{\varepsilon}\}\end{aligned} \tag{31}$$

where $\text{Var}_d\{\hat{\varepsilon}\}$ is the same one as (22). On the other hand, (30) can be approximated as

$$\begin{aligned}\text{Cov}\{\hat{\alpha}_j^{(i)}\hat{\alpha}_k^{(l)}\} &\cong \frac{1}{4\pi^2}\int_{S_x}\int_{S_y}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\cdot E_d\{\Delta\boldsymbol{h}(X)\,\Delta\boldsymbol{h}(Y)\}\,e^{j\omega_1 h(X)}\\ &\quad\cdot e^{j\omega_2 h(Y)}p_i(X)\,p_l(Y)\,d\omega_1 d\omega_2\,dX\,dY\\ &= \int_{h(X)=0}\int_{h(Y)=0}E_d\{\Delta\boldsymbol{h}(X)\,\Delta\boldsymbol{h}(Y)\}\cdot p_i(X)\,p_l(Y)\,dX\,dY\\ &\sim \frac{1}{\mathfrak{N}}.\end{aligned} \tag{32}$$

Equation (32) is proportional to $1/\mathfrak{N}$ because $E_d\{\Delta\boldsymbol{h}(X)\,\Delta\boldsymbol{h}(Y)\}$ is proportional to $1/\mathfrak{N}$.

Substituting (29)–(32) into (28), and ignoring the terms proportional to $1/N_i\mathfrak{N}$

$$\text{Var}\{\hat{\varepsilon}\} \cong P_1^2\frac{\bar{\epsilon}_1(1 - \bar{\epsilon}_1)}{N_1} + P_2^2\frac{\bar{\epsilon}_2(1 - \bar{\epsilon}_2)}{N_2} + \text{Var}_d\{\hat{\varepsilon}\}. \tag{33}$$

As we discussed in Section II-C, $\mathrm{Var}_d\{\hat{\varepsilon}\}$ is proportional to $1/\mathfrak{N}^2$ when the Bayes classifier is used for Gaussian distributions. Therefore, $\mathrm{Var}\{\hat{\varepsilon}\}$ of (33) is dominated by the first two terms which are due to the finite test set. A comparison of (33) and (13) shows that the effect of the finite design set appears in $\bar{\epsilon}_1$ and $\bar{\epsilon}_2$ of (33) instead of ϵ_1 and ϵ_2 of (13). That is, the bias due to the finite design set increases the variance proportionally. However, since $\bar{\epsilon}_1 - \epsilon_1 \sim 1/\mathfrak{N}$, this effect can be ignored. It should be noted that $\mathrm{Var}_d\{\hat{\varepsilon}\}$ could be proportional to $1/\mathfrak{N}$ if the classifier is not the Bayes.

Thus, we can draw the following conclusions from (27) and (33). When both design and test sets are finite,

1) the bias of the classification error comes entirely from the finite design set, and

2) the variance comes predominantly from the finite test set.

III. Dependent Design and Test Sets

In the previous section, we assumed that the design and test sets were finite and independent. When only one set of samples is available, independence can be achieved by using either the holdout method or the leave-one-out method. In the holdout method, the available sample set is divided into two groups; one group is used for designing the classifier and the other for testing the classifier. The ratio of design sample size to test sample size must be determined by the desired degradation and variance of the estimated error, as derived in Section II-D. On the other hand, in the leave-one-out method, each sample is tested by the classifier which was designed using the remaining samples [2]. With N available samples, the test sample size is 1, the design sample size is $N-1$ ($\cong N$), and the procedure is repeated N times.

Theoretically, due to the independence of the design and test sets, the analyses of the holdout and leave-one-out methods are the same. For the leave-one-out method, the empirical distribution of (6) becomes a single impulse anchored at the test sample. However, since the procedure is repeated N times, the summation across all N samples is still performed. Thus, for the holdout method, the first line of (7) has the summation inside the integration with respect to X and, for the leave-one-out method, the summation is outside of the integration. Using linearity, they both reduce to the second line of (7). The discriminant function is random, so (19) is used together with (7) to generate (23). Since independence is maintained within the summation (which now indexes both the discriminant function and the test sample), the arguments of Section II-D follow. Of course, for a fixed total available data set, the holdout method reduces the size of both the design ($\mathfrak{N}$) and test (N) sets, degrading its performance relative to the leave-one-out method.

It has been shown that the above procedures tend to give a larger error than the true one. The true error is the error of the classifier designed using the true distributions, tested with the true distributions. On the other hand, an error smaller than the true one can be obtained by the resubstitution method, in which all available samples are used to design the classifier and the same sample set is used to test the classifier. Since the resubstitution and leave-one-out methods can be carried out simultaneously without additional computation time [13], it is a common practice to compute both estimates to obtain upper and lower bounds of the true error.

When the resubstitution method is used, the design and test sample sets are no longer independent. In this section, we would like to address the dependency of the design and test sample sets. The expected value and variance of the resubstitution error and the statistical properties of the bias between the resubstitution and leave-one-out errors depend on the classifiers to be used. Therefore, in this section, we limit our discussions to parametric classifiers such as the quadratic and linear classifiers. Extending this discussion to other types of classifiers could be handled in a similar way. (In related but much less general discussions, Foley [1] and Raudys [14] address the resubstitution method for linear and Euclidean distance classifiers.)

A. Modifications of M and Σ

Let us assume that the expected vector M and covariance matrix Σ of a distribution are estimated from the available sample set, $X_1, \cdots, X_{N-1}$ by the sample mean and sample covariance as

$$\hat{M} = \frac{1}{N-1}\sum_{i=1}^{N-1} X_i \tag{34}$$

$$\hat{\Sigma} = \frac{1}{N-2}\sum_{i=1}^{N-1} (X_i - \hat{M})(X_i - \hat{M})^T. \tag{35}$$

When an additional sample Y is used, the above estimates are modified as

$$\hat{M}_R = \frac{1}{N}[(N-1)\hat{M} + Y] = \hat{M} + \frac{1}{N}(Y - \hat{M}) \tag{36}$$

or

$$Y - \hat{M}_R = \frac{N-1}{N}(Y - \hat{M}) \tag{37}$$

and

$$\hat{\Sigma}_R = \frac{1}{N-1}\left[\sum_{i=1}^{N-1}(X_i - \hat{M}_R)(X_i - \hat{M}_R)^T + (Y - \hat{M}_R)(Y_i - \hat{M}_R)^T\right]$$

$$= \hat{\Sigma} - \frac{1}{N-1}\hat{\Sigma} + \frac{1}{N}(Y - \hat{M})(Y - \hat{M})^T. \tag{38}$$

The deviations of these estimates from the true parameters, M and Σ, are

$$\Delta M_R = \Delta M + \frac{1}{N}(Y - M - \Delta M)$$

$$\cong \Delta M + \frac{1}{N}(Y - M) \tag{39}$$

$$\Delta\Sigma_R = \Delta\Sigma - \frac{1}{N-1}(\Sigma + \Delta\Sigma) + \frac{1}{N}(Y - M - \Delta M)(Y - M - \Delta M)^T \cong \Delta\Sigma - \frac{1}{N}\Sigma + \frac{1}{N}(Y - M)(Y - M)^T. \quad (40)$$

ΔM and $\Delta\Sigma$ assumed to be proportional to $1/N$ and approximations were made by ignoring $1/N^2$ and higher-order terms.

With this approximation, a function of $\hat{M}_R$ and $\hat{\Sigma}_R$, $f(\hat{M}_R, \hat{\Sigma}_R)$, can be expanded around $f(M, \Sigma)$ as

$$f(\hat{M}_R, \hat{\Sigma}_R) \cong f(M, \Sigma) + \frac{\partial f^T}{\partial M}\Delta M_R + \operatorname{tr}\frac{\partial f}{\partial \Sigma}\Delta\Sigma_R. \quad (41)$$

In the general Taylor series expansion, components of the second-order terms are also proportional to $1/N$. Using (39) and (40),

$$\Delta M_R \Delta M_R^T \cong \Delta M \Delta M^T + \frac{2}{N}(Y - M)\Delta M^T \quad (42)$$

$$\Delta\Sigma_R\Delta\Sigma_R^T \cong \Delta\Sigma\Delta\Sigma^T - \frac{2}{N}[\Sigma - (Y - M)(Y - M)^T]\Delta\Sigma^T \quad (43)$$

$$\Delta M_R \Delta\Sigma_R^T \cong \Delta M\Delta\Sigma^T - \frac{1}{N}\Delta M\Sigma^T + \frac{1}{N}\Delta M(Y - M)(Y - M)^T + \frac{1}{N}(Y - M)\Delta\Sigma^T. \quad (44)$$

In the above expression, each $1/N$ term contains a random variable which is assumed to be proportional to $1/N$, making the entire term proportional to $1/N^2$. Thus, (41) is consistent with the approximations made by ignoring $1/N^2$ and higher-order terms.

Substituting (39) and (40) into (41),

$$f(\hat{M}_R, \hat{\Sigma}_R) \cong \left[f(M, \Sigma) + \frac{\partial f^T}{\partial M}\Delta M + \operatorname{tr}\frac{\partial f}{\partial\Sigma}\Delta\Sigma\right] + \frac{1}{N}\left[\frac{\partial f^T}{\partial M}(Y - M) + \operatorname{tr}\frac{\partial f}{\partial\Sigma}\{(Y - M)(Y - M)^T - \Sigma\}\right] = f(\hat{M}, \hat{\Sigma}) + \frac{1}{N}\left[\frac{\partial f^T}{\partial M}(Y - M) + \operatorname{tr}\frac{\partial f}{\partial\Sigma}\{(Y - M)(Y - M)^T - \Sigma\}\right]. \quad (45)$$

Note that the difference between the two random variables $f(\hat{M}_R, \hat{\Sigma}_R)$ and $f(\hat{M}, \hat{\Sigma})$ is not random, as long as Y is fixed and the first-order approximation is valid.

Example: Let us examine the case where f is given by

$$f(M, \Sigma) = \tfrac{1}{2}(Y - M)^T\Sigma^{-1}(Y - M) + \tfrac{1}{2}\ln|\Sigma|. \quad (46)$$

Then,

$$\frac{\partial f}{\partial M} = -\Sigma^{-1}(Y - M) \quad (47)$$

$$\frac{\partial f}{\partial\Sigma} = -\frac{1}{2}\Sigma^{-1}(Y - M)(Y - M)^T\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}. \quad (48)$$

Therefore,

$$f(\hat{M}_R, \hat{\Sigma}_R) - f(\hat{M}, \hat{\Sigma}) \cong -\frac{1}{2N}[d^4(Y) + n] \quad (49)$$

where

$$d^2(Y) = (Y - M)^T\Sigma^{-1}(Y - M). \quad (50)$$

B. Quadratic Classifiers

In this section, the quadratic classifier of (14) is discussed. Using (46), (14) can be rewritten as

$$h(X) = f(M_1, \Sigma_1) - f(M_2, \Sigma_2). \quad (51)$$

When a sample X from ω_1 is tested in the resubstitution method,

$$\hat{h}_R(X) = f(\hat{M}_{1R}, \hat{\Sigma}_{1R}) - f(\hat{M}_2, \hat{\Sigma}_2) \cong f(\hat{M}_1, \hat{\Sigma}_1) - \frac{1}{2N_1}[d_1^4(X) + n] - f(\hat{M}_2, \hat{\Sigma}_2) = \hat{h}_L(X) - \frac{1}{2N_1}[d_1^4(X) + n] \quad \text{for } X \in \omega_1. \quad (52)$$

Likewise, when X comes from ω_2,

$$\hat{h}_R(X) \cong \hat{h}_L(X) + \frac{1}{2N_2}[d_2^4(X) + n] \quad \text{for } X \in \omega_2 \quad (53)$$

where $\hat{h}_R(X)$ and $\hat{h}_L(X)$ are the discriminant functions for the resubstitution and leave-one-out methods, N_i is the sample size for ω_i and $d_i^2(X) = (X - M_i)^T\Sigma_i^{-1}(X - M_i)$.

Now, the resubstitution error can be computed by (23) and (24) with $\hat{h}$ of (24) replaced by $\hat{h}_R$ of either (52) or (53) depending on $i = 1$ or 2. The result is

$$\hat{\varepsilon}_R = \frac{1}{2} + \frac{P_1}{N_1}\sum_{j=1}^{N_1}\frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{e^{j\omega\hat{h}_R(X_j^{(1)})}}{j\omega}d\omega - \frac{P_2}{N_2}\sum_{j=1}^{N_2}\frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{e^{j\omega\hat{h}_R(X_j^{(2)})}}{j\omega}d\omega \cong \hat{\varepsilon}_L - \left[\frac{P_1}{N_1^2}\sum_{j=1}^{N_1}\hat{\beta}_j^{(1)} + \frac{P_2}{N_2^2}\sum_{j=1}^{N_2}\hat{\beta}_j^{(2)}\right] \quad (54)$$

where

$$\hat{\beta}_j^{(i)} = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \frac{d_i^4(X_j^{(i)}) + n}{2} e^{j\omega \hat{h}_L(X_j^{(i)})}\, d\omega. \quad (55)$$

In order to obtain the second line of (54), an approximation of $e^{j\omega a/N} \cong 1 + j\omega a/N$ is used. Also, note that, in the leave-one-out method, the design and test samples are independent and, therefore, the discussion of Section II-D can be applied without modification. However, in the leave-one-out method, the number of design samples ($\mathcal{U}_i$) is always the same as the number of test samples (N_i).

Now, the statistical properties of the bias, $\hat{\varepsilon}_b = \hat{\varepsilon}_L - \hat{\varepsilon}_R$, can be studied. The expected value of $\hat{\varepsilon}_b$ is

$$E\{\hat{\varepsilon}_b\} \cong \frac{P_1}{N_1}\bar{\beta}_1 + \frac{P_2}{N_2}\bar{\beta}_2 \quad (56)$$

where

$$\bar{\beta}_i = \frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty} \frac{d_i^4(X) + n}{2} E_d\{e^{j\omega \hat{h}_L(X)}\}\, p_i(X)\, d\omega\, dX. \quad (57)$$

And, the variance of $\hat{\varepsilon}_b$ is

$$\begin{aligned}\mathrm{Var}\{\hat{\varepsilon}_b\} = {} & \frac{P_1^2}{N_1^2}\left[\frac{1}{N_1}\mathrm{Var}\{\hat{\beta}_j^{(1)}\} + \left(1 - \frac{1}{N_1}\right)\right. \\ & \left.\cdot\, \mathrm{Cov}\{\hat{\beta}_j^{(1)}\hat{\beta}_k^{(1)}\}\right] \\ & + \frac{P_2^2}{N_2^2}\left[\frac{1}{N_2}\mathrm{Var}\{\hat{\beta}_j^{(2)}\} + \left(1 - \frac{1}{N_2}\right)\right. \\ & \left.\cdot\, \mathrm{Cov}\{\hat{\beta}_j^{(2)}\hat{\beta}_k^{(2)}\}\right] \\ & + \frac{2P_1P_2}{N_1N_2}\mathrm{Cov}\{\hat{\beta}_j^{(1)}\hat{\beta}_k^{(2)}\}. \end{aligned} \quad (58)$$

The explicit expression for $\bar{\beta}_i$ of (57) can be obtained by using the same technique used to compute $\bar{\epsilon}$ in [12], if two distributions are Gaussian with $M_1 = 0$, $M_2 = M$ and $\Sigma_1 = \Sigma_2 = I$ and the quadratic classifier of (14) is used. For $N_1 = N_2 = N$

$$E_d\{e^{j\omega \hat{h}_L(X)}\} \cong e^{j\omega h(X)}\left[1 + \frac{1}{N}a\right] \cong e^{j\omega h(X)} \quad (59)$$

$$e^{j\omega h(X)}p_1(X) = \frac{\sqrt{2\pi}}{\sqrt{M^TM}} e^{-M^TM/8} N_\omega\left(-\frac{j}{2}, \frac{1}{M^TM}\right) \cdot N_x(j\omega M, I) \quad (60)$$

$$e^{j\omega h(X)}p_2(X) = \frac{\sqrt{2\pi}}{\sqrt{M^TM}} e^{-M^TM/8} N_\omega\left(\frac{j}{2}, \frac{1}{M^TM}\right) \cdot N_x((1 + j\omega)M, I) \quad (61)$$

where a is a constant given in [12]. $N_\omega(d, k)$ and $N_x(D, K)$ are Gaussian density functions of ω and X with the expected value d and variance k for N_ω, and the expected vector D and covariance matrix K for N_X. Thus, the integration of (57) merely involves computing the moments of the Gaussian distributions of (60) and (61), resulting in

$$\bar{\beta}_i \cong \frac{1}{2\sqrt{2\pi M^TM}} e^{-M^TM/8}\left[n^2 + (1 + M^TM/2)n + \left[(M^TM)^2/16 - M^TM/2 - 1\right]\right]. \quad (62)$$

The first lines of Table II show the values of $E\{\hat{\varepsilon}_b\}$ computed from (56) and (62) with $M^TM = 2.56^2$ and $P_1 = P_2 = 0.5$ for various k ($= N/n$) and n. The theoretical values are compared with the experimental ones in the second lines. The experiments were conducted by generating N samples, estimating M_i and Σ_i, designing the quadratic classifier of (14), estimating the resubstitution and leave-one-out errors and computing the bias between them. The experiment was repeated 10 times and the average and standard deviation of the estimated biases are listed in the second and third lines. As Table II shows, the first and second lines are close, confirming the validity of our discussion.

An important fact is that, from (56) and (62), $E\{\hat{\varepsilon}_b\}$ is roughly proportional to n^2/N for large n. A simpler explanation for this fact can be obtained by observing (57) more closely. Assuming (59) and carrying through the integration of (57) with respect to ω,

$$\begin{aligned}\bar{\beta}_i &\cong \int_S \frac{d_i^4(X) + n}{2}\delta(h(X))\, p_i(X)\, dX \\ &= \int_{h(X)=0} \frac{d_i^4(X) + n}{2} p_i(X)\, dX. \end{aligned} \quad (63)$$

It is well known that $d_i^2(X)$ is χ^2-distributed with an expected value of n and standard deviation of $\sqrt{2n}$, if X is Gaussianly distributed. Particularly when n is large, $d_i^2(X)$ on the classification boundary should be n times some number not far from 1. That is, $d_i^4(X)$ is close to n^2. Thus, $\bar{\beta}_i$ should be proportional to n^2.

The analysis of the variance (58) is more complex. Although the order of magnitude may not be immediately clear from (58), our experimental results, presented in Fig. 2 and the third line of Table II, show that the standard deviation is roughly proportional to $1/N$. The intuitive explanation should be the same as that presented in Section II-C.

C. Effect of Outliers

It is widely believed in the pattern recognition field that classifier performance can be improved by removing outliers, points far from a class's inferred mean which seem to distort the distribution. The approach used in Section III-A to analyze the difference between the resubstitution and leave-one-out parameters can be extended to handle the effect of a single point of the design set on classifier performance.

As in (34)–(38), assume that $N - 1$ samples have been used to estimate a distribution's parameters ($\hat{M}$, $\hat{\Sigma}$) and

TABLE II
BIAS BETWEEN LEAVE-ONE-OUT AND RESUBSTITUTION FOR I-I (%)

		n			
	4	8	16	32	64
3	9.00	13.79	23.03	41.34	77.87
	13.33	15.42	19.69	22.86	30.29
	7.03	5.22	4.12	4.26	3.40
5	5.40	8.27	13.82	24.80	46.72
	7.50	9.25	10.75	17.75	24.47
	4.56	3.24	2.28	2.69	1.53
10	2.70	4.14	6.91	12.40	23.36
	2.25	4.63	6.34	9.58	16.01
	1.84	2.02	1.59	1.61	1.24
15	1.80	2.76	4.61	8.27	15.57
	1.33	3.13	4.42	7.44	11.92
k	0.90	1.29	0.87	0.47	1.18
20	1.35	2.07	3.45	6.20	11.68
	1.38	2.09	3.14	5.05	9.56
	1.05	1.00	0.64	0.53	0.45
30	0.90	1.38	2.30	4.13	7.79
	0.63	1.58	2.39	3.94	6.41
	0.45	0.52	0.41	0.35	0.33
40	0.67	1.03	1.73	3.10	5.84
	0.44	1.08	1.55	2.96	5.21
	0.30	0.39	0.30	0.30	0.36
50	0.54	0.83	1.38	2.48	4.67
	0.30	0.75	1.38	2.29	4.27
	0.23	0.23	0.37	0.25	0.25

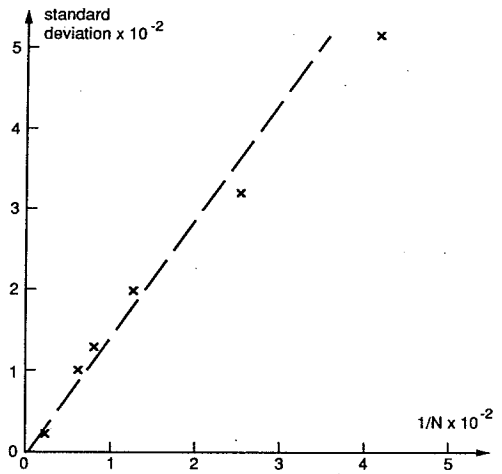


Fig. 2. Bias between leave-one-out and resubstitution errors for I-I (standard deviation versus $1/N$ for $n = 8$).

that these estimates will now be modified by including a new point Y. These new estimates $(\hat{M}_y, \hat{\Sigma}_y)$ are defined by (36) and (38). The approximations in (39)–(44) are still valid, so (45) can also be used. For the quadratic classifier, (47) and (48) can be substituted into (45) to yield

$$\begin{aligned} & f(\hat{M}_y, \hat{\Sigma}_y) - f(\hat{M}, \hat{\Sigma}) \\ &= \frac{1}{2N}\left[-2(Y-M)^T\Sigma^{-1}(X-M)\right. \\ &\quad + (Y-M)^T\Sigma^{-1}(Y-M) - n \\ &\quad - \left\{(Y-M)^T\Sigma^{-1}(X-M)\right\}^2 \\ &\quad \left. + (X-M)^T\Sigma^{-1}(X-M)\right] \\ &= \frac{1}{N}g(X-Y). \end{aligned} \tag{64}$$

The corresponding change in the discriminant function for $Y \in \omega_1$ can be found by inserting (64) into (51)

$$\begin{aligned} \hat{h}_y(X) &= f(\hat{M}_{1y}, \hat{\Sigma}_{1y}) - f(\hat{M}_2, \hat{\Sigma}_2) \\ &\cong f(\hat{M}_1, \hat{\Sigma}_1) + \frac{1}{N}g_1(X, Y) - f(\hat{M}_2, \hat{\Sigma}_2) \\ &= \hat{h}(X) + \frac{1}{N}g_1(X, Y) \qquad \text{for } Y \in \omega_1. \end{aligned} \tag{65}$$

Likewise, when Y comes from ω_2,

$$\hat{h}_y(X) = \hat{h}(X) - \frac{1}{N}g_2(X, Y) \qquad \text{for } Y \in \omega_2 \tag{66}$$

where g_i indicates that M_i and Σ_i are used instead of M and Σ in (64).

When this modified classifier is used on an independent set of test samples, the result is, using (19),

$$\begin{aligned} \hat{\varepsilon}_y &= \frac{1}{2} + \frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}\frac{e^{j\omega\hat{h}_y(X)}}{j\omega}\tilde{p}(X)\,d\omega\,dX \\ &\cong \frac{1}{2} + \frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}\frac{e^{j\omega\hat{h}(X)}}{j\omega}\left[1 \pm \frac{j\omega}{N}g_i(X, Y)\right] \\ &\quad \cdot \tilde{p}(X)\,d\omega\,dX \\ &= \hat{\varepsilon} \pm \frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}e^{j\omega\hat{h}(X)}\frac{1}{N}g_i(X, Y)\,\tilde{p}(X)\,d\omega\,dX \\ &\cong \hat{\varepsilon} \pm \frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}e^{j\omega h(X)}\frac{1}{N}g_i(X, Y)\,\tilde{p}(X)\,d\omega\,dX \end{aligned} \tag{67}$$

where $+$ and $i = 1$ are used for $Y \in \omega_1$ and $-$ and $i = 2$ are for $Y \in \omega_2$. The approximation in the last line involves expressing $e^{j\omega\hat{h}(X)}$ in terms of $e^{j\omega h(X)}$ and ignoring terms smaller than $1/N$. Unlike the case of the resubstitution error, (67) keeps $\tilde{p}(X)$ in its integrand. This makes the integral in (67) particularly easy to handle. If the quadratic classifier is the Bayes classifier, the integration with respect to ω results in

$$\Delta\hat{\varepsilon}_y = \pm\int_S \delta(h(X))\frac{1}{N}g_i(X, Y)\,\tilde{p}(X)\,dX = 0. \tag{68}$$

That is, as long as $\tilde{p}(X) = 0$ at $h(X) = 0$, the effect of an individual sample is negligible. Even if the quadratic classifier is not optimal, $\Delta\hat{\varepsilon}_y$ is dominated by a $1/N$ term. Thus, as one would expect, as the number of samples becomes larger, the effect of an individual sample diminishes.

These results were confirmed in three sets of experiments. The first was the mean difference case used earlier. In the second experiment, the two classes share a mean, but have different covariances (I for ω_1, $4I$ for ω_2). The third experiment used Standard Data from [13] where the classes differ widely in both the mean and the covariance. Eight-dimensional data was used in each case.

The experiments were run in the following manner. N samples were generated for each class. Then, an addi-

TABLE III
BIAS BETWEEN ERROR WITHOUT OUTLIER AND ERROR INCLUDING OUTLIER FOR VARIOUS OUTLIER DISTANCES FROM THE CLASS 1 MEAN

(a) FOR I-I ($\epsilon^* = 10$ percent)
(b) FOR I-4I ($\epsilon^* = 9$ percent)
(c) FOR STANDARD DATA ($\epsilon^* = 1.9$ percent)

N	ERROR WITHOUT OUTLIER (%)	BIAS BETWEEN ERROR WITHOUT OUTLIER AND ERROR INCLUDING OUTLIER (%)			
		$d^2 = n/2$	$d^2 = n$	$d^2 = 2n$	$d^2 = 3n$
24	20.18	0.519	0.689	0.769	0.762
40	15.61	0.124	0.211	0.279	0.274
80	12.04	0.029	0.035	0.027	0.018
120	11.71	0.008	0.012	0.011	0.003
160	11.04	0.006	0.010	0.014	0.013
240	10.74	0.004	0.006	0.010	0.001
320	10.53	0.004	0.006	0.009	0.011
400	10.34	-.001	-.001	-.003	-.001

(a)

N	ERROR WITHOUT OUTLIER (%)	BIAS BETWEEN ERROR WITHOUT OUTLIER AND ERROR INCLUDING OUTLIER (%)			
		$d^2 = n/2$	$d^2 = n$	$d^2 = 2n$	$d^2 = 3n$
24	23.53	0.792	1.213	1.451	1.356
40	16.19	0.222	0.423	0.619	0.658
80	11.79	0.025	0.060	0.091	0.083
120	10.83	0.015	0.032	0.047	0.045
160	10.32	-.003	0.005	0.014	0.013
240	9.92	0.003	0.012	0.025	0.034
320	9.52	0.003	0.006	0.012	0.015
400	9.41	0.000	0.000	0.001	-.001

(b)

N	ERROR WITHOUT OUTLIER (%)	BIAS BETWEEN ERROR WITHOUT OUTLIER AND ERROR INCLUDING OUTLIER (%)			
		$d^2 = n/2$	$d^2 = n$	$d^2 = 2n$	$d^2 = 3n$
24	5.58	0.374	0.555	0.664	0.673
40	3.70	0.054	0.088	0.110	0.103
80	2.54	0.005	0.007	0.008	0.003
120	2.35	0.005	0.007	0.007	0.005
160	2.25	-.001	0.000	0.001	0.001
240	2.14	0.001	0.002	0.003	0.004
320	2.08	0.000	0.000	0.001	0.001
400	2.05	0.000	0.000	0.000	0.000

(c)

tional sample Y was generated from class 1 and scaled to a specific normalized distance from the mean. Classifiers were designed with and without Y and were tested on the true distributions using Novak's program computing the performance of a classifier with a given test distribution [4]. This procedure was repeated 10 times for each particular value of N. The entire process was run a number of times with varying distances. Experimental results are presented in Table III. Notice that even when the squared distance is much larger than its expected value n, the outlier's effect is still negligible.

IV. BOOTSTRAP METHODS

As an alternative to the holdout and leave-one-out error estimates, Efron [3] has suggested using a bootstrap technique to estimate the optimistic bias of the resubstitution error and, in turn, to estimate the expected error rate for a given decision rule. In the bootstrap procedure, one assumes that the existing sample set represents the true distributions. That is, these density functions consist of impulses located at the existing sample points

$$p_i^*(X) = \frac{1}{N_i} \sum_{j=1}^{N_i} \delta(X - X_j^{(i)}) \qquad i = 1, 2 \tag{69}$$

where * indicates something related to the bootstrap operation. Note that in this section, $X_j^{(i)}$ is considered a given fixed vector and is not random as it was in the previous sections.

When samples are drawn from $p_i^*(X)$ randomly, we select only the existing sample points with random frequencies. Thus, the N_i samples drawn from $p_i^*(X)$ form a density function

$$\hat{p}_i^*(X) = \sum_{j=1}^{N_i} \theta_j^{(i)} \delta(X - X_j^{(i)}) \qquad i = 1, 2. \tag{70}$$

Within each class, the $\theta_j^{(i)}$'s are identically distributed under the condition $\Sigma_{j=1}^{N_i} \theta_j^{(i)} = 1$. Their statistical properties are known [3]:

$$E\{\theta_j^{(i)}\} = \frac{1}{N_i} \tag{71}$$

$$E\{\theta_j^{(i)}\theta_k^{(i)}\} = \frac{1}{N_i^2}\delta_{jk} - \frac{1}{N_i^3} \tag{72}$$

$$E\{\theta_j^{(i)}\theta_k^{(l)}\} = 0, \qquad \text{for } i \neq l. \tag{73}$$

The error estimate with independent design and test samples in the bootstrap procedure (which, for lack of a better term, we will call the holdout error), $\hat{\varepsilon}_H^*$, is obtained by generating samples, designing a classifier based on $\hat{p}_i^*(X)$ and testing $\hat{p}_i^*(X)$. On the other hand, the resubstitution error $\hat{\varepsilon}_R^*$ is computed by testing $\hat{p}_i^*(X)$. The bias between them can be expressed by

$$\begin{aligned}\hat{\varepsilon}_b^* &= \hat{\varepsilon}_H^* - \hat{\varepsilon}_R^* \\ &= \frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty} \frac{e^{j\omega\hat{h}^*(X)}}{j\omega}\left[P_1\sum_{j=1}^{N_1}\left(\frac{1}{N_1} - \theta_j^{(1)}\right)\right. \\ &\quad \cdot \delta(X - X_j^{(1)}) \\ &\quad \left. - P_2\sum_{j=1}^{N_2}\left(\frac{1}{N_2} - \theta_j^{(2)}\right)\delta(X - X_j^{(2)})\right] d\omega\, dX \\ &= P_1\sum_{j=1}^{N_1}\gamma_j^{(1)} - P_2\sum_{j=1}^{N_2}\gamma_j^{(2)}\end{aligned} \tag{74}$$

where

$$\gamma_j^{(i)} = -\frac{\Delta\theta_j^{(i)}}{2\pi}\int_{-\infty}^{+\infty}\frac{e^{j\omega\hat{h}^*(X_j^{(i)})}}{j\omega}\,d\omega \tag{75}$$

and $\Delta\theta_j^{(i)} = \theta_j^{(i)} - (1/N_i)$.

When a quadratic classifier is used, $\hat{h}^*(X)$ in (75) becomes

$$\hat{h}^*(X) = f(\hat{M}_1^*, \hat{\Sigma}_1^*) - f(\hat{M}_2^*, \hat{\Sigma}_2^*) \tag{76}$$

where $f(\cdot, \cdot)$ is defined in (46). The bootstrap parameters, $\hat{M}_i^*$ and $\hat{\Sigma}_i^*$ are

$$\hat{M}_i^* = \sum_{j=1}^{N_i} \theta_j^{(i)} X_j^{(i)} \tag{77}$$

$$\hat{\Sigma}_i^* = \sum_{j=1}^{N_i} \theta_j^{(i)} (X_j^{(i)} - \hat{M}_i)(X_j^{(i)} - \hat{M}_i)^T. \tag{78}$$

Note that $\hat{M}_i = (1/N_i) \sum_{j=1}^{N_i} X_j^{(i)}$ is used to compute $\hat{\Sigma}_i^*$. $\hat{M}_i$ is available in the bootstrap operation and the use of $\hat{M}_i$ instead of $\hat{M}_i^*$ simplifies the discussion significantly. Their expectations are

$$E_*\{\hat{M}_i^*\} = \sum_{j=1}^{N_i} E\{\theta_j^{(i)}\} X_j^{(i)} = \frac{1}{N} \sum_{j=1}^{N_i} X_j^{(i)} = \hat{M}_i \tag{79}$$

$$E_*\{\hat{\Sigma}_i^*\} = \sum_{j=1}^{N_i} E\{\theta_j^{(i)}\} (X_j^{(i)} - \hat{M}_i)(X_j^{(i)} - \hat{M}_i)^T = \frac{N_i - 1}{N_i} \hat{\Sigma}_i \cong \hat{\Sigma}_i \tag{80}$$

where E_* indicates the expectation with respect to the θ's. $f(\hat{M}_i^*, \hat{\Sigma}_i^*)$ can be expanded around $f(\hat{M}_i, \hat{\Sigma}_i)$ by the Taylor series as

$$f(\hat{M}_i^*, \hat{\Sigma}_i^*) \cong f(\hat{M}_i, \hat{\Sigma}_i) + \frac{\partial f^T}{\partial \hat{M}_i} \Delta \hat{M}_i + \operatorname{tr} \frac{\partial f}{\partial \hat{\Sigma}_i} \Delta \hat{\Sigma}_i \tag{81}$$

where $\Delta \hat{M}_i = \hat{M}_i^* - \hat{M}_i$ and $\Delta \hat{\Sigma}_i = \hat{\Sigma}_i^* - \hat{\Sigma}_i$. Since $\hat{h}(X) = f(\hat{M}_1, \hat{\Sigma}_1) - f(\hat{M}_2, \hat{\Sigma}_2)$,

$$\Delta \hat{h}(X) = \hat{h}^*(X) - \hat{h}(X) \cong \frac{\partial f^T}{\partial \hat{M}_1} \Delta \hat{M}_1 - \frac{\partial f^T}{\partial \hat{M}_2} \Delta \hat{M}_2 + \operatorname{tr}\left(\frac{\partial f}{\partial \hat{\Sigma}_1} \Delta \hat{\Sigma}_1 - \frac{\partial f}{\partial \hat{\Sigma}_2} \Delta \hat{\Sigma}_2 \right). \tag{82}$$

The partial derivatives of (82) can be obtained by (47) and (48).

A. Bootstrap Expectation

Using the approximation of (21), (75) can be approximated as

$$\gamma_j^{(i)} \cong -\frac{\Delta \theta_j^{(i)}}{2\pi} \int_{-\infty}^{+\infty} \frac{e^{j\omega \hat{h}(X_j^{(i)})}}{j\omega} \left[1 + j\omega \Delta \hat{h}(X_j^{(i)}) + \frac{(j\omega)^2}{2} \Delta \hat{h}^2(X_j^{(i)}) \right] d\omega. \tag{83}$$

The third term contains third-order moments with the combination of $\Delta\theta_j^{(i)}$ and $\Delta \hat{h}^2$ and can be ignored. Thus, our analysis will focus on the first and second terms. With this in mind, substituting (77), (78) and (82) into (83) produces

$$\begin{aligned} \gamma_j^{(i)} \cong & -\frac{\Delta \theta_j^{(i)}}{2\pi} \int_{-\infty}^{+\infty} \frac{e^{j\omega \hat{h}(X_j^{(i)})}}{j\omega} d\omega \\ & - \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{j\omega \hat{h}(X_j^{(i)})} \left[\frac{\partial f^T}{\partial \hat{M}_1} \sum_{k=1}^{N_1} \Delta \theta_j^{(i)} \Delta \theta_k^{(1)} X_k^{(1)} \right. \\ & - \frac{\partial f}{\partial \hat{M}_2} \sum_{k=1}^{N_2} \Delta \theta_j^{(i)} \Delta \theta_k^{(2)} X_k^{(2)} \\ & + \sum_{k=1}^{N_1} \Delta \theta_j^{(i)} \Delta \theta_k^{(1)} (X_k^{(1)} - \hat{M}_1)^T \frac{\partial f}{\partial \hat{\Sigma}_1} (X_k^{(1)} - \hat{M}_1) \\ & - \sum_{k=1}^{N_2} \Delta \theta_j^{(i)} \Delta \theta_k^{(2)} (X_k^{(2)} - \hat{M}_2)^T \frac{\partial f}{\partial \hat{\Sigma}_2} \\ & \left. \cdot (X_k^{(2)} - \hat{M}_2) \right] d\omega. \end{aligned} \tag{84}$$

Using the partial derivatives of (47) and (48) and the expectations in (71)–(73), $E_*\{\gamma_j^{(i)}\}$ becomes

$$E_*\{\gamma_j^{(i)}\} \cong \frac{(-1)^{i+1}}{N_i^2} \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{\hat{d}_i^4(X_j^{(i)}) + n}{2} \cdot e^{j\omega \hat{h}(X_j^{(i)})} d\omega \tag{85}$$

where

$$\hat{d}_i^2(X) = (X - \hat{M}_i)^T \hat{\Sigma}_i^{-1} (X - \hat{M}_i). \tag{86}$$

In the derivation of (85), we utilized the relationship that

$$\begin{aligned} \frac{1}{N_i} \sum_{k=1}^{N_i} (X_k^{(i)} - \hat{M}_i)^T \hat{\Sigma}_i^{-1} (X_k^{(i)} - \hat{M}_i) &= \operatorname{tr} \hat{\Sigma}_i^{-1} \frac{1}{N_i} \sum_{k=1}^{N_i} (X_k^{(i)} - \hat{M}_i)(X_k^{(i)} - \hat{M}_i)^T \\ &= \frac{N_i - 1}{N_i} \operatorname{tr} \hat{\Sigma}_i^{-1} \hat{\Sigma}_i \\ &= \frac{N_i - 1}{N_i} n \cong n. \end{aligned}$$

Thus, the expectation of the bootstrap bias for a quadratic classifier given a sample set $S = \{X_1^{(1)}, \cdots, X_{N_1}^{(1)}, X_1^{(2)}, \cdots, X_{N_2}^{(2)}\}$ becomes

$$E_*\{\hat{\varepsilon}_b^* \mid S\} = \frac{P_1}{N_1^2} \sum_{j=1}^{N_1} \hat{\beta}_j^{*(1)} + \frac{P_2}{N_2^2} \sum_{j=1}^{N_2} \hat{\beta}_j^{*(2)} \tag{87}$$

where

$$\hat{\beta}_j^{*(i)} = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{\hat{d}_i^4(X_j^{(i)}) + n}{2} e^{j\omega \hat{h}(X_j^{(i)})} d\omega. \tag{88}$$

Note that (55) and (88) are very similar. The differences are $\hat{d}_i^2$ of (86) versus d_i^2 of (50) and $\hat{h}$ versus $\hat{h}_L$. $\hat{h}$ is the discriminant function designed with $\hat{M}_i$ and $\hat{\Sigma}_i$, the sample mean and sample covariance of the sample set S. The test samples $X_j^{(i)}$ are the members of the same set S. Therefore, $\hat{h}$ is the same as the resubstitution discriminant

TABLE IV
BOOTSTRAP RESULTS

(a) FOR I-I ($\epsilon^* = 10$ PERCENT)

	CONVENTIONAL LEAVE-ONE-OUT & RESUBSTITUTION			BOOTSTRAP		
N	$E\{\epsilon_R\}$	$E\{\epsilon_L^*\}$	$E\{\epsilon_L - \epsilon_R\}$	$E_S\{\epsilon_R + E_*\{\epsilon_b^* \mid S\}\}$	$E_S E_*\{\epsilon_b^* \mid S\}$	$E_S \mathrm{Var}_*\{\epsilon_b^* \mid S\}$
24	3.54	17.08	13.54	12.77	9.23	0.18
	0.11	4.89	3.14	4.17	1.38	0.04
40	5.75	13.38	7.63	11.68	5.92	0.08
	0.07	6.04	3.88	4.44	1.90	0.02
80	7.13	11.19	4.06	10.67	3.55	0.04
	0.04	2.47	1.29	2.50	0.56	0.01
120	9.04	11.79	2.75	11.45	2.41	0.03
	0.06	2.97	1.01	2.79	0.43	0.01
160	9.13	11.28	2.16	11.17	2.05	0.02
	0.03	2.35	1.09	1.94	0.44	0.00
240	8.27	9.35	1.08	9.46	1.19	0.01
	0.02	1.61	0.51	1.61	0.15	0.00
320	9.78	10.67	0.89	10.78	1.00	0.01
	0.01	0.80	0.37	0.91	0.11	0.00
400	9.18	9.78	0.60	9.96	0.77	0.01
	0.01	0.91	0.26	0.84	0.10	0.00

(a)

function $\hat{h}_R$ of the previous sections, while $\hat{h}_L$ is the leave-one-out discriminant function. As is shown in (52) and (53), the difference between $\hat{h}_L$ and $\hat{h}_R$ is proportional to $1/N$. Thus, the difference between $e^{j\omega\hat{h}_L}$ and $e^{j\omega\hat{h}_R}$ is proportional to $1/N$. Also, as (50) suggests, it can be shown that the difference between $\hat{d}_i^2$ and d_i^2 is proportional to $1/N$. Thus, ignoring terms with $1/N$, $\hat{\varepsilon}_b$ of (54) and $E_*\{\hat{\varepsilon}_b^* \mid S\}$ of (87) (note that S is now a random set) become equal and have the same statistical properties. Practically, this means that estimating the expected error rate using the leave-one-out and bootstrap methods should yield the same results.

These conclusions have been confirmed experimentally. For several values of N_i, 8-dimensional sample vectors were generated from the Gaussian distributions used in Section III. The generated samples were bootstrapped and used to design a quadratic classifier. This classifier was then tested on the original sample set ($\hat{\varepsilon}_H^*$) and the bootstrap sample set ($\hat{\varepsilon}_R^*$). Each sample set (S) was bootstrapped 100 times and the results were averaged to simulate the bootstrap expectation ($E_*\{\hat{\varepsilon}_b^* \mid S\}$.) The whole procedure was repeated 10 times to estimate the expectation with respect to the training sample set ($E_S E_*\{\hat{\varepsilon}_b^* \mid S\}$.) Results are presented in Table IV. In columns 3–7, the first line of each entry is the mean of 10 trials and the second line is the standard deviation. In column 2, the first line is still the mean, but the variance is presented in the second line.

When N_i is particularly small, our approximations might not be valid and the leave-one-out and bootstrap methods may produce different results. Although the bootstrap bias estimate does seem to have a slightly smaller standard deviation (column 4 versus column 6 of Table IV), both our results and those presented in Jain, Dubes, and Chen [15] show that the leave-one-out and bootstrap methods are equivalent (column 3 versus column 5 of Table IV).

B. Bootstrap Variance

The variance with respect to the bootstrap can be evaluated in a fashion similar to (58)

$$\begin{aligned}\mathrm{Var}_*\{\hat{\varepsilon}_b^* \mid S\} = P_1^2 \Bigg[\sum_{j=1}^{N_1} \mathrm{Var}_*\{\gamma_j^{(1)}\} + \sum_{\substack{j=1 \\ j\neq k}}^{N_1} \sum_{k=1}^{N_1} \cdot \mathrm{Cov}_*\{\gamma_j^{(1)}\gamma_k^{(1)}\} \Bigg] \\ + P_2^2 \Bigg[\sum_{j=1}^{N_2} \mathrm{Var}_*\{\gamma_j^{(2)}\} + \sum_{\substack{j=1 \\ j\neq k}}^{N_1} \sum_{k=1}^{N_2} \cdot \mathrm{Cov}_*\{\gamma_j^{(2)}\gamma_k^{(2)}\} \Bigg] \\ - 2P_1P_2 \sum_{j=1}^{N_1} \sum_{k=1}^{N_2} \mathrm{Cov}_*\{\gamma_j^{(1)}\gamma_k^{(2)}\}. \end{aligned} \tag{89}$$

TABLE IV (*Continued.*)

(b) FOR I-4I ($\epsilon^* = 9$ PERCENT)
(c) FOR STANDARD DATA ($\epsilon^* = 1.9$ PERCENT)

	CONVENTIONAL LEAVE-ONE-OUT & RESUBSTITUTION			BOOTSTRAP		
N	$E\{\epsilon_R\}$	$E\{\epsilon_L\}$	$E\{\epsilon_L-\epsilon_R\}$	$E_S\{\epsilon_R+E_*\{\epsilon_b^* \mid S\}\}$	$E_S E_*\{\epsilon_b^* \mid S\}$	$E_S \mathrm{Var}_*\{\epsilon_b^* \mid S\}$
24	3.54	18.33	14.79	15.08	11.54	0.21
	0.12	4.79	3.86	4.35	1.26	0.03
40	4.88	13.75	8.88	12.10	7.22	0.12
	0.06	3.23	2.97	2.27	0.92	0.03
80	7.19	11.19	4.00	10.82	3.63	0.04
	0.08	2.72	1.56	3.12	0.54	0.01
120	8.25	10.75	2.50	10.86	2.61	0.03
	0.03	2.14	1.23	2.04	0.37	0.01
160	7.59	9.88	2.28	9.56	1.96	0.02
	0.01	1.58	0.68	1.23	0.33	0.00
240	8.38	9.75	1.38	9.80	1.42	0.02
	0.03	1.94	0.49	1.83	0.22	0.00
320	9.11	10.09	0.98	10.14	1.03	0.01
	0.71	0.83	0.40	0.77	0.15	0.00
400	9.09	9.99	0.90	9.99	0.90	0.01
	0.01	0.95	0.24	0.89	0.13	0.00

(b)

	CONVENTIONAL LEAVE-ONE-OUT & RESUBSTITUTION			BOOTSTRAP		
N	$E\{\epsilon_R\}$	$E\{\epsilon_L\}$	$E\{\epsilon_L-\epsilon_R\}$	$E_S\{\epsilon_R+E_*\{\epsilon_b^* \mid S\}\}$	$E_S E_*\{\epsilon_b^* \mid S\}$	$E_S \mathrm{Var}_*\{\epsilon_b^* \mid S\}$
24	0.63	5.00	4.38	4.14	3.52	0.10
	0.01	3.43	3.02	1.69	0.84	0.02
40	1.88	3.63	1.75	3.74	1.86	0.03
	0.02	1.99	1.21	1.95	0.67	0.02
80	1.44	2.31	0.88	2.26	0.82	0.01
	0.01	1.10	0.94	1.08	0.24	0.00
120	1.75	2.71	0.96	2.31	0.56	0.01
	0.01	1.04	0.48	1.05	0.19	0.00
160	1.94	2.34	0.41	2.35	0.42	0.01
	0.00	0.90	0.36	0.80	0.17	0.00
240	2.21	2.50	0.29	2.50	0.29	0.00
	0.00	0.71	0.26	0.60	0.13	0.00
320	2.00	2.17	0.17	2.18	0.18	0.00
	0.00	0.48	0.14	0.53	0.07	0.00
400	2.01	2.24	0.23	2.21	0.19	0.00
	0.00	0.45	0.16	0.38	0.07	0.00

(c)

Because the samples from each class were bootstrapped independently, $\mathrm{Cov}_*\{\gamma_j^{(1)}\gamma_k^{(2)}\} = 0$.

Using a property of the inverse Fourier transform,

$$\gamma_j^{(i)} = -\frac{\Delta\theta_j^{(i)}}{2\pi}\int_{-\infty}^{+\infty}\frac{e^{j\omega\hat{h}^*(X_j^{(i)})}}{j\omega}\,d\omega = -\frac{1}{2}\operatorname{sgn}\left(\hat{h}^*(X_j^{(i)})\right)\Delta\theta_j^{(i)}. \tag{90}$$

Thus, the variance of $\gamma_j^{(i)}$ is

$$\mathrm{Var}_*\{\gamma_j^{(1)}\} = E_*\{\gamma_j^{(i)2}\} - E_*^2\{\gamma_j^{(i)}\} = \frac{1}{4}E\{\Delta\theta_j^{(i)2}\} - E_*^2\{\gamma_j^{(i)}\} \cong \frac{1}{4}\left(\frac{1}{N_i^2} - \frac{1}{N_i^3}\right) \tag{91}$$

where $E_*^2\{\gamma_j^{(i)}\}$ is proportional to $1/N_i^4$ from (85) and therefore can be ignored. $\mathrm{Cov}_*\{\gamma_j^{(i)}\gamma_k^{(i)}\}$ may be approximated by using the first term only of (84). Again, using (90),

$$\begin{aligned}\mathrm{Cov}_*\{\gamma_j^{(i)}\gamma_k^{(i)}\} &= E_*\{\gamma_j^{(i)}\gamma_k^{(i)}\} - E_*\{\gamma_j^{(i)}\}E_*\{\gamma_k^{(i)}\} \\ &\cong \frac{1}{4}\operatorname{sgn}\left(\hat{\boldsymbol{h}}(X_j^{(i)})\right)\operatorname{sgn}\left(\hat{\boldsymbol{h}}(X_k^{(i)})\right)E\{\Delta\boldsymbol{\theta}_j^{(i)}\Delta\boldsymbol{\theta}_k^{(i)}\} \\ &\quad - E_*\{\gamma_j^{(i)}\}E\{\gamma_k^{(i)}\} \\ &\cong -\frac{1}{4N_i^3}\operatorname{sgn}\left(\hat{\boldsymbol{h}}(X_j^{(i)})\right)\operatorname{sgn}\left(\hat{\boldsymbol{h}}(X_k^{(i)})\right) \qquad (92)\end{aligned}$$

where $E\{\Delta\boldsymbol{\theta}_j^{(i)}\Delta\boldsymbol{\theta}_k^{(i)}\} = -1/N_i^3$ for $j \neq k$ by (72), and $E_*\{\gamma_j^{(i)}\}E_*\{\gamma_k^{(i)}\}$ is proportional to $1/N_i^4$ by (85) and therefore can be ignored.

Thus, substituting (91) and (92) into (89) and using $\mathrm{Cov}_*\{\gamma_j^{(1)}\gamma_k^{(2)}\} = 0$,

$$\begin{aligned}\mathrm{Var}_*\{\hat{\varepsilon}_b^* \mid S\} &\cong \frac{1}{4}\sum_{i=1}^{2}\frac{P_i}{N_i}\left[1 - \sum_{j=1}^{N_i}\frac{\operatorname{sgn}\left(\hat{\boldsymbol{h}}(X_j^{(i)})\right)}{N_i}\right. \\ &\quad \left.\cdot\sum_{k=1}^{N_i}\frac{\operatorname{sgn}\left(\hat{\boldsymbol{h}}(X_k^{(i)})\right)}{N_i}\right] \\ &= \frac{1}{4}\sum_{i=1}^{2}\frac{P_i}{N_i}\left[1 - (1 - 2\hat{\varepsilon}_{Ri})(1 - 2\hat{\varepsilon}_{Ri})\right] \\ &= \sum_{i=1}^{2}P_i\frac{\hat{\varepsilon}_{Ri}(1 - \hat{\varepsilon}_{Ri})}{N_i}. \qquad (93)\end{aligned}$$

Note that $\Sigma\ \operatorname{sgn}\left(\hat{\boldsymbol{h}}(X_j^{(i)})\right)/N_i = (-1)^i[(\#\text{ of correctly classified }\omega_i\text{ samples by }\hat{\boldsymbol{h}} \lesseqgtr_{\omega_2}^{\omega_1} 0)/N_i - (\#\text{ of misclassified }\omega_i\text{ samples by }\hat{\boldsymbol{h}} \lesseqgtr_{\omega_2}^{\omega_1} 0)/N_i] = (-1)^i[(1 - \hat{\varepsilon}_{Ri}) - \hat{\varepsilon}_{Ri}] = (-1)^i(1 - 2\hat{\varepsilon}_{Ri})$. Since $\hat{\boldsymbol{h}}$ is the resubstitution discriminant function for the original sample set, the resulting error is the resubstitution error.

Note that (93) is the variance expression of the resubstitution error estimate. This is seen in Table IV (second line of column 2 versus first line of column 7) and theoretically substantiates a claim of Efron [3]. Also, note that, since (93) only involves bootstrap operations, this value can be estimated using just one set of samples. When $\boldsymbol{S}$ becomes a random set, $\mathrm{Var}_*\{\hat{\varepsilon}_b^* \mid \boldsymbol{S}\}$ varies with $\hat{\varepsilon}_{Ri}(1 - \hat{\varepsilon}_{Ri}) \cong \hat{\varepsilon}_{Ri}$.

V. Conclusions

The object of this paper was to apply the error expression derived in [12] to various classifier test procedures in order to theoretically analyze their estimates of the expected classifier performance. It was shown that the design samples alone account for the degradation in a classifier's performance, while the test samples dominate the variance of the error estimate. These results have been known. But, this paper offers a new theoretical approach to understanding how design and test sample sizes affect the performance of classifiers. A general expression showing the relationship between the resubstitution and leave-one-out estimates of functions of Gaussian parameters was derived. As an example, the statistical properties of the difference between the resubstitution and leave-one-out error estimates for the quadratic classifier were investigated. The difference was found to be inversely proportional to the number of design samples and roughly proportional to n^2. In a related discussion, the effect of outlier design samples was found to be negligible, other than their effective reduction of the number of design samples in the training set. Finally, Efron's bootstrap estimate of the optimistic bias of the resubstitution error was analyzed. The resulting error estimate was shown to be statistically equivalent to the leave-one-out error estimate under reasonable design conditions.

Although not exhaustive, this study should provide a better understanding of the role of dependent and independent design and test samples in classifier design and evaluation. Hopefully, the tools and methodology can be applied to other statistical testing procedures and may help propose new ones.

Appendix 1
Proof of $\hat{\varepsilon} \geqq \epsilon$

The first step is to prove that the first-order variation of (19) is zero regardless of $\Delta\boldsymbol{h}(X)$. From (21), the first-order variation of (19) is

$$\begin{aligned}&\frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}\Delta\boldsymbol{h}(X)\,e^{j\omega h(X)}\,\tilde{p}(X)\,d\omega\,dX \\ &\quad = \int_S \Delta\boldsymbol{h}(X)\delta(h(X))\,\tilde{p}(X)\,dX \\ &\quad = \int_{h(X)=0}\Delta\boldsymbol{h}(X)\,\tilde{p}(X)\,dX \\ &\quad = 0. \qquad (A1)\end{aligned}$$

The last equality comes from the fact that $\tilde{p}(X) = 0$ at $h(X) = 0$ if $h(X)$ is the Bayes classifier of $P_1p_1(X)$ and $P_2p_2(X)$.

The second step involves showing that the second-order variation of (19) is positive regardless of $\Delta\boldsymbol{h}(X)$. Again from (21)

$$\begin{aligned}&\frac{1}{2\pi}\int_S\int_{-\infty}^{+\infty}\frac{j\omega}{2}\Delta\boldsymbol{h}^2(X)\,e^{j\omega h(X)}\,\tilde{p}(X)\,d\omega\,dX \\ &\quad = \frac{1}{2}\int_S\Delta\boldsymbol{h}^2(X)\frac{d\delta(h)}{dh}\tilde{p}(X)\,dX. \qquad (A2)\end{aligned}$$

In the region very close to $h(X) = 0$, $d\delta(h)/dh > 0$ and $\tilde{p}(X) > 0$ for $h < 0$, while $d\delta(h)/dh < 0$ and $\tilde{p}(X) < 0$ for $h > 0$. Since $\Delta\boldsymbol{h}^2(X) > 0$ regardless of $\Delta\boldsymbol{h}(X)$, (A2) is always positive.

APPENDIX 2
DERIVATION OF VAR $\{\hat{\epsilon}\}$

Keeping up to the second-order terms of $\Delta \boldsymbol{h}$,

$$e^{j\omega_1\hat{h}(X)}e^{j\omega_2\hat{h}(Y)}$$
$$= e^{j\omega_1 h(X)}e^{j\omega_2 h(Y)}e^{j\omega_1\Delta h(X)}e^{j\omega_2\Delta h(Y)}$$
$$\cong e^{j\omega_1 h(X)}e^{j\omega_2 h(Y)}[1 + j\omega_1\Delta\zeta_1(X) + j\omega_2\Delta\zeta_2(Y) - \omega_1\omega_2\Delta h(X)\,\Delta h(Y)] \quad (A3)$$

where

$$\Delta\zeta_i(X) = \Delta h(X) + \frac{j\omega_i}{2}\Delta h^2(X). \quad (A4)$$

Thus, the first line of (22) can be expanded to

$$\mathrm{Var}_d\{\hat{\epsilon}\}$$
$$\cong \frac{1}{2\pi}\int_{S_x}\int_{-\infty}^{+\infty}\frac{e^{j\omega_1 h(X)}}{j\omega_1}\tilde{p}(X)\,d\omega_1\,dX \cdot \frac{1}{2\pi}\int_{S_y}\int_{-\infty}^{+\infty}\frac{e^{j\omega_2 h(Y)}}{j\omega_2}\tilde{p}(Y)\,d\omega_2\,dY$$
$$+ \frac{1}{2\pi}\int_{S_x}\int_{-\infty}^{+\infty}E_d\{\Delta\zeta_1(X)\}\,e^{j\omega_1 h(X)}\tilde{p}(X)\,d\omega_1\,dX \cdot \frac{1}{2\pi}\int_{S_y}\int_{-\infty}^{+\infty}\frac{e^{j\omega_2 h(Y)}}{j\omega_2}\tilde{p}(Y)\,d\omega_2\,dY$$
$$+ \frac{1}{2\pi}\int_{S_x}\int_{-\infty}^{+\infty}\frac{e^{j\omega_1 h(X)}}{j\omega_1}\tilde{p}(X)\,d\omega_1\,dX \cdot \frac{1}{2\pi}\int_{S_y}\int_{-\infty}^{+\infty}E_d\{\Delta\zeta_2(Y)\}\,e^{j\omega_2 h(Y)}\tilde{p}(Y)\,d\omega_2\,dY$$
$$+ \frac{1}{4\pi^2}\int_{S_y}\int_{S_x}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}E_d\{\Delta h(X)\,\Delta h(Y)\} \cdot e^{j\omega_1 h(X)}e^{j\omega_2 h(Y)}\tilde{p}(X)\tilde{p}(Y)\,d\omega_1 d\omega_2\,dX\,dY$$
$$- \left(\bar{\epsilon} - \frac{1}{2}\right)^2. \quad (A5)$$

The first line of (A5) is $(\epsilon - \frac{1}{2})^2$ from (4), and the second and third lines are each $(\epsilon - \frac{1}{2})\overline{\Delta\varepsilon}$ from (21). Furthermore, the summation of the first, second, third, and fifth lines is $(\epsilon - \frac{1}{2})^2 + 2(\epsilon - \frac{1}{2})\overline{\Delta\varepsilon} - (\bar{\epsilon} - \frac{1}{2})^2 = \overline{\Delta\varepsilon}^2$ where $\bar{\epsilon} = \epsilon + \overline{\Delta\varepsilon}$. Since $\overline{\Delta\varepsilon}$ is proportional to $E_d\{\Delta h(X) + (j\omega/2)\,\Delta h^2(X)\}$ $(\sim 1/\mathcal{N})$ from (21), $\overline{\Delta\varepsilon}^2$ is proportional to $1/\mathcal{N}^2$ and can be neglected. Thus, only the fourth line remains uncancelled which is the second line of (22).

REFERENCES

[1] D. H. Foley, "Considerations of sample and feature size," *IEEE Trans. Inform. Theory*, vol. IT-18, no. 5, pp. 618–626, Sept. 1972.

[2] P. A. Lachenbruch and R. M. Mickey, "Estimation of error rates in discriminant analysis," *Technometrics*, vol. 10, no. 1, pp. 1–11, 1968.

[3] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Statist.*, vol. 7, pp. 1–26, 1979.

[4] L. M. Novak, "On the sensitivity of Bayes and Fisher classifiers in radar target detection," in *Proc. 18th Asilomar Conf. Circuits, Systems, and Computers*, Nov. 5–7, 1984.

[5] S. Raudys and V. Pikelis, "On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-2, no. 3, pp. 242–252, May 1980.

[6] G. T. Toussaint, "Bibliography on estimation of misclassification," *IEEE Trans. Inform. Theory*, vol. IT-20, no. 4, pp. 472–479, July 1974.

[7] D. J. Hand, "Recent advances in error rate estimation," *Pattern Recognition Lett.*, vol. 4, pp. 335–346, 1986.

[8] A. K. Jain and B. Chandrasekaran, "Dimensionality and sample size considerations in pattern recognition practice," *Handbook of Statistics*, vol. 2, P. R. Krishnaiah and L. N. Kanal, Eds. Amsterdam, The Netherlands: North-Holland, 1982, pp. 835–855.

[9] C. P. Han, "Distribution of discriminant function in circular models," *Inst. Statist., Math. Ann.*, vol. 22, no. 1, pp. 117–125, 1970.

[10] G. J. McLachlan, "Some expected values for the error rates of the sample quadratic discriminant function," *Australian J. Statist.*, vol. 17, no. 3, pp. 161–165, 1975.

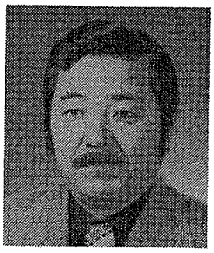
[11] S. John, "Errors in discrimination," *Ann. Math. Statist.*, vol. 32, pp. 1125–1144, 1961.

[12] K. Fukunaga and R. R. Hayes, "Effects of sample size in classifier design," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, no. 8, pp. 873–885, Aug. 1989.

[13] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic, 1972.

[14] S. Raudys, "Comparison of the estimates of the probability of misclassification," in *Proc. VIJCPR*, Kyoto, Japan, 1978, pp. 280–282.

[15] A. K. Jain, R. C. Dubes, and C.-C. Chen, "Bootstrap techniques for error estimation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-9, no. 5, pp. 628–633, Sept. 1987.



Keinosuke Fukunaga (M'66–SM'74–F'79) received the B.S. degree in electrical engineering from Kyoto University, Japan, in 1953, the M.S.E.E. degree from the University of Pennsylvania, Philadelphia, in 1959, and the Ph.D. degree from Kyoto University in 1962.

From 1953 to 1966 he was with the Mitsubishi Electric Company, Japan, first with the Central Research Laboratories working on computer applications in control systems, and then with the Computer Division where he was in charge of hardware development. Since 1966 he has been with Purdue University, West Lafayette, IN, where he is currently a Professor of Electrical Engineering. In the summers he has worked with a number of organizations. Also, he has served as a consultant to various government agencies and private companies.

Dr. Fukunaga was an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY for pattern recognition from 1977 to 1980. He is the author of *Introduction to Statistical Pattern Recognition*. He is a member of Eta Kappa Nu.

Raymond R. Hayes (S'86–M'87) received the B.S. degree in computer and electrical engineering under the Bell Labs Engineering Scholarship Program and the Ph.D. degree in electrical engineering under the IBM Resident Study Program, both from Purdue University, West Lafayette, IN, in 1984 and 1988, respectively.

He is now a Staff Member of the IBM Palo Alto Scientific Center. His current interests are in classification techniques and knowledge acquisition.

Dr. Hayes is a member of Tau Beta Pi, Eta Kappa Nu, AAAI, and the IEEE Computer Society.