Effects of Sample Size in Classifier Design

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

Abstract—**This paper discusses the effect of finite sample size on parameter estimates and their subsequent use in functions. General and parameter-specific expressions for the expected bias and variance of the functions are derived. These expressions are then applied to the Bhattacharyya distance and the analysis of the linear and quadratic classifiers, providing valuable insight into the relationship between the number of features and the number of training samples. Also, because of the functional form of the expressions, we present an empirical approach which will enable asymptotic performance to be accurately estimated using a very small number of samples. Results are experimentally verified using artificial data in controlled cases and using real, high-dimensional data.**

Index Terms—**Bhattacharyya distance, classifier degradation, classifier robustness, linear classifier, parameter estimation errors, quadratic classifier.**

I. Introduction

In practical pattern recognition problems, the parameters of the underlying distributions are unknown and the number of training samples available frequently is small. The size of this set of samples, relative to the number of features used, determines the accuracy of the parameter estimates and the similarity between the sample set and the true distribution.

In this paper, we will look at the effect of sample size on functions of the distributions' parameters. By viewing the estimated parameters as random variables, the expected value of a criterion can be computed by taking the expectation of the parameters over all possible N-size sets of training samples. This leads to a general expression for the expected bias and variance of the function, isolating the effects of functional form from the underlying distribution.

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 some performance criterion. A number of authors, including El-Sheikh and Wacker [1], have investigated the optimal number of features for a given finite design sample size in order to combat the "peaking phenomenon," the apparent loss of classifiability which accompanies an increase in the number of features without an increase in the number of training samples. An excellent review of work done in this area is presented in Jain and Chandrasekaran [2]. Another group of authors has looked into the effect of the ratio of training sample size to feature set dimensionality on the expected performance of an empirically designed classifier used on the true test distribution. In [3], Raudys and Pikelis catalog the development of a number of approximate expressions for the expected performance of the linear classifier and an exact expression for the quadratic classifier. Asymptotic expansions for the quadratic classifier have also been developed by Han [4] and McLachlan [5]. Unfortunately, these expressions are too complex to provide valuable insight and their accuracy has not been experimentally verified. Thus, the relationship between sample size and dimensionality has been inferred through simulation (e.g., [3] and [6]), the investigation of related criteria (e.g., [7] for Bhattacharyya distance and [8] for divergence), and a look at the performance of these classifiers tested on the design set [9].

By applying our general expression to the Bhattacharyya distance and the classifier error equation, we have developed a useful framework for the analysis of classifier performance, design, and testing procedures. This provides valuable insight into the relationship between dimensionality and sample size and the importance of mean and covariance shifts in measuring separability. Also, we have developed an empirical approach which will allow the designer to estimate the asymptotic performance of a particular type of classifier. This can be used to evaluate tradeoffs in classifier complexity and performance, and to determine the ratio of design samples to dimensionality needed for a particular performance level.

II. Bias and Variance Expressions

A. General Formulation

Let us consider the problem of estimating $f(y_1, \cdots, y_L)$ by $f(\hat{y}_1, \cdots, \hat{y}_L)$ where f is a given function, y_i's are the true parameter values and $\hat{y}_i$'s are their estimates. In this section, we will derive expressions for the expected value and variance of $f(\hat{y}_1, \cdots, \hat{y}_L)$, and propose a new method to estimate $f(y_1, \cdots, y_L)$.

Assuming that the deviation of $\hat{y}_i$ from y_i is small, $f(\hat{Y})$ can be expanded by a Taylor series up to the second-order terms as

$$\hat{f} \triangleq f(\hat{Y}) \cong f(Y) + \frac{\partial f^T}{\partial Y}\Delta Y + \frac{1}{2}\operatorname{tr}\left(\frac{\partial^2 f}{\partial Y^2}\Delta Y \Delta Y^T\right) \quad (1)$$

Manuscript received October 12, 1987; revised May 9, 1988. Recommended for acceptance by A. K. Jain. This work was supported in part by the National Science Foundation under Grants ECS-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 8928491.

where $Y = [y_1 \cdots y_L]^T$ and $\hat{Y} = [\hat{y}_1 \cdots \hat{y}_L]^T$ are the column vectors of the true parameters and their estimates, respectively. $\Delta Y = \hat{Y} - Y$, V^T indicates the transpose of the vector V, and tr A is the trace of a matrix A.

If the estimates are unbiased,

$$E\{\Delta Y\} = 0 \tag{2}$$

and subsequently

$$E\{\hat{f}\} \cong f + \frac{1}{2}\,\mathrm{tr}\left(\frac{\partial^2 f}{\partial Y^2} E\{\Delta Y \Delta Y^T\}\right). \tag{3}$$

Similarly, the variance of $\hat{f}$ can be derived as

$$\begin{aligned}\mathrm{Var}\{\hat{f}\} &\cong E\left\{\left[\frac{\partial f^T}{\partial Y}\Delta Y + \frac{1}{2}\,\mathrm{tr}\left(\frac{\partial^2 f}{\partial Y^2}\Delta Y \Delta Y^T\right) - \frac{1}{2}\,\mathrm{tr}\left(\frac{\partial^2 f}{\partial Y^2} E\{\Delta Y \Delta Y^T\}\right)\right]^2\right\} \\ &\cong E\left\{\left(\frac{\partial f^T}{\partial Y}\Delta Y\right)^2\right\} \\ &= \frac{\partial f^T}{\partial Y} E\{\Delta Y \Delta Y^T\} \frac{\partial f}{\partial Y}\end{aligned} \tag{4}$$

where the approximation from the first line to the second line was made by discarding terms higher than second-order.

Equation (3) shows that $\hat{f}$ is a biased estimate in general and that the bias depends on $\partial^2 f/\partial Y^2$ and $E\{\Delta Y \Delta Y^T\}$, where $\partial^2 f/\partial Y^2$ is determined by the functional form of f and $E\{\Delta Y \Delta Y^T\}$ is determined by $p(\hat{Y})$, the density function of $\hat{Y}$, and N, the number of samples used to compute $\hat{Y}$. Likewise, the variance depends on $\partial f/\partial Y$ and $E\{\Delta Y \Delta Y^T\}$.

For many estimators, the effects of $p(\hat{Y})$ and N on $E\{\Delta Y \Delta Y^T\}$ can be separated as

$$E\{\Delta Y \Delta Y^T\} = g(N)\, K(p(\hat{Y})) \tag{5}$$

where the scalar g and the matrix K are functions determined by how $\hat{Y}$ is computed. Substituting (5) into (3),

$$E\{\hat{f}\} \cong f + cg(N) \tag{6}$$

where $c = \frac{1}{2}\,\mathrm{tr}\,(\partial^2 f/\partial Y^2\, K(p(\hat{Y})))$ is independent of N and treated as a constant determined by a given underlying problem. This leads to the following procedure to estimate f:

1) Change the sample size N as $N_1, N_2, \cdots, N_l$. For each N_i, compute $\hat{Y}$ and subsequently $\hat{f}$ empirically. Repeat the experiment M times, and approximate $E\{\hat{f}\}$ with the sample mean of the M experimental results.

2) Plot these empirical points $E\{\hat{f}\}$ versus $g(N)$. Then, find the line best fitted to these points. The slope of this line is c and the y-intercept is the improved estimate of f. There are many possible ways of selecting a line. The standard procedure would be the minimum mean-square error approach.

B. Parametric Formulation

In pattern recognition, most of the expressions we would like to estimate are functions of the expected vectors and covariance matrices. In this section, we will show how the general discussion of the previous section can be applied to this particular family of parameters.

Assume that N samples are drawn from each of two n-dimensional Gaussian distributions with their expected vectors and covariance matrices given by

$$\begin{aligned} M_1 &= 0, \quad \Sigma_1 = I \\ M_2 &= M, \quad \Sigma_2 = \Lambda. \end{aligned} \tag{7}$$

Without loss of generality, any two covariance matrices can be simultaneously diagonalized to I and Λ, and a coordinate shift can bring the expected vector of one class to zero.

M_i and Σ_i can be estimated without bias by the sample mean and sample covariance

$$\begin{aligned} \hat{M}_i &= \frac{1}{N}\sum_{j=1}^{N} X_j^{(i)} \\ \hat{\Sigma}_i &= \frac{1}{N-1}\sum_{j=1}^{N} (X_j^{(i)} - \hat{M}_i)(X_j^{(i)} - \hat{M}_i)^T \end{aligned} \tag{8}$$

where $X_j^{(i)}$ is the jth sample vector from class i. Thus, the parameter vector $\hat{Y}$ of (1) consists of $2(n + n^2)$ components

$$\begin{aligned}\hat{Y} = [&\hat{m}_1^{(1)} \cdots \hat{m}_n^{(1)} \quad \hat{m}_1^{(2)} \cdots \hat{m}_n^{(2)} \quad \hat{\alpha}_{11}^{(1)} \cdots \hat{\alpha}_{nn}^{(1)} \\ &\hat{\alpha}_{11}^{(2)} \cdots \hat{\alpha}_{nn}^{(2)}]^T\end{aligned} \tag{9}$$

where $\hat{m}_i^{(r)}$ is the ith component of $\hat{M}_r$, and $\hat{\alpha}_{ij}^{(r)}$ is the ith row and jth column component of $\hat{\Sigma}_r$.

The random variables of (9) satisfy the following statistical properties, where $\Delta m_i^{(r)} = \hat{m}_i^{(r)} - m_i^{(r)}$ and $\Delta\alpha_{ij}^{(r)} = \hat{\alpha}_{ij}^{(r)} - \alpha_{ij}^{(r)}$:

1) The sample mean and covariance are unbiased:

$$E\{\Delta m_i^{(r)}\} = 0, \qquad E\{\Delta\alpha_{ij}^{(r)}\} = 0. \tag{10}$$

2) Samples from different classes are independent:

$$\begin{aligned} E\{\Delta m_i^{(1)}\Delta m_j^{(2)}\} &= E\{\Delta m_i^{(1)}\}\, E\{\Delta m_j^{(2)}\} = 0 \\ E\{\Delta\alpha_{ij}^{(1)}\Delta\alpha_{kl}^{(2)}\} &= E\{\Delta\alpha_{ij}^{(1)}\}\, E\{\Delta\alpha_{kl}^{(2)}\} = 0 \\ E\{\Delta m_i^{(r)}\Delta\alpha_{kl}^{(s)}\} &= E\{\Delta m_i^{(r)}\}\, E\{\Delta\alpha_{kl}^{(s)}\} \\ &= 0 \quad \text{for } r \neq s. \end{aligned} \tag{11}$$

3) Diagonal Σ_1 and Σ_2 cause the mean estimate covariances to be diagonal:

$$\begin{aligned} E\{\Delta m_i^{(r)}\Delta m_j^{(r)}\} &= 0, \quad \text{for } i \neq j \\ E\{\Delta m_i^{(1)2}\} &= \frac{1}{N} \\ E\{\Delta m_i^{(2)2}\} &= \frac{\lambda_i}{N} \end{aligned} \tag{12}$$

where λ_i is the ith diagonal component of Λ.

4) The third-order central moments of a Gaussian distribution are zero:

$$E\{\Delta m_i^{(r)} \Delta \alpha_{kl}^{(r)}\} = 0. \tag{13}$$

5) The fourth-order central moments of a Gaussian distribution are known:

$$E\{\Delta \alpha_{ij}^{(1)} \Delta \alpha_{kl}^{(1)}\}$$

$$= \begin{cases} \dfrac{1}{N} & \text{for } (i \neq j, i = k, j = l) \\ & \quad \text{or } (i \neq j, i = l, j = k) \\ \dfrac{2}{N-1} \cong \dfrac{2}{N} & \text{for } i = j = k = l \\ 0 & \text{otherwise} \end{cases}$$

$$E\{\Delta \alpha_{ij}^{(2)} \Delta \alpha_{kl}^{(2)}\}$$

$$= \begin{cases} \dfrac{\lambda_i \lambda_j}{N} & \text{for } (i \neq j, i = k, j = l) \\ & \quad \text{or } (i \neq j, i = l, j = k) \\ \dfrac{2\lambda_i^2}{N-1} \cong \dfrac{2\lambda_i^2}{N} & \text{for } i = j = k = l \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Note that in the equal index cases of (14) $N - 1$ is replaced by N for simplicity.

Substituting (9)–(14) into (3), the bias term of the estimate, $E\{\Delta f\} = E\{\hat{f}\} - f$, becomes

$$E\{\Delta f\} \cong \frac{1}{2} \operatorname{tr}\left(\frac{\partial^2 f}{\partial Y^2} E\{\Delta Y \Delta Y^T\}\right)$$

$$= \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \frac{\partial^2 f}{\partial y_i \partial y_j} E\{\Delta y_i \Delta y_j\}$$

$$= \frac{1}{2} \sum_{r=1}^{2} \left[\sum_{i=1}^{n} \frac{\partial^2 f}{\partial m_i^{(r)2}} E\{\Delta m_i^{(r)2}\} \right.$$

$$+ \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} \frac{\partial^2 f}{\partial \alpha_{ij}^{(r)} \partial \alpha_{ij}^{(r)}} E\{\Delta \alpha_{ij}^{(r)} \Delta \alpha_{ij}^{(r)}\}$$

$$+ \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} \frac{\partial^2 f}{\partial \alpha_{ij}^{(r)} \partial \alpha_{ji}^{(r)}} E\{\Delta \alpha_{ij}^{(r)} \Delta \alpha_{ji}^{(r)}\}$$

$$\left. + \sum_{i=1}^{n} \frac{\partial^2 f}{\partial \alpha_{ii}^{(r)2}} E\{\Delta \alpha_{ii}^{(r)2}\} \right]$$

$$\cong \frac{1}{2N} \left[\sum_{i=1}^{n} \frac{\partial^2 f}{\partial m_i^{(1)2}} + \sum_{i=1}^{n} \frac{\partial^2 f}{\partial m_i^{(2)2}} \lambda_i \right.$$

$$+ \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} \left(\frac{\partial^2 f}{\partial \alpha_{ij}^{(1)} \partial \alpha_{ij}^{(1)}} + \frac{\partial^2 f}{\partial \alpha_{ij}^{(1)} \partial \alpha_{ji}^{(1)}} \right)$$

$$+ \sum_{i=1}^{n} \frac{\partial^2 f}{\partial \alpha_{ii}^{(1)2}} 2$$

$$+ \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} \left(\frac{\partial^2 f}{\partial \alpha_{ij}^{(2)} \partial \alpha_{ij}^{(2)}} + \frac{\partial^2 f}{\partial \alpha_{ij}^{(2)} \partial \alpha_{ji}^{(2)}} \right) \lambda_i \lambda_j$$

$$\left. + \sum_{i=1}^{n} \frac{\partial^2 f}{\partial \alpha_{ii}^{(2)2}} 2\lambda_i^2 \right]. \tag{15}$$

Note that the effect of N is successfully separated, and that $g(N)$ of (5) becomes $1/N$. This is true for any functional form of f, provided f is a function of the expected vectors and covariance matrices of two Gaussian distributions. This conclusion can be extended to non-Gaussian cases in which (13) is satisfied and $E\{\Delta \alpha_{ij}^{(r)} \Delta \alpha_{kl}^{(r)}\}$ of (14) is proportional to $1/N$.

Similarly, the variance can be computed from (4), resulting in

$$\operatorname{Var}\{\hat{f}\} \cong \frac{1}{N} \left[\sum_{i=1}^{n} \left(\frac{\partial f}{\partial m_i^{(1)}} \right)^2 + \sum_{i=1}^{n} \left(\frac{\partial f}{\partial m_i^{(2)}} \right)^2 \lambda_i \right.$$

$$+ \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} \left\{ \left(\frac{\partial f}{\partial \alpha_{ij}^{(1)}} \right)^2 + \frac{\partial f}{\partial \alpha_{ij}^{(1)}} \frac{\partial f}{\partial \alpha_{ji}^{(1)}} \right\}$$

$$+ \sum_{i=1}^{n} \left(\frac{\partial f}{\partial \alpha_{ii}^{(1)}} \right)^2 2 + \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} \left\{ \left(\frac{\partial f}{\partial \alpha_{ij}^{(2)}} \right)^2 \right.$$

$$\left. \left. + \frac{\partial f}{\partial \alpha_{ij}^{(2)}} \frac{\partial f}{\partial \alpha_{ji}^{(2)}} \right\} \lambda_i \lambda_j + \sum_{i=1}^{n} \left(\frac{\partial f}{\partial \alpha_{ii}^{(2)}} \right)^2 2\lambda_i^2 \right]. \tag{16}$$

Note that, in order to calculate the bias and variance, we only need to compute $\partial f / \partial m_i^{(r)}$, $\partial f / \partial \alpha_{ij}^{(r)}$, $\partial^2 f / \partial m_i^{(r)2}$, $\partial^2 f / \partial \alpha_{ij}^{(r)} \partial \alpha_{ij}^{(r)}$ and $\partial^2 f / \partial \alpha_{ij}^{(r)} \partial \alpha_{ji}^{(r)}$ for $r = 1, 2$.

III. Bhattacharyya Distance Between Two Distributions

A popular measure of similarity between two distributions is the Bhattacharyya distance [10]

$$B = \frac{1}{8} (M_2 - M_1)^T \left(\frac{\Sigma_1 + \Sigma_2}{2} \right)^{-1} (M_2 - M_1)$$

$$+ \frac{1}{2} \ln \frac{\left| \dfrac{\Sigma_1 + \Sigma_2}{2} \right|}{\sqrt{|\Sigma_1|}\sqrt{|\Sigma_2|}} \tag{17}$$

Since B is a function of M_1, M_2, Σ_1 and Σ_2, it is a member of the family of functions discussed previously.

If two distributions are Gaussian, the Bhattacharyya distance gives an upper bound on the Bayes error, ϵ^*,

$$\epsilon^* \leqq \epsilon_u = \sqrt{P_1 P_2} e^{-B}. \tag{18}$$

where P_i is the *a priori* probability of class i. The first and second terms of (17), B_1 and B_2, measure the difference between the two distributions due to the mean and covariance shifts respectively.

When $\hat{M}_i$ and $\hat{\Sigma}_i$ of (8) are used to compute B, the resulting $\hat{B}$ differs from its true value. The bias and variance of $\hat{B}$ can be obtained using (15) and (16).

A. First Term B_1:

Since

$$B_1 = \frac{1}{8} \sum_{i=1}^{n} \frac{2}{1 + \lambda_i} (m_i^{(2)} - m_i^{(1)})^2,$$

$\partial B_1 / \partial m_i^{(r)}$ and $\partial^2 B_1 / \partial m_i^{(r)2}$

can be easily obtained as

$$\frac{\partial B_1}{\partial m_i^{(1)}} = -\frac{m_i^{(2)} - m_i^{(1)}}{2(1 + \lambda_i)}, \quad \frac{\partial B_1}{\partial m_i^{(2)}} = \frac{m_i^{(2)} - m_i^{(1)}}{2(1 + \lambda_i)} \tag{19}$$

$$\frac{\partial^2 B_1}{\partial m_i^{(1)2}} = \frac{\partial^2 B_1}{\partial m_i^{(2)2}} = \frac{1}{2(1 + \lambda_i)}. \tag{20}$$

The computation of $\partial B_1 / \partial \alpha_{ij}^{(r)}$, $\partial^2 B_1 / \partial \alpha_{ij}^{(r)} \partial \alpha_{ij}^{(r)}$ and $\partial^2 B_1 / \partial \alpha_{ij}^{(r)} \partial \alpha_{ji}^{(r)}$ are more complex and presented in Appendix 1. The results are

$$\frac{\partial B_1}{\partial \alpha_{ij}^{(1)}} = \frac{\partial B_1}{\partial \alpha_{ij}^{(2)}} = -\frac{m_i m_j}{4(1 + \lambda_i)(1 + \lambda_j)} \tag{21}$$

$$\frac{\partial^2 B_1}{\partial \alpha_{ij}^{(1)} \alpha_{ij}^{(1)}} = \frac{\partial^2 B_1}{\partial \alpha_{ij}^{(2)} \partial \alpha_{ij}^{(2)}} = 0 \quad \text{for } i \neq j \tag{22}$$

$$\frac{\partial^2 B_1}{\partial \alpha_{ij}^{(1)} \alpha_{ji}^{(1)}} = \frac{\partial^2 B_1}{\partial \alpha_{ij}^{(2)} \partial \alpha_{ji}^{(2)}} = \frac{1}{4(1 + \lambda_i)(1 + \lambda_j)} \left[\frac{m_i^2}{1 + \lambda_i} + \frac{m_j^2}{1 + \lambda_j} \right] \tag{23}$$

where $m_i = m_i^{(2)} - m_i^{(1)}$.

Substituting (19)–(23) into (15) and (16),

$$E\{\Delta B_1\} \cong \frac{1}{4N} \left[n + \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{m_j^2 (1 + \lambda_i \lambda_j)}{(1 + \lambda_j)^2 (1 + \lambda_i)} + \sum_{i=1}^{n} \frac{m_i^2 (1 + \lambda_i^2)}{(1 + \lambda_i)^3} \right] \tag{24}$$

$$\text{Var}\{\hat{B}_1\} \cong \frac{1}{4N} \left[\sum_{i=1}^{n} \frac{m_i^2}{1 + \lambda_i} + \sum_{i=1}^{n} \sum_{j=1}^{n} \cdot \frac{m_i^2 m_j^2 (1 + \lambda_i \lambda_j)}{2(1 + \lambda_i)^2 (1 + \lambda_j)^2} \right]. \tag{25}$$

B. Second Term B_2:

Similarly, the partial derivatives for B_2 are derived in Appendix 2. They are listed as follows:

$$\frac{\partial B_2}{\partial m_i^{(r)}} = 0 \quad \text{and} \quad \frac{\partial^2 B_2}{\partial m_i^{(r)2}} = 0 \quad \text{for } r = 1, 2 \tag{26}$$

$$\frac{\partial B_2}{\partial \alpha_{ij}^{(1)}} = \frac{\delta_{ij}}{2(1 + \lambda_i)} - \frac{\delta_{ij}}{4} \tag{27}$$

$$\frac{\partial B_2}{\partial \alpha_{ij}^{(2)}} = \frac{\delta_{ij}}{2(1 + \lambda_i)} - \frac{\delta_{ij}}{4\lambda_i} \tag{28}$$

$$\frac{\partial^2 B_2}{\partial \alpha_{ij}^{(1)} \partial \alpha_{ij}^{(1)}} = \frac{1}{4} - \frac{1}{2(1 + \lambda_i)(1 + \lambda_j)} \tag{29}$$

$$\frac{\partial^2 B_2}{\partial \alpha_{ij}^{(2)} \partial \alpha_{ij}^{(2)}} = \frac{1}{4\lambda_i \lambda_j} - \frac{1}{2(1 + \lambda_i)(1 + \lambda_j)} \tag{30}$$

$$\frac{\partial^2 B_2}{\partial \alpha_{ij}^{(1)} \partial \alpha_{ji}^{(1)}} = \frac{\partial^2 B_2}{\partial \alpha_{ij}^{(2)} \partial \alpha_{ji}^{(2)}} = 0 \quad \text{for } i \neq j. \tag{31}$$

Substituting (26)–(31) into (15) and (16),

$$E\{\Delta B_2\} \cong \frac{1}{4N} \left[n(n + 1) - \sum_{i=1}^{n} \sum_{j=1}^{n} \cdot \frac{1 + \lambda_i \lambda_j}{(1 + \lambda_i)(1 + \lambda_j)} - \sum_{i=1}^{n} \frac{1 + \lambda_i^2}{(1 + \lambda_i)^2} \right] \tag{32}$$

$$\text{Var}\{\hat{B}_2\} \cong \frac{1}{2N} \sum_{i=1}^{n} \left[\left(\frac{1}{1 + \lambda_i} - \frac{1}{2} \right)^2 + \left(\frac{1}{1 + \lambda_i} - \frac{1}{2\lambda_i} \right)^2 \lambda_i^2 \right]. \tag{33}$$

C. Discussions and Experimental Verification

Table I shows the dependence of $E\{\Delta B_1\}$ and $E\{\Delta B_2\}$ on n and k $(= N/n)$ for three different cases. In Case 1, samples from both classes are drawn from the same source $N(0, I)$, a Gaussian distribution with zero mean and identity covariance matrix. In Case 2, the two distributions share a covariance matrix but differ in the means. In Case 3, the means are the same, but the covariances are different. As Table I indicates, for all three cases, $E\{\Delta B_1\}$ is proportional to $1/k$ while $E\{\Delta B_2\}$ is proportional to $(n + 1)/k$. Also, note that $E\{\Delta B_1\}$ is the same for Cases 1 and 3 because the sources have the same mean. Similarly, $E\{\Delta B_2\}$ is the same for Cases 1 and 2 because the sources share a covariance matrix.

TABLE I
SAMPLE BIAS EXPRESSIONS FOR THE BHATTACHARYYA DISTANCE

	Case 1 N(0,I) N(0,I)	Case 2 N(0,I) N(M,I)	Case 3 N(0,I) N(0,4I)
		$m_1 = 2.56$	
m_i	$m_i = 0$	$m_i = 0$ $(i \neq 1)$	$m_i = 0$
λ_i	$\lambda_i = 1$	$\lambda_i = 1$	$\lambda_i = 4$
B_1	0	0.82	0
B_2	0	0	0.11 n
ϵ^*	50%	10%	Depends on n
$E\{\Delta B_1\}$	$\frac{0.25}{k}$	$\frac{0.35}{k}$	$\frac{0.25}{k}$
$E\{\Delta B_2\}$	$0.125\frac{n+1}{k}$	$0.125\frac{n+1}{k}$	$0.08\frac{n+1}{k}$

TABLE II
VALUES OF k AND N REQUIRED TO MAINTAIN $E\{\hat{B}\} \leqq 0.223$

n	4	8	16	32	64	128
k	3.9	6.2	10.7	19.6	39.6	73.4
N=nk	16	50	172	628	2407	9396

TABLE III
STATISTICS FOR CASE 4

i	1	2	3	4	5	6	7	8
λ_i	8.41	12.06	0.12	0.22	1.49	1.77	0.35	2.73
m_i	3.86	3.10	0.84	0.84	1.64	1.08	0.26	0.01

Since the trend is the same for all three cases, let us study Case 1 closely. Table I demonstrates that in high dimensional space ($n >> 1$) the distortion due to the covariance estimate ($E\{\Delta B_2\} = 0.125\ (n + 1)/k$) dominates that caused by the mean estimate ($E\{\Delta B_1\} = 0.25/k$). Also, since $E\{\Delta B_2\} = 0.125\ (n + 1)/k$, an increasingly large value of k is required to maintain a constant value of $E\{\hat{B}\}$ ($= E\{\hat{B}_1\} + E\{\hat{B}_2\}$) as the dimensionality increases. For example, Table II shows the value of k required to keep the value of $E\{\hat{B}\}$ less than 0.223. The true Bayes error for this case is 50 percent, and $E\{\hat{B}\} = 0.223$ gives an upper bound of 40 percent using (18). Only 16 samples (3.9 times the dimensionality) are needed to achieve $E\{\hat{B}\} = 0.223$ in a four-dimensional space, while 9396 samples (73.4 times the dimensionality) are needed in a 128-dimensional space. This result is sharply contrasted with the common belief that a fixed multiple of the dimensionality such as 5 or 10 could be used to determine the sample size.

Since the theoretical results of (24) and (32) for bias and (25) and (33) for variance are approximations, we have conducted three sets of experiments to verify these results. The first two cases are Cases 2 and 3 of Table I, while the third, which will be called Case 4, uses both mean and covariance differences. Case 4 uses an eight-dimensional Gaussian data set taken from [6] with a Bayes error of 1.9 percent, and λ_i's and m_i's listed in Table III.

For Cases 2 and 3, the dimensionality ranged from 4 to 64 in powers of 2, and k was selected as 3, 5, 10, 15, 20, 30, 40, and 50. $N(= nk)$ samples were generated from each class according to the given mean and covariance, and $\hat{B}_1$ and $\hat{B}_2$ were computed. This procedure was repeated 10 times independently, and the mean and standard deviation were computed. Tables IV, V, and VI present a comparison of the theoretical predictions (first lines) and the means of the 10 trials (second lines) for Cases 2, 3, and 4 respectively. These tables show that the theoretical predictions of the biases match the experimental results very closely.

The third lines of Tables IV, V, and VI present the standard deviations of the 10 trials. Table VII shows the theoretical predictions computed from (25) and (33) for $\hat{B}_1$ of Case 2 and $\hat{B}_2$ of Case 3. Again the theoretical predictions match the experimental results closely. It should be noted that the variances for $\hat{B}_2$ of Case 2 and $\hat{B}_1$ of Case 1 are zero theoretically. This suggests that the variances for these cases come from the Taylor expansion terms higher than second-order and therefore are expected to be smaller than the variances for the other cases. This is confirmed by comparing the variances between $\hat{B}_1$ and $\hat{B}_2$ of Cases 2 and 3. Also, note that the variances of $\hat{B}_2$ for Case 3 are independent of n.

In addition to the experimental verification, when $n = 1$, our theoretical predictions agree with those presented for univariate Gaussian densities in Jain [7]. Note that, because of the presence of cross-terms (e.g., $\lambda_i\lambda_j$), Jain's univariate expression cannot be applied to the multivariate case by summing the contributions of each feature even when these features are mutually independent.

TABLE IV
BIASES OF $\hat{B}$ FOR CASE 2: (a) $\hat{B}_1$ ($B_1 = 0.82$), (b) $\hat{B}_2$ ($B_2 = 0$)

(a)

	n	4	8	16	32	64
		1.1101	1.0758	1.0587	1.0502	1.0459
	3	1.0730	0.9933	1.0502	1.0754	1.0825
		0.4688	0.3791	0.2221	0.1551	0.0955
		0.9946	0.9740	0.9638	0.9586	0.9561
	5	1.0941	1.0702	1.0396	0.9659	0.9764
		0.3867	0.2745	0.1542	0.1091	0.0733
		0.9080	0.8977	0.8926	0.8900	0.8887
	10	0.9593	0.9277	0.8421	0.9128	0.8911
		0.2240	0.1424	0.1045	0.0720	0.0709
		0.8791	0.8723	0.8688	0.8671	0.8663
	15	0.8802	0.8705	0.8909	0.8634	0.8730
k		0.1634	0.1493	0.1053	0.0794	0.0493
		0.8647	0.8595	0.8570	0.8557	0.8551
	20	0.8778	0.8891	0.8261	0.8685	0.8361
		0.1356	0.1060	0.0929	0.0455	0.0387
		0.8502	0.8468	0.8451	0.8443	0.8438
	30	0.7901	0.8477	0.8583	0.8436	0.8373
		0.0702	0.0992	0.0712	0.0361	0.0366
		0.8430	0.8405	0.8392	0.8385	0.8382
	40	0.7917	0.8251	0.8578	0.8343	0.8444
		0.0786	0.1118	0.0522	0.0283	0.0271
		0.8387	0.8366	0.8356	0.8351	0.8348
	50	0.8524	0.8383	0.8364	0.8301	0.8290
		0.1060	0.0404	0.0515	0.0475	0.0287

(b)

	n	4	8	16	32	64
		0.2083	0.3750	0.7083	1.3750	2.7083
	3	0.2546	0.4106	0.8930	1.7150	3.2875
		0.0787	0.0653	0.0588	0.0776	0.1083
		0.1250	0.2250	0.4250	0.8250	1.6250
	5	0.1133	0.2791	0.5244	0.9252	1.8035
		0.0266	0.0785	0.0581	0.0302	0.0775
		0.0625	0.1125	0.2125	0.4125	0.8125
	10	0.0803	0.1179	0.2280	0.4365	0.8578
		0.0339	0.0191	0.0218	0.0279	0.0234
		0.0417	0.0750	0.1417	0.2750	0.5417
	15	0.0437	0.0742	0.1416	0.2894	0.5566
k		0.0243	0.0146	0.0143	0.0257	0.0170
		0.0313	0.0563	0.1063	0.2063	0.4063
	20	0.0389	0.0566	0.1079	0.2099	0.4129
		0.0101	0.0140	0.0132	0.0154	0.0058
		0.0208	0.0375	0.0708	0.1375	0.2708
	30	0.0190	0.0344	0.0707	0.1416	0.2777
		0.0063	0.0082	0.0097	0.0098	0.0062
		0.0156	0.0281	0.0531	0.1031	0.2031
	40	0.0170	0.0282	0.0561	0.1034	0.2061
		0.0072	0.0084	0.0086	0.0046	0.0063
		0.0125	0.0225	0.0425	0.0825	0.1625
	50	0.0102	0.0219	0.0417	0.0831	0.1650
		0.0037	0.0062	0.0041	0.0060	0.0057

(1st line: Theoretical prediction,
2nd line: The mean of 10 trials,
3rd line: The standard deviation of 10 trials)

TABLE V
BIASES OF $\hat{B}$ FOR CASE 3: (a) $\hat{B}_1 (B_1 = 0)$, (b) $\hat{B}_2 (B_2 = 0.11\,n)$

k	n = 4	8	16	32	64
3	0.0833	0.0833	0.0833	0.0833	0.0833
	0.1435	0.1212	0.1051	0.1118	0.1061
	0.0971	0.0633	0.0415	0.0385	0.0160
5	0.0500	0.0500	0.0500	0.0500	0.0500
	0.0489	0.0709	0.0579	0.0545	0.0605
	0.0284	0.0314	0.0141	0.0209	0.0071
10	0.0250	0.0250	0.0250	0.0250	0.0250
	0.0192	0.0267	0.0266	0.0276	0.0262
	0.0151	0.0124	0.0066	0.0079	0.0035
15	0.0167	0.0167	0.0167	0.0167	0.0167
	0.0159	0.0155	0.0207	0.0166	0.0181
	0.0078	0.0049	0.0106	0.0046	0.0036
20	0.0125	0.0125	0.0125	0.0125	0.0125
	0.0135	0.0156	0.0139	0.0120	0.0141
	0.0055	0.0071	0.0036	0.0038	0.0025
30	0.0083	0.0083	0.0083	0.0083	0.0083
	0.0050	0.0097	0.0085	0.0087	0.0085
	0.0037	0.0050	0.0030	0.0014	0.0013
40	0.0063	0.0063	0.0063	0.0063	0.0063
	0.0066	0.0082	0.0056	0.0062	0.0065
	0.0045	0.0050	0.0021	0.0014	0.0010
50	0.0050	0.0050	0.0050	0.0050	0.0050
	0.0042	0.0040	0.0054	0.0049	0.0052
	0.0037	0.0017	0.0015	0.0008	0.0009

(a)

k	n = 4	8	16	32	64
3	0.5796	1.1326	2.2385	4.4503	8.8739
	0.7129	1.0732	2.4527	4.7841	9.3263
	0.1447	0.1653	0.2332	0.1893	0.1642
5	0.5263	1.0366	2.0572	4.0983	8.1806
	0.5081	1.0063	2.1341	4.1041	8.4000
	0.1119	0.1546	0.1129	0.0868	0.1209
10	0.4863	0.9646	1.9212	3.8343	7.6606
	0.4901	0.9463	1.9345	3.8014	7.6630
	0.1016	0.0722	0.0759	0.0702	0.1206
15	0.4730	0.9406	1.8758	3.7463	7.4873
	0.5085	0.9675	1.9030	3.7952	7.5133
	0.0686	0.0350	0.0567	0.0306	0.0658
20	0.4663	0.9286	1.8532	3.7023	7.4006
	0.4708	0.9331	1.8277	3.7019	7.4049
	0.0658	0.0686	0.0966	0.0394	0.0672
30	0.4596	0.9166	1.8305	3.6583	7.3139
	0.4478	0.9033	1.8656	3.7053	7.3493
	0.0328	0.0646	0.0411	0.0884	0.0531
40	0.4473	0.9106	1.7886	3.5769	7.1536
	0.4713	0.8937	1.8058	3.6374	7.2596
	0.0444	0.0328	0.0353	0.0563	0.0392
50	0.4543	0.9070	1.8124	3.6231	7.2446
	0.4456	0.8872	1.8116	3.6279	7.2212
	0.0562	0.0506	0.0362	0.0449	0.0610

(b)

(1st line: Theoretical prediction,
2nd line: The mean of 10 trials,
3rd line: The standard deviation of 10 trials)

TABLE VI
BIASES OF $\hat{B}$ FOR CASE 4: (a) $\hat{B}_1$, (b) $\hat{B}_2$

k	Theoretical	Experimental Mean	Experimental St. Dev.
3	1.6453	1.5056	0.4995
5	1.4951	1.5104	0.1650
10	1.3824	1.3864	0.1997
15	1.3448	1.3365	0.1886
20	1.3261	1.3266	0.1712
30	1.3073	1.2884	0.1136
40	1.2979	1.3104	0.0658
50	1.2923	1.2997	0.0769

(a)

k	Theoretical	Experimental Mean	Experimental St. Dev.
3	1.4431	1.5695	0.2081
5	1.3002	1.2287	0.1446
10	1.1929	1.1638	0.0766
15	1.1572	1.1497	0.0523
20	1.1393	1.1255	0.0539
30	1.1214	1.1005	0.0337
40	1.1125	1.1093	0.0405
50	1.1071	1.1063	0.0276

(b)

D. Verification of the Proposed Estimation Procedure

The proposed estimation procedure following (6) was tested on a set of 66-dimensional, millimeter-wave radar data. The samples were collected by rotating a Camaro and a Dodge Van on a turntable and taking approximately 8800 readings. Sixty-six range bins were selected and the resulting 66 dimensional vectors were normalized by energy. The vectors were then selected at each half-degree to form 720-sample sets. The Bhattacharyya distance estimated from 720 samples, $\hat{B}_{720}$, was 2.29 which corresponds to an upper bound of the Bayes error of 5.1% (ϵ_u = 5.1 percent). These 720 samples per class were then divided into two sets of 360 samples. Since two sets were available from each class, there were 4 possible combinations of selecting one set from each class and forming a two-class problem. $\hat{B}$ was computed for each combination and the average of the 4 cases was taken. The resulting $\hat{B}_{360}$ was 3.27 (ϵ_u = 1.9 percent). Since $g(N)$ of (6) is $1/N$ for this case, two equations, $\hat{B}_{720} = 2.29 = B + c/720$ and $\hat{B}_{360} = 3.27 = B + c/360$, were set up and solved for B. Note that we replaced $E\{\hat{B}_{720}\}$ by $\hat{B}_{720}$ because Var $\{\hat{B}_{720}\}$ was expected to be small from the experimental results for Cases 2 and 3. The resulting B was 1.31 (ϵ_u = 13.5 percent). On the other hand, when all available 8800 samples per class were used, $\hat{B}_{8800}$ was 1.51 (ϵ_u = 11.0 percent).

Although the radar data is not guaranteed to be Gaussian, the above results indicate that the prediction of the true B from a relatively small number of samples (720 per class for the 66-dimensional space) seems possible. Also, note that $\hat{B}_{360}$, $\hat{B}_{720}$, and $\hat{B}_{8800}$ are significantly different. Without the proposed compensation, $\hat{B}_{360}$ and $\hat{B}_{720}$ could not provide a useful upper bound of the Bayes error.

TABLE VII
PREDICTED STANDARD DEVIATIONS

k	$\hat{B}_1$ for Case 2, n = 4	8	16	32	64	$\hat{B}_2$ for Case 3, for all n
3	0.3531	0.2497	0.1765	0.1248	0.0883	0.1732
5	0.2735	0.1934	0.1368	0.0967	0.0684	0.1342
10	0.1934	0.1368	0.0967	0.0684	0.0483	0.0949
15	0.1579	0.1117	0.0790	0.0558	0.0395	0.0775
20	0.1368	0.0967	0.0684	0.0483	0.0342	0.0671
30	0.1117	0.0790	0.0558	0.0395	0.0279	0.0548
40	0.0967	0.0684	0.0483	0.0342	0.0242	0.0474
50	0.0865	0.0612	0.0432	0.0306	0.0216	0.0424

IV. CLASSIFIER DEGRADATION

An even more important measurement in pattern recognition is the expected performance of a given classifier. The discriminant functions for some of the popular classifiers, including the linear and quadratic classifiers, are functions of M_1, M_2, Σ_1, and Σ_2. Thus, they are the members of the family of functions presented in Section II. However, unlike the Bhattacharyya distance, the degradation of the expected classifier performance due to a finite sample size comes from two sources: the finite sample set used for design and the finite number of test samples. Thus, we need to study their effects separately.

A. Effect of Test Sample Size

When the design and test samples are independent, the effect of test sample size is well-understood. Let us assume that a classifier is given and ϵ_i (i = 1, 2) is the true

probability of error from class $i(\omega_i)$. In order to estimate ϵ_i, N_i samples from ω_i are drawn and tested by the given classifier and the number of misclassified samples $\hat{\tau}_i$ is counted. The random variables $\hat{\tau}_1$ and $\hat{\tau}_2$ are independent and each is binomially distributed as [10]:

$$\Pr\{\hat{\tau}_1 = \tau_1, \hat{\tau}_2 = \tau_2\} = \prod_{i=1}^{2} \Pr\{\hat{\tau}_1 = \tau_i\} = \prod_{i=1}^{2} \binom{N_i}{\tau_i} \epsilon_i^{\tau_i}(1-\epsilon_i)^{N_i-\tau_i} \tag{34}$$

ϵ_i is estimated by $\hat{\tau}_i/N_i$ and subsequently, the total probability of error is estimated by

$$\hat{\epsilon} = \sum_{i=1}^{2} P_i \frac{\hat{\tau}_i}{N_i} \tag{35}$$

where P_i is the *a priori* probability of ω_i. The expected value and variance are known:

$$E\{\hat{\epsilon}\} = \epsilon \tag{36}$$

$$\text{Var}\{\hat{\epsilon}\} = \sum_{i=1}^{2} P_i^2 \frac{\epsilon_i(1-\epsilon_i)}{N_i}. \tag{37}$$

B. Expression of ϵ

The effect of design sample size is much harder to analyze. In order to discuss this subject, we need to express the probability of error ϵ in terms of the classifier. Let us assume that the classifier is defined as

$$h(X) \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 0. \tag{38}$$

The characteristic function of h for ω_i is

$$\phi_i(\omega) = E\{e^{j\omega h(X)} | \omega_i\} = \int_S e^{j\omega h(X)} p_i(X)\, dX \tag{39}$$

where S indicates the entire n-dimensional space and $p_i(X)$ is the density function of X for ω_i. Since the characteristic function of h is the Fourier transform of the density function of h (except for the sign of $j\omega$), the density function of h for ω_i, $q_i(X)$, can be obtained by the inverse Fourier transform as

$$q_i(h) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \phi_i(\omega) e^{-j\omega h}\, d\omega. \tag{40}$$

The probabilities of error for ω_1 and ω_2 are

$$\epsilon_1 = \int_0^{\infty} q_1(h)\, dh = 1 - \int_{-\infty}^{0} q_1(h)\, dh \tag{41}$$

$$\epsilon_2 = \int_{-\infty}^{0} q_2(h)\, dh. \tag{42}$$

According to Fourier transform theory, the integration in the h-space can be converted to multiplication by $1/j\omega$ in the ω-space. That is,

$$g_i(t) = \int_{-\infty}^{t} q_i(h)\, dh = \frac{\phi_i(0)}{2} - \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{\phi_i(\omega)}{j\omega} e^{-j\omega t}\, d\omega. \tag{43}$$

Inserting $g_i(0)$ into (41) and (42), and realizing that (39) guarantees $\phi_i(0) = 1$,

$$\begin{aligned}\epsilon &= P_1\epsilon_1 + P_2\epsilon_2 \\ &= \frac{1}{2} + P_1 \frac{1}{2\pi} \int_{-\infty}^{+\infty} \frac{\phi_1(\omega)}{j\omega} d\omega - P_2 \frac{1}{2\pi} \cdot \int_{-\infty}^{+\infty} \frac{\phi_2(\omega)}{j\omega} d\omega \\ &= \frac{1}{2} + \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_S \frac{e^{j\omega h(X)}}{j\omega} \cdot [P_1 p_1(X) - P_2 p_2(X)]\, dX\, d\omega.\end{aligned} \tag{44}$$

When the design sample size is finite, the parameters Y of the distributions are estimated and the discriminant function is based on these estimated parameters $\hat{Y}$. That is, $\hat{h}(X) = h(X, \hat{Y})$ is a random variable shifted from $h(X, Y)$. Taking the expectation with respect to $\hat{Y}$,

$$E\{\hat{\epsilon}\} = \frac{1}{2} + \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_S \frac{E\{e^{j\omega\hat{h}(X)}\}}{j\omega} \cdot [P_1 p_1(X) - P_2 p_2(X)]\, dX\, d\omega. \tag{45}$$

Treating $e^{j\omega\hat{h}(X)}$ as $\hat{f}$ in (3)

$$\begin{aligned}E\{e^{j\omega\hat{h}(X)}\} &\cong e^{j\omega h(X)} + \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \frac{\partial^2 e^{j\omega h(X)}}{\partial y_i \partial y_j} E\{\Delta y_i \Delta y_j\} \\ &= e^{j\omega h(X)} + \frac{j\omega}{2} e^{j\omega h(X)} \sum_{i=1}^{L} \sum_{j=1}^{L} \cdot \left[\frac{\partial^2 h(X)}{\partial y_i \partial y_j} + j\omega \frac{\partial h(X)}{\partial y_i} \frac{\partial h(X)}{\partial y_j}\right] E\{\Delta y_i \Delta y_j\}.\end{aligned} \tag{46}$$

Substituting (46) into (45) and realizing $E\{\Delta\epsilon\} = E\{\hat{\epsilon}\} - \epsilon$,

$$\begin{aligned}E\{\Delta\epsilon\} \cong \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_S \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \left[\frac{\partial^2 h(X)}{\partial y_i \partial y_j} + j\omega \frac{\partial h(X)}{\partial y_i} \frac{\partial h(X)}{\partial y_j}\right] E\{\Delta y_i \Delta y_j\} \\ \times e^{j\omega h(X)} [P_1 p_1(X) - P_2 p_2(X)]\, dX\, d\omega.\end{aligned} \tag{47}$$

Equation (47) is a very general expression for $E\{\Delta\epsilon\}$ which is valid regardless of the selection of $h(X)$, P_i and $p_i(X)$. The term $E\{\Delta y_i \Delta y_j\}$ gives the effect of the sam-

ple size N. Therefore, if (5) is satisfied, $E\{\Delta\epsilon\}$ can be expressed by $cg(N)$ where c is determined by $h(X)$, P_i, and $p_i(X)$, and the proposed estimation procedure following (6) can be applied. Furthermore, if $h(X)$ is a function of M_i and Σ_i, $g(N)$ becomes $1/N$.

C. The Quadratic Classifier for Gaussian Distributions

When a quadratic classifier is designed from N training samples, drawn from two simultaneously diagonalized distributions, $N(0, I)$ and $N(M, \Lambda)$, with *a priori* probabilities $P_1 = P_2 = 0.5$, the discriminant function can be found as

$$\hat{h}(X) = \tfrac{1}{2}(X - \hat{M}_1)^T\hat{\Sigma}_1^{-1}(X - \hat{M}_1) - \tfrac{1}{2}(X - \hat{M}_2)^T\hat{\Sigma}_2^{-1}(X - \hat{M}_2) + \tfrac{1}{2}\ln|\hat{\Sigma}_1| - \tfrac{1}{2}\ln|\hat{\Sigma}_2| \quad (48)$$

where $\hat{M}_i$ and $\hat{\Sigma}_i$ are estimated by (8). Forming $\hat{Y}$ as in (9), we only need to compute $\partial h/\partial m_i^{(r)}$, $\partial h/\partial \alpha_{ij}^{(r)}$, $\partial^2 h/\partial m_i^{(r)2}$, $\partial^2 h/\partial \alpha_{ij}^{(r)}\partial \alpha_{ij}^{(r)}$ and $\partial^2 h/\partial \alpha_{ij}^{(r)}\partial \alpha_{ji}^{(r)}$, since $E\{\Delta y_i \Delta y_j\} = 0$ for other combinations.

With M_i and Σ_i given in (7), these partial derivatives can be easily computed and are listed in Appendix 3. Substituting these results into (47),

$$\begin{aligned}\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L}[\cdot]E\{\Delta y_i \Delta y_j\} &\cong \frac{1}{2N}\left[(n+1)\sum_{i=1}^{n}\left\{x_i^2 - \frac{(x_i - m_i)^2}{\lambda_i}\right\}\right. \\ &\quad + j\omega\left[n + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\right. \\ &\quad \left.\left.\cdot\left\{x_i^2 x_j^2 + \frac{(x_i - m_i)^2(x_j - m_j)^2}{\lambda_i\lambda_j}\right\}\right]\right] \\ &\triangleq \frac{1}{N} f_q(X, \omega). \end{aligned} \quad (49)$$

Thus, (47) may be rewritten as

$$E\{\hat{\epsilon}\} \cong \epsilon + \frac{c_q}{N} \quad (50)$$

where

$$c_q = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\int_S f_q(X, \omega)e^{j\omega h(X)} \cdot [P_1 p_1(X) - P_2 p_2(X)]\, dX\, d\omega. \quad (51)$$

That is, c_q is determined by the underlying distributions, and stays constant for experiments with various sample sizes. Thus, as was proposed in Section II, we may choose various values of N as $N_1, \cdots, N_l$, and measure $\hat{\epsilon}$. Computing $E\{\hat{\epsilon}\}$ from several independent trials, we may solve (50) for ϵ and c_q by a line fit technique.

The above technique was applied to the radar data. The entire 8800-sample set was divided into two groups, each consisting of 4400 samples. When one group was used to design a quadratic classifier and the other was used for testing, the error $\hat{\epsilon}_{4400}$ was 17.2 percent. Then, 720 samples were selected from the design group and used to design a quadratic classifier. The entire 4400 samples of the test group were tested, resulting in $\hat{\epsilon}_{720} = 21.4$ percent. Such a large number of test samples was used to eliminate the variation of $\hat{\epsilon}$ due to test sample size. The same experiment was performed for 360 samples. Since there were two groups of 360 samples from 720 samples for each class, four error estimates were obtained; they were averaged, resulting in $\hat{\epsilon}_{360} = 25.4$ percent. $\hat{\epsilon}_{720}$ and the averaged $\hat{\epsilon}_{360}$ were used to obtain ϵ by solving (50), resulting in $\epsilon = 17.4$ percent. This result is very close to $\hat{\epsilon}_{4400} = 17.2$ percent, and confirms that we can predict the potential performance of the quadratic classifier even if the available sample size is relatively small for a high dimensional space.

Although we do not need to know the value of c_q to conduct the above experiment to estimate ϵ, c_q can be computed by carrying through the integration of (51). Let us consider the simplest case, Case 2 of Table I, in which $p_1(X)$ and $p_2(X)$ are Gaussian $N(0, I)$ and $N(M, I)$, respectively. Then, $e^{j\omega h}(X)\, p_i(X)$ may be rewritten as

$$e^{j\omega h(X)}p_1(X) = \frac{\sqrt{2\pi}}{\sqrt{\beta}}e^{-\beta/8}N_\omega\left(-\frac{j}{2}, \frac{1}{\beta}\right)N_X(j\omega M, I) \quad (52)$$

$$e^{j\omega h(X)}p_2(X) = \frac{\sqrt{2\pi}}{\sqrt{\beta}}e^{-\beta/8}N_\omega\left(\frac{j}{2}, \frac{1}{\beta}\right) \cdot N_X((1 + j\omega)M, I) \quad (53)$$

where $\beta = M^TM$. $N_\omega(a, b)$ and $N_X(D, K)$ are Gaussian density functions of ω and X with the expected value of a and variance b for N_ω, and the expected vector D and covariance matrix K for N_X.

Since $f_q(\omega, X)$ is a linear combination of $x_i^a x_j^b (a, b \leqq 4)$ as is seen in (49), $\int f_q(X, \omega)\, N_X(\cdot, \cdot)\, dX$ is the linear combination of the moments of $N_X(\cdot, \cdot)$. The result of the integration becomes a polynomial in ω

$$\gamma_i(\omega) = \frac{\beta^2}{2}(j\omega)^5 \mp \beta^2(j\omega)^4 + \frac{\beta}{2}(n + 5 + 3\beta)(j\omega)^3 \mp \frac{\beta}{2}(n + 5 + 2\beta)(j\omega)^2 + \frac{1}{4}[n(n + 7) + (5n + 9)\beta + \beta^2](j\omega) \mp \frac{(n + 1)\beta}{2} \quad (54)$$

where $-$ and $+$ of $\mp$ are for $i = 1$ and 2, respectively. Again, the $\int \gamma_i(\omega)\, N_\omega(\cdot, \cdot)\, d\omega$ is a linear combination of the moments of $N_\omega(\cdot, \cdot)$. Thus, c_q for $P_1 = P_2 = 0.5$ is

$$c_q = \frac{1}{4\sqrt{2\pi\beta}}\, e^{-\beta/8} \cdot \left[n^2 + \left(1 + \frac{\beta}{2}\right)n + \left(\frac{\beta^2}{16} - \frac{\beta}{2} - 1\right)\right]. \quad (55)$$

$E\{\hat{\epsilon}\}$ can be predicted theoretically by $\epsilon + c_q/N$. Table VIII lists the theoretical predictions for various N and $k\ (= N/n)$ for the distribution parameters given in Case 2 of Table I. These parameters yield $\beta = 2.56^2$ and $\epsilon = 0.1$ (10 percent). Also shown in Table VIII are experimental results verifying these predictions. For each combination of n and k, N samples were generated from each class and used to design a quadratic classifier which was then tested on true distributions. Novak developed an algorithm which numerically computes the error of any discriminant function with a quadratic form tested on two Gaussian distributions [11]. This procedure was repeated 10 times. The second and third lines in Table VIII show the means and standard deviations of the experimental results. The theoretical prediction accurately reflects the experimental trends. Also, the standard deviations are small. Notice that as n gets larger, k must increase to maintain the same performance, since c_q is proportional to n^2 for $n \gg 1$. This conclusion agrees with Pipberger's experimental results [6] and the numerical tables in Raudys and Pikelis [3]. Together, these present design guidelines significantly different from the traditional rules of thumb which suggest a particular fixed value of k for all values of n.

TABLE VIII
QUADRATIC CLASSIFIER DEGRADATION FOR CASE 2

	n = 4	8	16	32	64
	0.1450	0.1689	0.2115	0.3067	0.4894
3	0.1668	0.2041	0.2204	0.2673	0.3131
	0.0351	0.0235	0.0289	0.0195	0.0133
	0.1270	0.1414	0.1691	0.2240	0.3336
5	0.1403	0.1640	0.1734	0.2081	0.2554
	0.0211	0.0186	0.0091	0.0057	0.0074
	0.1135	0.1207	0.1345	0.1620	0.2168
10	0.1152	0.1240	0.1366	0.1573	0.1934
	0.0081	0.0061	0.0070	0.0054	0.0085
	0.1090	0.1138	0.1230	0.1413	0.1779
15	0.1086	0.1184	0.1232	0.1415	0.1658
k	0.0044	0.0061	0.0042	0.0053	0.0042
	0.1067	0.1103	0.1173	0.1310	0.1584
20	0.1077	0.1105	0.1190	0.1393	0.1513
	0.0021	0.0023	0.0051	0.0022	0.0032
	0.1045	0.1069	0.1115	0.1207	0.1389
30	0.1054	0.1071	0.1114	0.1307	0.1365
	0.0019	0.0021	0.0020	0.0019	0.0022
	0.1034	0.1052	0.1086	0.1155	0.1292
40	0.1037	0.1057	0.1087	0.1150	0.1275
	0.0024	0.0013	0.0013	0.0013	0.0018
	0.1027	0.1041	0.1069	0.1124	0.1234
50	0.1025	0.1044	0.1068	0.1125	0.1221
	0.0013	0.0010	0.0013	0.0009	0.0007

(1st line: Theoretical prediction, 2nd line: The mean of 10 trials, 3rd line: The standard deviation of 10 trials)

D. The Linear Classifier for Gaussian Distributions

The analysis of the linear classifier proceeds in a similar fashion. Fisher's linear discriminant function is

$$h(X) = (M_2 - M_1)^T \bar{\Sigma}^{-1} X + \tfrac{1}{2}(M_1^T \bar{\Sigma}^{-1} M_1 - M_2^T \bar{\Sigma}^{-1} M_2) \quad (56)$$

where $\bar{\Sigma} = (\Sigma_1 + \Sigma_2)/2$. Again, we assume, without loss of generality, that $M_1 = 0$, $M_2 = M$, $\Sigma_1 = I$ and $\Sigma_2 = \Lambda$.

The partial derivatives of h can be computed as is shown in Appendix 4. Thus, (47) becomes

$$E\{\hat{\epsilon}\} \cong \epsilon + \frac{c_l}{N} \quad (57)$$

$$c_l = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_S f_l(X, \omega)\, e^{j\omega h(X)} \cdot [P_1 p_1(X) - P_2 p_2(X)]\, dX\, d\omega \quad (58)$$

$$f_l(X, \omega) = \sum_{i=1}^{n} \left[\frac{1 - \lambda_i}{1 + \lambda_i} + (2x_i - m_i)\left\{\frac{(1 + \lambda_i^2)m_i}{(1 + \lambda_i)^3} + \frac{m_i}{(1 + \lambda_i)^2} \sum_{j=1}^{n} \frac{1 + \lambda_i\lambda_j}{1 + \lambda_j}\right\}\right] + \frac{j\omega}{2}\left[4 \sum_{i=1}^{n}\left\{\frac{x_i^2}{(1 + \lambda_i)^2} + \frac{(x_i - m_i)^2\lambda_i}{(1 + \lambda_i)^2}\right\} + \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{m_i(2x_j - m_j)(1 + \lambda_i\lambda_j)}{(1 + \lambda_i)^2(1 + \lambda_j)^2} \cdot \{m_i(2x_j - m_j) + m_j(2x_i - m_i)\}\right]. \quad (59)$$

Again, c_l is determined by the underlying distributions, and ϵ can be estimated from experiments with various N. Also, since $f_l(X, \omega)$ is a linear combination of $x_i^a (a \leqq 2)$, c_l can be theoretically computed for Case 2 of Table I, resulting in

$$c_l = \frac{1}{2\sqrt{2\pi\beta}}\, e^{-\beta/8}\left[\left(1 + \frac{\beta}{4}\right)n - 1\right]. \quad (60)$$

TABLE IX
LINEAR CLASSIFIER DEGRADATION FOR CASE 2

		n			
	4	8	16	32	64
	0.1273	0.1287	0.1294	0.1298	0.1300
3	0.1437	0.1436	0.1336	0.1302	0.1319
	0.0365	0.0174	0.0135	0.0081	0.0040
	0.1164	0.1172	0.1177	0.1179	0.1180
5	0.1165	0.1223	0.1207	0.1199	0.1207
	0.0128	0.0153	0.0071	0.0048	0.0041
	0.1082	0.1086	0.1088	0.1089	0.1090
10	0.1050	0.1089	0.1093	0.1086	0.1092
	0.0030	0.0041	0.0024	0.0021	0.0019
	0.1055	0.1057	0.1059	0.1060	0.1060
15	0.1048	0.1080	0.1064	0.1058	0.1064
k	0.0030	0.0032	0.0027	0.0013	0.0012
	0.1041	0.1043	0.1044	0.1045	0.1045
20	0.1039	0.1039	0.1058	0.1040	0.1045
	0.0021	0.0018	0.0026	0.0011	0.0008
	0.1027	0.1029	0.1029	0.1030	0.1030
30	0.1036	0.1033	0.1027	0.1033	0.1028
	0.0023	0.0021	0.0009	0.0006	0.0006
	0.1020	0.1022	0.1022	0.1022	0.1022
40	0.1022	0.1027	0.1021	0.1023	0.1022
	0.0021	0.0014	0.0009	0.0005	0.0004
	0.1016	0.1017	0.1018	0.1018	0.1018
50	0.1016	0.1021	0.1018	0.1018	0.1017
	0.0011	0.0007	0.0005	0.0004	0.0003

(1st line: Theoretical prediction, 2nd line: The mean of 10 trials, 3rd line: The standard deviation of 10 trials)

Equation (60) was experimentally verified in the same manner as (55). The results are shown in Table IX.

Comparison of (55) and (60) reveals an important distinction between quadratic and linear classifiers. For Case 2, the two covariances are the same. Thus, if the true underlying parameters are used, the quadratic classifier of (48) becomes identical to the linear classifier of (56). However, when the estimated covariances are used, $\hat{\Sigma}_1 \neq \hat{\Sigma}_2$ even though $\Sigma_1 = \Sigma_2$. Thus, the classifier of (48) differs from that of (56). As a result, $E\{\Delta\epsilon\}$ for quadratic is proportional to $n^2/N\ (= n/k)$ while $E\{\Delta\epsilon\}$ for linear is proportional to $n/N\ (= 1/k)$ as in (55) and (60) when $n >> 1$. Although it depends on the values of n and β, we may generally conclude that c_q is larger than c_l for $n >> 1$. This implies that many more samples are needed to properly design a quadratic classifier than a linear classifier. Novak reported in [11] that the linear classifier is more robust (less sensitive to parameter estimation errors) than the quadratic classifier, particularly in high dimensional spaces. Our results support his claim both theoretically and experimentally.

Also note that for large n, c_l/N is proportional to $1/k$. This indicates that, as far as the design of a linear classifier is concerned, a fixed multiple could be used to determine the sample size from the dimensionality. This coincides with the conclusions of many reports in the past. However, (60) indicates that the value of the multiple depends on β, which measures the separability between two distributions with a common covariance matrix.

V. CONCLUSIONS

The main purpose of this paper was to investigate the effect of finite sample size parameter estimates on the evaluation of a family of functions. To this end, we have presented general expressions for the expected bias and variance in terms of the statistical properties of the parameter estimates.

Applying these expressions to the Bhattacharyya distance has provided insight into the relationship between the dimensionality and the number of training samples and their effect on measuring separability due to mean and covariance shifts. Applying them to classifier evaluation equations, we have derived explicit expressions for the degradation of the quadratic and linear classifiers. This provides a new guideline for the selection of the number of samples or features necessary for a certain level of classifier performance. We have provided theoretical evidence that, as the dimensionality increases, covariance-based similarity measures and the quadratic classifier require an increasing multiple of samples. We have also presented support for the claim that the linear classifier is more robust.

Finally, the form of the bias expression allows the dependence on the sample size to be separated from the distribution-specific terms. Since the distribution and dimension are fixed for a given sample set, an empirical approach was employed to use estimates of expected performance for different sized samples to find an estimate of the asymptotic performance. This allows relatively small sample sets to provide accurate, unbiased estimates.

APPENDIX 1
COMPUTATION OF THE DERIVATIVES OF B_1

In order to compute the derivatives of B_1, we need the following formula for matrix differentiation [12].

$$\frac{\partial A^{-1}}{\partial a_{ij}} = -A^{-1}\frac{\partial A}{\partial a_{ij}}A^{-1} = -A^{-1}I(i,j)A^{-1} \quad \text{(A1)}$$

where a_{ij} is the i, j component of a matrix A, and $I(i, j)$ is a matrix with an i, j component of 1 and all other components equal to 0. The s, t component of (A1) is

$$\left[\frac{\partial A^{-1}}{\partial a_{ij}}\right]_{st} = -[A^{-1}]_{si}[A^{-1}]_{jt}. \quad \text{(A2)}$$

Applying (A1) repeatedly,

$$\begin{aligned}\frac{\partial^2 A^{-1}}{\partial a_{ij}\partial a_{kl}} &= A^{-1}\frac{\partial A}{\partial a_{kl}}A^{-1}I(i,j)A^{-1} \\ &\quad + A^{-1}I(i,j)A^{-1}\frac{\partial A}{\partial a_{kl}}A^{-1} \\ &= A^{-1}I(k,l)\,A^{-1}I(i,j)A^{-1} \\ &\quad + A^{-1}I(i,j)\,A^{-1}I(k,l)A^{-1} \quad \text{(A3)}\end{aligned}$$

and

$$\left[\frac{\partial^2 A^{-1}}{\partial a_{ij}\partial a_{kl}}\right]_{st} = [A^{-1}]_{sk}[A^{-1}]_{li}[A^{-1}]_{jt} + [A^{-1}]_{si}[A^{-1}]_{jk}[A^{-1}]_{lt}. \quad \text{(A4)}$$

In the computation of the derivatives of B_1 with respect to $\alpha_{ij}^{(r)}$, let $A = \bar{\Sigma} = (\Sigma_1 + \Sigma_2)/2$ and $M = M_2 - M_1$ from (7).

From (17), (A1), and (A2)

$$\begin{aligned}\frac{\partial B_1}{\partial \alpha_{ij}^{(r)}} &= -\frac{1}{8} M^T \bar{\Sigma}^{-1}\left(\frac{1}{2}\frac{\partial \Sigma_r}{\partial \alpha_{ij}^{(r)}}\right)\bar{\Sigma}^{-1}M \\ &= -\frac{1}{16} M^T \bar{\Sigma}^{-1} I(i,j)\bar{\Sigma}^{-1}M \\ &= -\frac{1}{16}\sum_{s=1}^{n}\sum_{t=1}^{n}[\bar{\Sigma}^{-1}]_{si}[\bar{\Sigma}^{-1}]_{jt}m_s m_t \\ &= -\frac{1}{16}\sum_{s=1}^{n}\sum_{t=1}^{n}\frac{2\delta_{si}}{1+\lambda_i}\frac{2\delta_{jt}}{1+\lambda_j}m_s m_t \\ &= -\frac{m_i m_j}{4(1+\lambda_i)(1+\lambda_j)} \end{aligned} \quad \text{(A5)}$$

where $\delta_{ij} = 0$ or 1 depending on $i \neq j$ or $i = j$ and m_i is the ith component of M.

Also from (17), (A3), and (A4)

$$\begin{aligned}\frac{\partial^2 B_1}{\partial \alpha_{ij}^{(r)}\partial \alpha_{ij}^{(r)}} &= \frac{1}{8} M^T\left[\bar{\Sigma}^{-1}\frac{I(i,j)}{2}\bar{\Sigma}^{-1}\frac{I(i,j)}{2}\bar{\Sigma}^{-1} + \bar{\Sigma}^{-1}\frac{I(i,j)}{2}\bar{\Sigma}^{-1}\frac{I(i,j)}{2}\bar{\Sigma}^{-1}\right]M \\ &= \frac{1}{16}\sum_{s=1}^{n}\sum_{t=1}^{n}[\bar{\Sigma}^{-1}]_{si}[\bar{\Sigma}^{-1}]_{ji}[\bar{\Sigma}^{-1}]_{jt}m_s m_t \\ &= \frac{1}{16}\sum_{s=1}^{n}\sum_{t=1}^{n}\frac{2\delta_{si}}{1+\lambda_i}\frac{2\delta_{ji}}{1+\lambda_i}\frac{2\delta_{jt}}{1+\lambda_j}m_s m_t \\ &= 0 \quad \text{for } i \neq j. \end{aligned} \quad \text{(A6)}$$

Likewise,

$$\begin{aligned}\frac{\partial^2 B_1}{\partial \alpha_{ij}^{(r)}\partial \alpha_{ji}^{(r)}} &= \frac{1}{8} M^T\left[\bar{\Sigma}^{-1}\frac{I(i,j)}{2}\bar{\Sigma}^{-1}\frac{I(j,i)}{2}\bar{\Sigma}^{-1} + \bar{\Sigma}^{-1}\frac{I(j,i)}{2}\bar{\Sigma}^{-1}\frac{I(i,j)}{2}\bar{\Sigma}^{-1}\right]M \\ &= \frac{1}{32}\sum_{s=1}^{n}\sum_{t=1}^{n}\left[\frac{2\delta_{si}}{1+\lambda_i}\frac{2\delta_{jj}}{1+\lambda_j}\frac{2\delta_{it}}{1+\lambda_i}m_s m_t + \frac{2\delta_{sj}}{1+\lambda_j}\frac{2\delta_{ii}}{1+\lambda_i}\frac{2\delta_{jt}}{1+\lambda_j}m_s m_t\right] \\ &= \frac{1}{4}\left[\frac{m_i^2}{(1+\lambda_i)^2(1+\lambda_j)} + \frac{m_j^2}{(1+\lambda_j)^2(1+\lambda_i)}\right]. \end{aligned} \quad \text{(A7)}$$

Equations (A5), (A6), and (A7) are shown in (21), (22), and (23) respectively.

Appendix 2
Computation of the Derivatives of B_2

From [12], if a matrix A is symmetric

$$\frac{\partial \ln |A|}{\partial A} = A^{-1^T} = A^{-1} \quad \text{(A8)}$$

or

$$\frac{\partial \ln |A|}{\partial a_{ij}} = [A^{-1}]_{ij}. \quad \text{(A9)}$$

Using (A1),

$$\frac{\partial^2 \ln |A|}{\partial A \partial a_{kl}} = \frac{\partial A^{-1}}{\partial a_{kl}} = -A^{-1}I(k,l)A^{-1} \quad \text{(A10)}$$

or

$$\frac{\partial^2 \ln |A|}{\partial a_{ij}\partial a_{kl}} = -[A^{-1}I(k,l)A^{-1}]_{ij} = -[A^{-1}]_{ik}[A^{-1}]_{lj}. \quad \text{(A11)}$$

Since $B_2 = \frac{1}{2}\ln|\bar{\Sigma}| - \frac{1}{4}\ln|\Sigma_1| - \frac{1}{4}|\Sigma_2|$ from (17) and $\bar{\Sigma} = (\Sigma_1 + \Sigma_2)/2$,

$$\frac{\partial B_2}{\partial \alpha_{ij}^{(1)}} = \frac{1}{2}\frac{1}{2}[\bar{\Sigma}^{-1}]_{ij} - \frac{1}{4}[\Sigma_1^{-1}]_{ij} = \frac{1}{4}\frac{2\delta_{ij}}{1+\lambda_i} - \frac{1}{4}\delta_{ij} \quad \text{(A12)}$$

$$\frac{\partial B_2}{\partial \alpha_{ij}^{(2)}} = \frac{1}{2}\frac{1}{2}[\bar{\Sigma}^{-1}]_{ij} - \frac{1}{4}[\Sigma_2^{-1}]_{ij} = \frac{1}{4}\frac{2\delta_{ij}}{1+\lambda_i} - \frac{1}{4}\frac{\delta_{ij}}{\lambda_i}. \quad \text{(A13)}$$

Equations (A12) and (A13) are shown in (27) and (28), respectively.

The second order derivatives of B_2 are obtained by using (A11):

$$\frac{\partial^2 B_2}{\partial \alpha_{ij}^{(r)}\partial \alpha_{ij}^{(r)}} = -\frac{1}{2}\frac{1}{4}[\bar{\Sigma}^{-1}]_{ii}[\bar{\Sigma}^{-1}]_{jj} + \frac{1}{4}[\Sigma_r^{-1}]_{ii}[\Sigma_r^{-1}]_{jj} \quad \text{(A14)}$$

$$\frac{\partial^2 B_2}{\partial \alpha_{ij}^{(r)}\partial \alpha_{ji}^{(r)}} = -\frac{1}{2}\frac{1}{4}[\bar{\Sigma}^{-1}]_{ij}[\bar{\Sigma}^{-1}]_{ji} + \frac{1}{4}[\Sigma_r^{-1}]_{ij}[\Sigma_r^{-1}]_{ji}. \quad \text{(A15)}$$

Therefore,

$$\frac{\partial^2 B_2}{\partial \alpha_{ij}^{(1)}\partial \alpha_{ij}^{(1)}} = -\frac{1}{8}\frac{2}{1+\lambda_i}\frac{2}{1+\lambda_j} + \frac{1}{4} \quad \text{(A16)}$$

$$\frac{\partial^2 B_2}{\partial \alpha_{ij}^{(2)}\partial \alpha_{ij}^{(2)}} = -\frac{1}{8}\frac{2}{1+\lambda_i}\frac{2}{1+\lambda_j} + \frac{1}{4}\frac{1}{\lambda_i}\frac{1}{\lambda_j} \quad \text{(A17)}$$

$$\frac{\partial^2 B_2}{\partial \alpha_{ij}^{(r)} \partial \alpha_{ji}^{(r)}} = 0 \qquad \text{for } i \neq j. \tag{A18}$$

Equations (A16)–(A18) are shown in (29)–(31).

Appendix 3
The Derivatives of h for the Quadratic Classifier

The derivatives of h with respect to $m_i^{(r)}$ can be obtained easily from (48) as follows:

$$\frac{\partial h(X)}{\partial M_1} = -\Sigma_1^{-1}(X - M_1), \quad \frac{\partial h(X)}{\partial M_2} = -\Sigma_2^{-1}(X - M_2) \tag{A19}$$

$$\frac{\partial^2 h(X)}{\partial M_1^2} = \Sigma_1^{-1}, \quad \frac{\partial^2 h(X)}{\partial M_2^2} = -\Sigma_2^{-1}. \tag{A20}$$

Using $M_1 = 0$, $M_2 = M$, $\Sigma_1 = I$, and $\Sigma_2 = \Lambda$ of (7),

$$\frac{\partial h(X)}{\partial m_i^{(1)}} = -x_i, \quad \frac{\partial h(X)}{\partial m_i^{(2)}} = \frac{x_i - m_i}{\lambda_i} \tag{A21}$$

$$\frac{\partial^2 h(X)}{\partial m_i^{(1)2}} = 1, \quad \frac{\partial^2 h(X)}{\partial m_i^{(2)2}} = -\frac{1}{\lambda_i}. \tag{A22}$$

In order to derive the derivatives with respect to $\alpha_{ij}^{(r)}$, we need the derivatives for matrix inversion as in Appendix 1 and the derivatives of the log-determinant as in Appendix 2. They can be computed as follows:

$$\begin{aligned}\frac{\partial h(X)}{\partial \alpha_{ij}^{(1)}} &= -\frac{1}{2}(X - M_1)^T \Sigma_1^{-1} I(i,j)\, \Sigma_1^{-1}(X - M_1) \\ &\quad + \frac{1}{2}[\Sigma_1^{-1}]_{ij} \\ &= -\frac{1}{2} x_i x_j + \frac{1}{2}\delta_{ij}\end{aligned} \tag{A23}$$

$$\begin{aligned}\frac{\partial h(X)}{\partial \alpha_{ij}^{(2)}} &= \frac{1}{2}(X - M_2)^T \Sigma_2^{-1} I(i,j)\, \Sigma_2^{-1}(X - M_2) \\ &\quad - \frac{1}{2}[\Sigma_2^{-1}]_{ij} \\ &= \frac{1}{2}\frac{(x_i - m_i)(x_j - m_j)}{\lambda_i \lambda_j} - \frac{1}{2}\frac{\delta_{ij}}{\lambda_i}\end{aligned} \tag{A24}$$

$$\begin{aligned}\frac{\partial^2 h(X)}{\partial \alpha_{ij}^{(1)} \partial \alpha_{ij}^{(1)}} &= \frac{1}{2}(X - M_1)^T [\Sigma_1^{-1} I(i,j)\, \Sigma_1^{-1} I(i,j) \Sigma_1^{-1} \\ &\quad + \Sigma_1^{-1} I(i,j)\, \Sigma_1^{-1} I(i,j) \Sigma_1^{-1}](X - M_1) \\ &\quad - \frac{1}{2}[\Sigma_1^{-1}]_{ii}[\Sigma_1^{-1}]_{jj} \\ &= \sum_{s=1}^{n}\sum_{t=1}^{n} \delta_{si}\delta_{ji}\delta_{jt} x_s x_t - \frac{1}{2}\delta_{ii}\delta_{jj} \\ &= -\frac{1}{2} \qquad \text{for } i \neq j\end{aligned} \tag{A25}$$

$$\begin{aligned}\frac{\partial^2 (X)}{\partial \alpha_{ij}^{(2)} \partial \alpha_{ij}^{(2)}} &= -\frac{1}{2}(X - M_2)^T [\Sigma_2^{-1} I(i,j)\, \Sigma_2^{-1} I(i,j) \Sigma_2^{-1} \\ &\quad + \Sigma_2^{-1} I(i,j)\, \Sigma_2^{-1} I(i,j) \Sigma_2^{-1}](X - M_2) \\ &\quad - \frac{1}{2}[\Sigma_2^{-1}]_{ii}[\Sigma_2^{-1}]_{jj} \\ &= \frac{1}{2}\frac{1}{\lambda_i \lambda_j} \qquad \text{for } i \neq j\end{aligned} \tag{A26}$$

$$\begin{aligned}\frac{\partial^2 h(X)}{\partial \alpha_{ij}^{(1)} \partial \alpha_{ji}^{(1)}} &= \frac{1}{2}(X - M_1)^T [\Sigma_1^{-1} I(i,j)\, \Sigma_1^{-1} I(j,i) \Sigma_1^{-1} \\ &\quad + \Sigma_1^{-1} I(j,i)\, \Sigma_1^{-1} I(i,j) \Sigma_1^{-1}](X - M_1) \\ &\quad - \frac{1}{2}[\Sigma_1^{-1}]_{ij}[\Sigma_1^{-1}]_{ji} \\ &= \frac{1}{2}\sum_{s=1}^{n}\sum_{t=1}^{n}(\delta_{si}\delta_{jj}\delta_{it}x_s x_t + \delta_{sj}\delta_{ii}\delta_{jt}x_s x_t) \\ &\quad - \frac{1}{2}\delta_{ij}\delta_{ji} \\ &= \frac{1}{2}(x_i^2 + x_j^2) - \frac{1}{2}\delta_{ij}\end{aligned} \tag{A27}$$

$$\begin{aligned}\frac{\partial^2 h(X)}{\partial \alpha_{ij}^{(2)} \partial \alpha_{ji}^{(2)}} &= -\frac{1}{2}(X - M_2)^T [\Sigma_2^{-1} I(i,j)\, \Sigma_2^{-1} I(j,i) \Sigma_2^{-1} \\ &\quad + \Sigma_2^{-1} I(j,i)\, \Sigma_2^{-1} I(i,j) \Sigma_2^{-1}](X - M_2) \\ &\quad - \frac{1}{2}[\Sigma_2^{-1}]_{ij}[\Sigma_2^{-1}]_{ji} \\ &= -\frac{1}{2}\sum_{s=1}^{n}\sum_{t=1}^{n}\left(\frac{\delta_{si}}{\lambda_i}\frac{\delta_{jj}}{\lambda_j}\frac{\delta_{it}}{\lambda_i}(x_s - m_s)(x_t - m_t)\right. \\ &\quad \left. + \frac{\delta_{sj}}{\lambda_j}\frac{\delta_{ii}}{\lambda_i}\frac{\delta_{jt}}{\lambda_j}(x_s - m_s)(x_t - m_t)\right) + \frac{1}{2}\frac{\delta_{ij}}{\lambda_i}\frac{\delta_{ji}}{\lambda_j} \\ &= -\frac{1}{2}\left(\frac{(x_i - m_i)^2}{\lambda_i^2 \lambda_j} + \frac{(x_j - m_j)^2}{\lambda_j^2 \lambda_i}\right) + \frac{1}{2}\frac{\delta_{ij}}{\lambda_i \lambda_j}.\end{aligned} \tag{A28}$$

Plugging all these equations into (47), we obtain (49).

Appendix 4
The Derivatives of h for the Linear Classifier

The derivatives of h with respect to $m_i^{(r)}$ can be obtained from (56) as follows:

$$\frac{\partial h(X)}{\partial M_1} = -\bar{\Sigma}^{-1}X + \bar{\Sigma}^{-1}M_1 \tag{A29}$$

$$\frac{\partial h(X)}{\partial M_2} = \bar{\Sigma}^{-1}X - \bar{\Sigma}^{-1}M_2 \tag{A30}$$

$$\frac{\partial^2 h(X)}{\partial M_1^2} = \bar{\Sigma}^{-1} \tag{A31}$$

$$\frac{\partial^2 h(X)}{\partial M_2^2} = -\bar{\Sigma}^{-1}. \tag{A32}$$

Using $M_1 = 0$, $M_2 = M$, $\Sigma_1 = I$, $\Sigma_2 = \Lambda$, and $\bar{\Sigma} = (I + \Lambda)/2$,

$$\frac{\partial h(X)}{\partial m_i^{(1)}} = -\frac{2x_i}{1+\lambda_i}, \quad \frac{\partial h(X)}{\partial m_i^{(2)}} = \frac{2(x_i - m_i)}{1+\lambda_i} \tag{A33}$$

$$\frac{\partial^2 h(X)}{\partial m_i^{(1)2}} = \frac{2}{1+\lambda_i}, \quad \frac{\partial^2 h(X)}{\partial m_i^{(2)2}} = -\frac{2}{1+\lambda_i}. \tag{A34}$$

The derivatives with respect to $\alpha_{ij}^{(r)}$ are computed as follows:

$$\begin{aligned}\frac{\partial h}{\partial \alpha_{ij}^{(r)}} &= -\frac{1}{4} M^T \bar{\Sigma}^{-1} I(i,j)\, \bar{\Sigma}^{-1}(2X - M) \\ &= -\frac{1}{4}\sum_{s=1}^{n}\sum_{t=1}^{n} \frac{2\delta_{si}}{1+\lambda_i}\frac{2\delta_{jt}}{1+\lambda_j} m_s(2x_t - m_t) \\ &= -\frac{m_i(2x_j - m_j)}{(1+\lambda_i)(1+\lambda_j)}\end{aligned} \tag{A35}$$

$$\begin{aligned}\frac{\partial^2 h}{\partial \alpha_{ij}^{(r)} \partial \alpha_{ij}^{(r)}} &= \frac{1}{8} M^T \big[\bar{\Sigma}^{-1} I(i,j)\, \bar{\Sigma}^{-1} I(i,j) \bar{\Sigma}^{-1} \\ &\quad + \bar{\Sigma}^{-1} I(i,j)\, \bar{\Sigma}^{-1} I(i,j) \bar{\Sigma}^{-1}\big](2X - M) \\ &= \frac{1}{4}\sum_{s=1}^{n}\sum_{t=1}^{n} \frac{2\delta_{si}}{1+\lambda_i}\frac{2\delta_{ji}}{1+\lambda_j}\frac{2\delta_{jt}}{1+\lambda_j} \\ &\quad \cdot m_s(2x_t - m_t) \\ &= 0 \qquad \text{for } i \neq j\end{aligned} \tag{A36}$$

$$\begin{aligned}\frac{\partial^2 h}{\partial \alpha_{ij}^{(r)} \partial \alpha_{ji}^{(r)}} &= \frac{1}{8} M^T \big[\bar{\Sigma}^{-1} I(i,j)\, \bar{\Sigma}^{-1} I(j,i) \bar{\Sigma}^{-1} \\ &\quad + \bar{\Sigma}^{-1} I(j,i)\, \bar{\Sigma}^{-1} I(i,j) \bar{\Sigma}^{-1}\big](2X - M) \\ &= \frac{1}{8}\sum_{s=1}^{n}\sum_{t=1}^{n} \left\{ \frac{2\delta_{si}}{1+\lambda_i}\frac{2\delta_{jj}}{1+\lambda_j}\frac{2\delta_{it}}{1+\lambda_i} \right. \\ &\quad \left. + \frac{2\delta_{sj}}{1+\lambda_j}\frac{2\delta_{ii}}{1+\lambda_i}\frac{2\delta_{jt}}{1+\lambda_j} \right\} m_s(2x_t - m_t) \\ &= \frac{m_i(2x_i - m_i)}{(1+\lambda_i)^2(1+\lambda_j)} + \frac{m_j(2x_j - m_j)}{(1+\lambda_j)^2(1+\lambda_i)}.\end{aligned} \tag{A37}$$

Plugging all these results into (47), we obtain (59).

Acknowledgment

The authors wish to thank Dr. L. M. Novak of MIT Lincoln Laboratory for the acquisition of the radar data used for verification and for the use of his program to compute the expected performance of a classifier given a test distribution.

References

[1] T. S. El-Sheikh and A. G. Wacker, "Effect of dimensionality and estimation on the performance of Gaussian classifiers," *Pattern Recognition*, vol. 12, pp. 115–126, 1980.

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

[3] 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.

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

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

[6] H. V. Pipberger, "Computer analysis of electrocardiogram," in *Clinical Electrocardiography and Computers*, C. A. Caceres and L. S. Dreifus, Eds. New York: Academic, 1970, pp. 109–119.

[7] A. K. Jain, "On an estimate of the Bhattacharyya distance," *IEEE Trans. Syst., Man, Cybern.*, pp. 763–766, Nov. 1976.

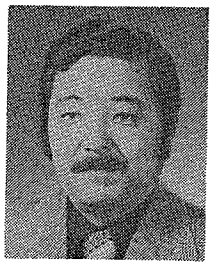
[8] H. M. Kalayeh and D. A. Landgrebe, "Predicting the required number of training samples," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, no. 6, pp. 664–667, Nov. 1983.

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

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

[11] L. 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.

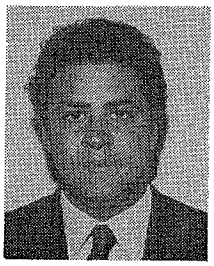
[12] W. Beyer, *CRC Standard Mathematical Tables*, 26th ed. Boca Raton, FL: CRC Press, 1981, pp. 44–45.



Keinosuke Fukunaga (M'66–SM'74–F'79) received the B.S. degree in electrical engineering from Kyoto University, Kyoto, 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.