Journal of Classification 11:101-120 (1994)

Error Rates in Quadratic Discrimination with Constraints on the Covariance Matrices

Bernhard W. Flury

Indiana University

Martin J. Schmid

Bundesamt für Sozialversicherung, Switzerland

A. Narayanan

The Procter & Gamble Company

The authors wish to thank the editor and three referees for their helpful comments on the first draft of this article. M. J. Schmid supported by grants no. 2.724-0.85 and 2.038-0.86 of the Swiss National Science Foundation.

Authors' Addresses: Bernhard W. Flury, Department of Mathematics, Indiana University, Bloomington, IN 47405, USA; Martin J. Schmid, Bundesamt für Sozialversicherung, 3000 Bern, Switzerland; and A. Narayanan, The Procter & Gamble Company, Ivorydale Technical Center, Cincinnati, Ohio 45217, USA. Address for correspondence: first author.

Abstract: In multivariate discrimination of several normal populations, the optimal classification procedure is based on quadratic discriminant functions. We compare expected error rates of the quadratic classification procedure if the covariance matrices are estimated under the following four models: (i) arbitrary covariance matrices, (ii) common principal components, (iii) proportional covariance matrices, and (iv) identical covariance matrices. Using Monte Carlo simulation to estimate expected error rates, we study the performance of the four discrimination procedures for five different parameter setups corresponding to "standard" situations that have been used in the literature. The procedures are examined for sample sizes ranging from 10 to 60, and for two to four groups. Our results quantify the extent to which a parsimonious method reduces error rates, and demonstrate that choosing a simple method of discrimination is often beneficial even if the underlying model assumptions are wrong.

Keywords: Common principal components; Linear Discriminant Function; Monte Carlo Simulation; Proportional Covariance Matrices

1. Introduction

Suppose the p-variate random vector $\mathbf{X}$ is measured in k populations, with density function $f_i(\mathbf{x})$ and prior probability π_i in the i-th group ($i = 1, \ldots, k$). Discriminant analysis is concerned with finding a partition $\mathbb{R}^p = R_1 \cup R_2 \cup \cdots \cup R_k$ such that an observation $\mathbf{x}$ with unknown group membership is classified into the i-th population if $\mathbf{x} \in R_i$. If we choose the partition such as to minimize the total probability of misclassification, each region of classification R_j contains all points $\mathbf{x} \in R^p$ such that $\pi_j f_j(\mathbf{x})$ is the maximum of all $\pi_i f_i(\mathbf{x})$, $i = 1, \ldots, k$; see Anderson (1984, Theorem 6.7.1). For k multivariate normal populations with mean vectors μ_i and nonsingular covariance matrices ψ_i, the classification region R_j is

$$R_j = \{\mathbf{x} \in \mathbb{R}^p : q_j(\mathbf{x}) \geq q_i(\mathbf{x}) \; \forall \; i = 1, \ldots, k\}$$

where the classification functions $q_i(\mathbf{x})$ are quadratic and given by

$$q_i(\mathbf{x}) = \mathbf{x}' \mathbf{A}_i \mathbf{x} + \mathbf{b}'_i \mathbf{x} + c_i \tag{1.1}$$

Here,

$$\begin{aligned} \mathbf{A}_i &= -\frac{1}{2} \psi_i^{-1} \\ \mathbf{b}_i &= \psi_i^{-1} \mu_i \\ c_i &= \log \pi_i - \frac{1}{2} \log \det(\psi_i) - \frac{1}{2} \mu'_i \psi_i^{-1} \mu_i \end{aligned} \tag{1.2}$$

In practical applications the parameters μ_i and ψ_i are estimated from training samples, and all parameters in (1.1) and (1.2) are then replaced by some

estimates $\hat{\mu}_i$ and $\hat{\psi}_i$, which yields estimated classification regions $\hat{R}_i$. Since these estimates are subject to sampling error, the resulting classification rule is no longer optimal, i.e., the total probability of misclassification based on $\hat{R}_1, \ldots, \hat{R}_k$ is not minimal. Evidently the performance of the classification procedure must depend on the precision of the estimates.

Whenever a large number of parameters is to be estimated, the variability of the estimates can be reduced by imposing valid constraints on the parameter space, be it by equating certain parameters, or by setting parameters equal to known (hypothetical) values. In normal theory discriminant analysis the classical assumption is that all covariance matrices ψ_i are identical, which means that the quadratic classification functions (1.1) become linear. It is well known (see Seber 1984, pp. 299-300) that linear discrimination, whenever appropriate, outperforms quadratic discrimination with no constraints imposed on the covariance matrices. However, this is not the whole story: even in situations where linear discrimination is theoretically wrong, it may outperform the theoretically correct quadratic method in terms of expected error rates. This happens typically for small samples, and indicates that the gain in precision due to reducing the number of parameters is more important than the bias introduced by imposing theoretically wrong constraints.

The performance of linear and quadratic discrimination rules has typically been assessed by simulation studies (e.g., Marks and Dunn 1974, Wahl and Kronmal 1977). More recently, asymptotic expansions have been used to obtain approximate analytical results. O'Neill (1992a) gives an expansion for the mean vector of the linear discriminant function coefficients when in fact the covariance matrices are not equal, and studies it in greater detail in the special case of proportional covariance matrices. The same author (O'Neill 1992b) studies asymptotic error rates of linear and quadratic discrimination rules and gives numerical results for the parameter configuration we call "O'Neill's model" in section 5. In particular, O'Neill gives approximate minimal sample sizes needed for quadratic discrimination to outperform linear discrimination, and finds that the linear discrimination rule is "quite robust to departures from the equal variances assumption (1992b, p. 177)," a result that our simulations confirm. A related study by Wakaki (1990) gives approximate error rates for linear and quadratic discrimination rules when the covariance matrices are in fact proportional.

What all the studies mentioned fail to do is suggest better methods of discrimination, i.e., methods that avoid both the overparameterization of the usual quadratic rule and the oversimplification of the linear rule. The present article is an attempt to fill the gap between linear and quadratic discrimination by imposing constraints (other than equality) on the covariance matrices. Related approaches are discussed in section 7.

2. Four Methods of Discrimination

We will now define four methods of discrimination, or rather, of estimation of the covariance matrices ψ_i. The estimates will then be substituted in (1.1) and (1.2) to obtain estimated classification regions. All four methods are based on samples of size N_i from the k populations, $\min_{1 \le i \le k} N_i > p$. The mean vectors μ_i will always be estimated by the sample mean vectors $\bar{\mathbf{x}}_i$, and the prior probabilities will be assumed to be known.

Method 1: Ordinary quadratic discrimination. The covariance matrices ψ_i are estimated by $\mathbf{S}_i$, the sample covariance matrices in their usual unbiased form.

Method 2: Common principal components. If the common principal component (CPC) model holds, all ψ_i have identical eigenvectors, i.e., there exists an orthogonal $p \times p$ matrix β such that $\psi_i = \beta \Lambda_i \beta'$, $i = 1, \ldots, k$, where $\Lambda_i = \text{diag}\,(\lambda_{i1}, \ldots, \lambda_{ip})$ (see Flury 1988, chapter 4). The classification procedure obtained by replacing the ψ_i in (1.2) by their maximum likelihood estimates under the CPC model will be referred to as CPC discrimination.

Method 3: Proportional covariance matrices. This method is based on the assumption that all ψ_i are proportional (with unknown proportionality factors). Replacing the ψ_i in (1.2) by their maximum likelihood estimates under proportionality (see Flury 1988, chapter 5) yields a classification procedure which we shall call proportional discrimination.

Method 4: Linear discrimination. Assuming equality of all covariance matrices ψ_i, the common covariance matrix ψ is estimated by the pooled sample covariance matrix.

We will sometimes refer to the four methods as (1) DIFF, (2) CPC, (3) PROP, and (4) EQU. The following table shows the number of functionally independent parameters to be estimated for covariance matrices under each of the four methods, for dimension p and number of groups k:

Method	Number of parameters
(1) DIFF	$kp(p-1)/2 + kp$
(2) CPC	$p(p-1)/2 + kp$
(3) PROP	$p(p-1)/2 + p + k - 1$
(4) EQU	$p(p-1)/2 + p$

For Methods 1 and 4, the number of parameters is evident. In Method 3, proportional discrimination, $k-1$ proportionality factors are estimated. In

Method 2, there are $p(p-1)/2$ parameters for the single orthogonal matrix, plus p variances (i.e., eigenvalues) for each of the k groups. This indicates that using more parsimonious methods of discrimination than DIFF may be particularly beneficial if either p or k is large. Method 3, proportional discrimination, appears particularly attractive because only a single parameter is added for each additional group. We will later see that proportional discrimination is indeed a useful method.

3. Some Theoretical Background

Due to the fact that finding exact or approximate expressions for expected error rates for the intermediate procedures 2 and 3 appears mathematically intractable, this paper uses stochastic simulation methods. However, some limited theoretical results will be helpful to identify situations in which a given method might be particularly successful. Asymptotic results comparing Methods 1 and 4 have been obtained, in a special parameter setup that we will discuss in section 5, by O'Neill (1984). Since it is assumed that "good estimation" will lead to small error rates, Flury and Schmid (1992) studied asymptotic variances of discriminant function coefficients under the same four methods as in the current paper, and for $k = 2$ groups. We give here only a very short summary of a few results that seem most relevant for the choice of parameter configurations to be studied; for a more thorough discussion see Flury and Schmid (1992, pp. 251-260).

(i) Assume the CPC model holds, then Methods 1 and 2 are theoretically correct. The asymptotic variances indicate that Method 2 does not necessarily yield discriminant function coefficients with smaller asymptotic variances than Method 1, depending on the eigenvalues (i.e., the diagonal elements of the Λ_i - matrices). For instance, if $\lambda_{1h} - \lambda_{1j} = \lambda_{2j} - \lambda_{2h}$ for all (j,h), then CPC discrimination and ordinary quadratic discrimination should do about equally well. On the other hand, if $\lambda_{ih}^{-1} - \lambda_{1j}^{-1} = \lambda_{2h}^{-1} - \lambda_{2j}^{-1}$ for all (j,h), then some of the quadratic coefficients have smaller asymptotic variances if estimated by Method 2.

(ii) If the ψ_i are proportional, then Methods 1 to 3 are theoretically correct. The asymptotic results indicate considerable potential advantages of proportional discrimination over both CPC - and ordinary quadratic discrimination, particularly if the dimension p is large.

(iii) If $\psi_1 = \psi_2$, then linear discrimination is obviously best, and the advantage of using Method 4 over 1 or 2 increases with the dimension p. Interestingly, the variances of discriminant function

coefficients estimated under Method 3 approach those obtained in linear discrimination when p increases. In other words, for large dimension p we may expect proportional discrimination to do almost as well as linear discrimination.

Evidently the asymptotic results leave some vital questions open, namely:

(a) what is the advantage of using a (correct) parsimonious method for small sample sizes?
(b) What are the effects of using different methods of estimation if the number of groups (k) is larger than 2?
(c) How do the constrained procedures 2 to 4 perform if they are applied to situations where they are theoretically wrong? For instance, how does linear discrimination perform if in fact the covariance matrices are proportional?

These are the questions we wish to answer, to some extent, by the simulation study reported in this paper.

4. Comparison of Classification Procedures and Computational Methods

Given a set of classification functions $\hat{q}_i(x)$, $i = 1, \ldots, k$, based on parameter estimates $\hat{\mu}_i$ and $\hat{\psi}_i$, we can determine the classification regions $\hat{R}_i$. The total probability of misclassification can then be written as

$$F(\hat{R}) = \sum_{i=1}^{k} \pi_i \int_{\mathbb{R}^p \setminus \hat{R}_i} f_i(x)\, dx \,. \tag{4.1}$$

Following Lachenbruch (1975, p. 30), we call $F(\hat{R})$ the *actual error rate.*

Different samples will yield different classification regions and thus different actual error rates. If we are interested in general properties of a classification procedure, and not just the performance of a given set of classification functions, the *expected actual error rate*, $E[F(\hat{R})]$, seems useful (see Lachenbruch 1975, p. 30). $E(F(\hat{R}))$ is the expectation of $F(\hat{R})$ under fixed distribution of populations and for fixed sample sizes. In this paper, we compare expected actual error rates of the Methods 1 - 4 under some specific parameter configurations. We will use simulation methods to determine these error rates because the mathematical difficulties seem prohibitive.

A FORTRAN program to approximate expected error rates by Monte Carlo simulation is described in Schmid (1987, Sec. 5.1 and Appendix II). In the present paper we give only a short outline of computational methods.

For a given parameter configuration, and for fixed sample size, we repeatedly generate the sample statistics S_i and $\bar{x}_i$, using the Wishart variate generator of Smith and Hocking (1972) and a modified Box-Muller transformation, respectively, which is described by Knuth (1969, p. 104).

From S_i we compute the appropriate estimators for ψ_i under all four methods, using the algorithms of Flury and Constantine (1985) for the CPC model and of Flury (1986) for the model of proportional covariance matrices. The actual error rates (4.1) are then computed for the four discrimination functions. The averages of repeatedly generated actual error rates serve as estimates for the expected error rates of the four discrimination methods. Standard errors of these means are computed for measuring the precision of the estimated expected error rates.

The actual error rate $F(\hat{R})$ for a given discriminant function is computed by two different methods, depending on the number of groups.

In the case of $k = 2$ groups, $F(\hat{R})$ can be computed as the sum of two one-dimensional integrals: After some transformations, we can apply a slightly modified method of Imhof (1961) to compute the distribution of quadratic forms in normal variables. This method and the solution of some difficulties in finding appropriate integration bounds are described by Schmid (1987, Sec. 5.1.3). A similar extension of Imhof's work can be found in Davies (1973, 1980).

In the case of $k \geq 3$ groups, $F(\hat{R})$ can be computed by Monte Carlo integration. To obtain reasonable precision of the integrals, this requires a much larger amount of computing than the method described for the case of two groups.

In this paper we give results obtained for $k = 2$ and for $k = 4$ groups. Each of the "situations" or "designs" described in the next section is defined by two mean vectors μ_1 and μ_2, and two covariance matrices ψ_1 and ψ_2, corresponding to a two-group case. An associated four group case was generated by using the parameters (μ_1, ψ_1), (μ_2, ψ_2), $(\mu_1 + \delta, \psi_1)$, and $(\mu_2 + \delta, \psi_2)$ to define four multivariate normal models, where δ is such that groups 3 and 4 are "totally separated" from 1 and 2. This means that the error rate is obtained entirely from the overlap of the densities of the first two groups, and the overlap of groups 3 and 4, which can be done using the numerical method of Imhof (1961) and Schmid (1987). In all calculations the numerical error of integration is strictly less than 10^{-3}.

5. Selection of Parameter Configurations

We will now describe and justify five different parameter configurations, or "designs," for which the Monte Carlo experiment was run

to estimate expected error rates. All five designs share the following characteristics:

- the analysis is done for $k = 2$ and $k = 4$ groups as described in section 4
- the dimension of all designs (i.e., the number of variables) is $p = 5$
- the sample sizes N_i are always equal, and range from $N_i = 10$ to $N_i = 60$ in steps of 5
- for each sample size, 1000 simulations are performed
- the prior probabilities are $\frac{1}{2}$ (if $k = 2$) and $\frac{1}{4}$ (if $k = 4$).

These choices reflect our experience from a larger number of simulations performed. The designs selected for presentation seemed to be the most informative ones. Evidently, the number of parameters that can be varied is, for all practical purposes, without limits, and so we constrained ourselves to a presentation of designs which have been described in the literature.

A detailed description of the five designs follows, all of them in terms of two covariance matrices ψ_1 and ψ_2, and one mean vector $\mu = \mu_1$, assuming without loss of generality that $\mu_2 = \mathbf{0}$.

Design 1: Efron's standard model.

Efron's (1975) standard model is a prototype for linear discrimination, defined by $\psi_1 = \psi_2 = I_5$ (the identity matrix of dimension 5×5), and $\mu = (\Delta,0,0,0,0)'$. We chose $\Delta = 2.5$, which is mostly arbitrary, but motivated by the idea that the four discrimination procedures should be compared in situations where good discrimination is possible but not trivial. The optimal error rate is $E[R] = .1056$. This design serves mostly the purpose of assessing the extent to which the four methods differ in an "optimal" situation. Note that optimal classification is based on the first variable alone; all other variables are pure noise.

Design 2: The proportional standard model.

This model, introduced by Flury and Schmid (1992, pp. 259), is very similar to Efron's standard model, the parameters being $\psi_1 = \mathbf{I}_5$, $\psi_2 = \gamma \cdot \mathbf{I}_5$, and $\mu = (\Delta,0,0,0,0)'$, where $\gamma > 0$ is a constant of proportionality. In the particular numerical example we chose $\Delta = 3$ and $\gamma = 4$. The optimal error rate is $E[R] = .076$.

Design 3: O'Neill's model.

This is a design based on a particular model studied by O'Neill (1984) for the purpose of comparing the performance of linear and (ordinary) quadratic classification rules. It is defined by the parameter setup $\psi_1 = \mathbf{I}_5$,

$\psi_2 = \text{diag}(\sigma^2, 1, 1, 1, 1)$, and $\mu = (\Delta, 0, 0, 0, 0)'$. Optimal classification is quadratic in the first variable, all other variables are just noise. In our particular numerical example we chose $\sigma = 3$ and $\Delta = 4.5$, yielding an optimal error rate of $E[R] = .107$.

O'Neill's model is interesting from two different points of view. First, as O'Neill (1984, 1992b) noticed, it takes a surprisingly large sample size for quadratic discrimination to outperform linear discrimination, even if the variances are quite different (such as $\sigma^2 = 9$ in our numerical example). Second, O'Neill's model is a special case of CPC, but not proportional. That is, both CPC and ordinary quadratic discrimination are theoretically correct. However, as the asymptotic calculations of Flury and Schmid (1992) show, CPC discrimination (at least in the case of $k = 2$ groups) is not expected to do much better than ordinary quadratic discrimination.

Design 4: A CPC model.

This design is based on the remark made already in section 3, that CPC discrimination should be better than quadratic discrimination if $\lambda_{1h}^{-1} - \lambda_{1j}^{-1} = \lambda_{2h}^{-1} - \lambda_{2j}^{-1}$ for all (j,h). Thus design 4 should provide an indication of the "maximum benefit" to be expected from using CPC rather than ordinary quadratic discrimination. Our (admittedly very arbitrary) parameter setup for the numerical example is $\psi_1 = \text{diag}\left[\frac{1}{5}, \frac{1}{2}, \frac{2}{3}, \frac{5}{6}, \frac{10}{11}\right]$, $\psi_2 = \text{diag}\left[\frac{1}{4}, 1, 2, 5, 10\right]$, and $\mu = (1,0,0,0,0)'$, yielding an optimal error rate of $E[R] = .067$.

Design 5: A model with "informative noise."

Evidently there are zillions of ways in which to generate situations where none of the constrained Methods 2 to 4 is theoretically correct. We chose to present a situation where the parameter setup is similar to design 1, but now the "noise" variables contain some information on discrimination due to differences in variability. At the same time, we wanted the CPC model to be "far from correct." A particular way to generate such models is to take

$$\psi_1 = \begin{bmatrix} 1 & 0 \\ 0 & \mathbf{E}_4(\rho) \end{bmatrix},$$

where $\mathbf{E}_4(\rho)$ is the equicorrelation matrix of dimension 4 by 4, with 1's on the diagonal, and ρ $(-\frac{1}{3} < \rho < 1)$ in all off-diagonal entries. Defining $\psi_2 = \text{diag}(1, 1 + 3\rho, 1 - \rho, 1 - \rho, 1 - \rho)$, we have a design where ψ_1 and ψ_2 have identical eigenvalues, but different eigenvectors. This is a situation that both CPC and proportional discrimination were not meant to deal with, and it

will therefore be interesting to see if they offer any advantage over ordinary quadratic discrimination.

In our particular numerical example we chose

$$\psi_1 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & .5 & .5 & .5 \\ 0 & .5 & 1 & .5 & .5 \\ 0 & .5 & .5 & 1 & .5 \\ 0 & .5 & .5 & .5 & 1 \end{bmatrix},$$

$\psi_2 = \text{diag}(1,2.5,.5,.5,.5)$, and $\mu = (2.5,0,0,0,0)'$. This yields an optimal error rate of $E[R] = .079$ (which is about three quarters of the optimal error rate in design 1).

6. Results

The numerical results obtained are summarized graphically for each of the five designs. There are two plots per design, one for $k = 2$ groups, and one for $k = 4$ groups. By construction (see section 4) of the four-group setup, the curve for Method 1 (DIFF) should be identical, up to sampling error, for $k = 2$ and $k = 4$. Since the 2-group and the 4-group results were obtained from independent runs of the simulation algorithm, this provides an additional check for plausibility of the results.

Design 1: (Efron's standard model; see Figure 1). As expected, linear discrimination performs consistently best for all sample sizes and both the 2-sample and the 4-sample case. Proportional discrimination is only slightly worse. CPC and ordinary quadratic discrimination perform distinctly worse, with expected error rates typically about twice as far from the optimal error rate as the corresponding expected error rates for EQU and PROP. Even in the 4-sample case the advantage of CPC over DIFF is only very slight.

In summary, if the assumption of equality of covariance matrices holds true, then linear discrimination is best (as is clear from the fact that it is the most parsimonious method), but not much is lost if proportional discrimination is used. If CPC or ordinary quadratic discrimination are used, roughly twice as many observations are needed to obtain the same expected error rate.

As for the numerical precision of the results, the standard errors for mean error rates in 1000 simulations ranged from .0002 ($N_i = 60$, $k = 4$, Method EQU) to .0041 ($N_i = 10$, $k = 4$, Method CPC). Similar ranges of standard errors were found in the four other designs as well, and will not be reported.

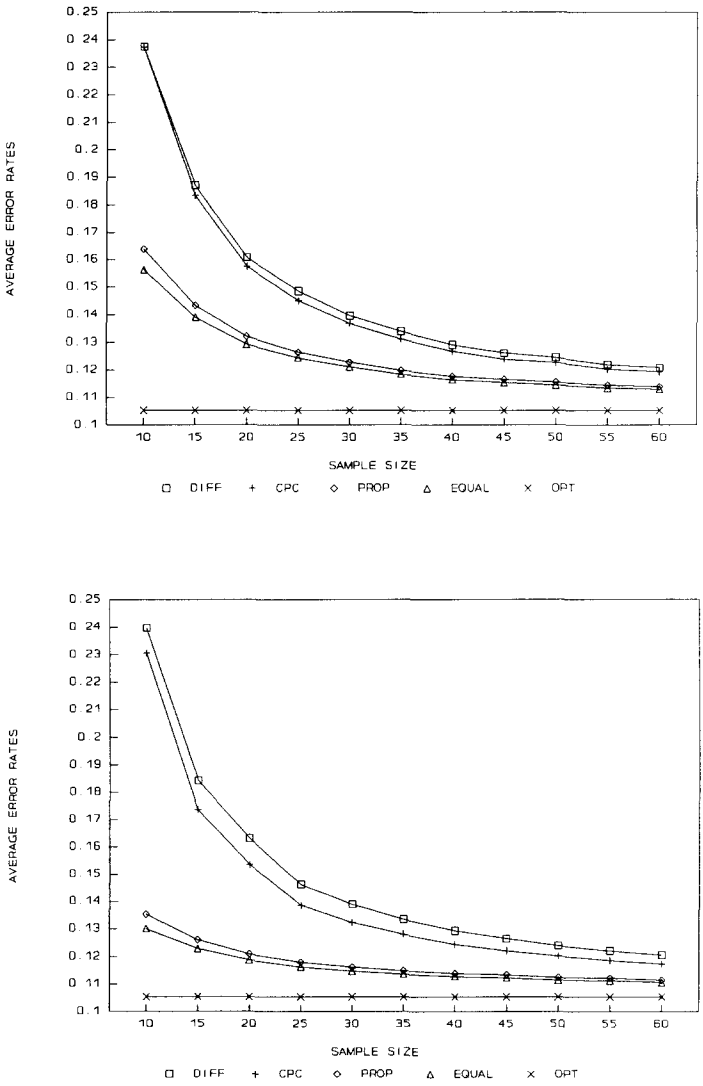

Figure 1. Average Error Rate for Design 1. (a) $k = 2$ groups; (b) $k = 4$ groups.

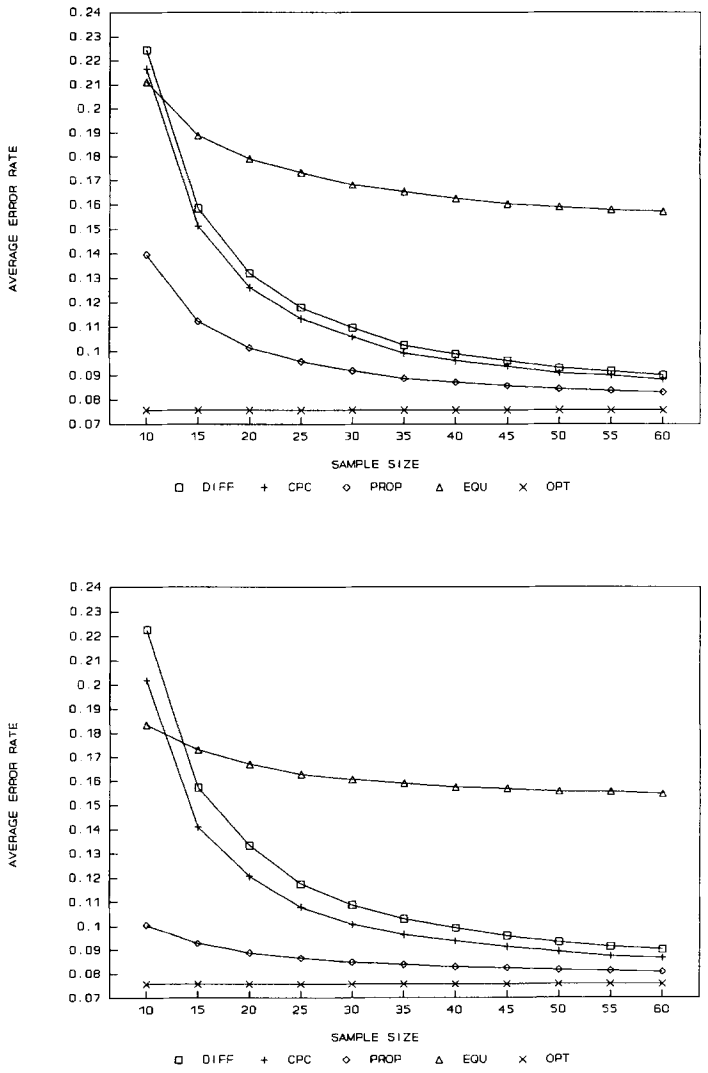

Figure 2. Average Error rate for Design 2. (a) $k = 2$ groups; (b) $k = 4$ groups.

Design 2: (Proportional standard model; see Figure 2). All methods except EQU are theoretically correct, but clearly PROP performs much better than either CPC, or DIFF. CPC performs somewhat better than DIFF, especially in the four-sample case. Remarkably, linear discrimination does as well or better than CPC or DIFF for very small sample sizes ($N_i = 10$), thus giving a case where using a theoretically wrong but parsimonious method may be better than using a correct but overparameterized one. This is particularly remarkable in view of the relatively large constant of proportionality ($\gamma = 4$) between the two covariance matrices. The main message from this example is, however, that proportional discrimination may provide a substantial improvement over ordinary quadratic discrimination, whenever it is appropriate.

Design 3: (O'Neill's model; see Figure 3). With the given numerical setup (i.e., a variance ratio of $\sigma^2 = 9$ for the first variable), one would expect the two theoretically correct Methods CPC and DIFF to do considerably better than the "inappropriate" Methods EQU and PROP. Interestingly, all four methods behave very similarly, and only at sample sizes around $N_i = 40$ DIFF starts to perform better than EQU. CPC discrimination appears to have a noticeable advantage over DIFF, particularly in the four-sample case. Interestingly, PROP performs worse than EQU for all sample sizes. A possible explanation for this somewhat unexpected phenomenon is that PROP introduces the "wrong" flexibility, compared to EQU. Proportional discrimination forces the boundaries of the classification regions to be genuinely quadratic, which is undesirable in this case.

O'Neill's model is the prototype of a situation where the direction of the mean difference vector in p-dimensional space is identical with the direction of the difference in variance. More precisely, $\mu_1 - \mu_2$ is proportional to the eigenvector of $\psi_1^{-1}\,\psi_2$ associated with the single characteristic root which is different from 1. The slow convergence of error rates in Figure 3 seems to indicate that none of the four methods of discrimination is well suited to handle this situation. Asymptotically, of course, both CPC and DIFF will reach the optimal error rate.

Design 4: (A CPC model; see Figure 4). Recall from section 5 that this design was "taylored" to favor CPC discrimination. Not surprisingly, CPC beats DIFF for all sample sizes, and by a larger margin for $k = 4$ groups than for $k = 2$ groups. The true surprise is the relatively good performance of proportional discrimination for all sample sizes. This is particularly astounding

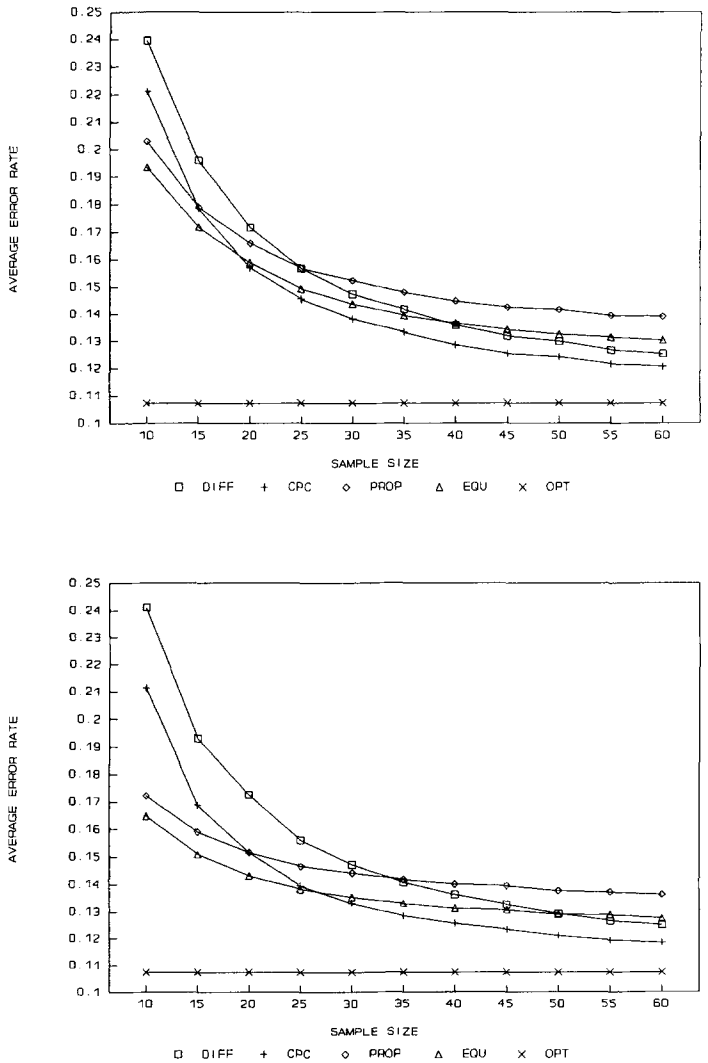

Figure 3. Average Error Rate for Design 3. (a) $k = 2$ groups; (b) $k = 4$ groups.

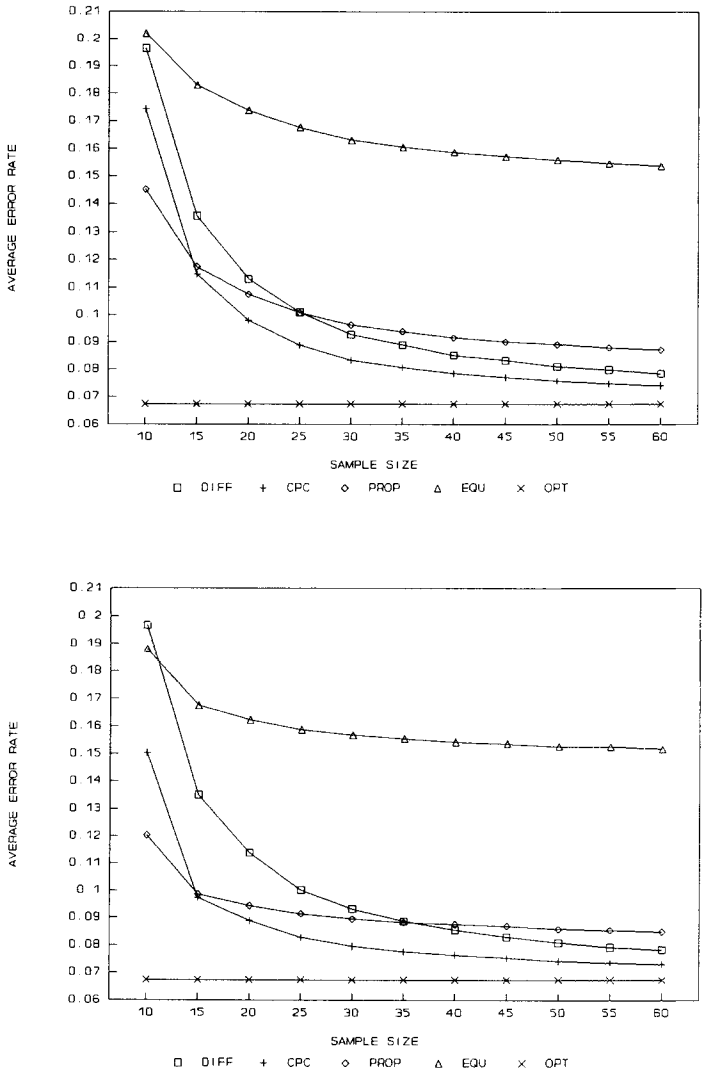

Figure 4. Average Error Rate for Design 4. (a) $k = 2$ groups; (b) $k = 4$ groups.

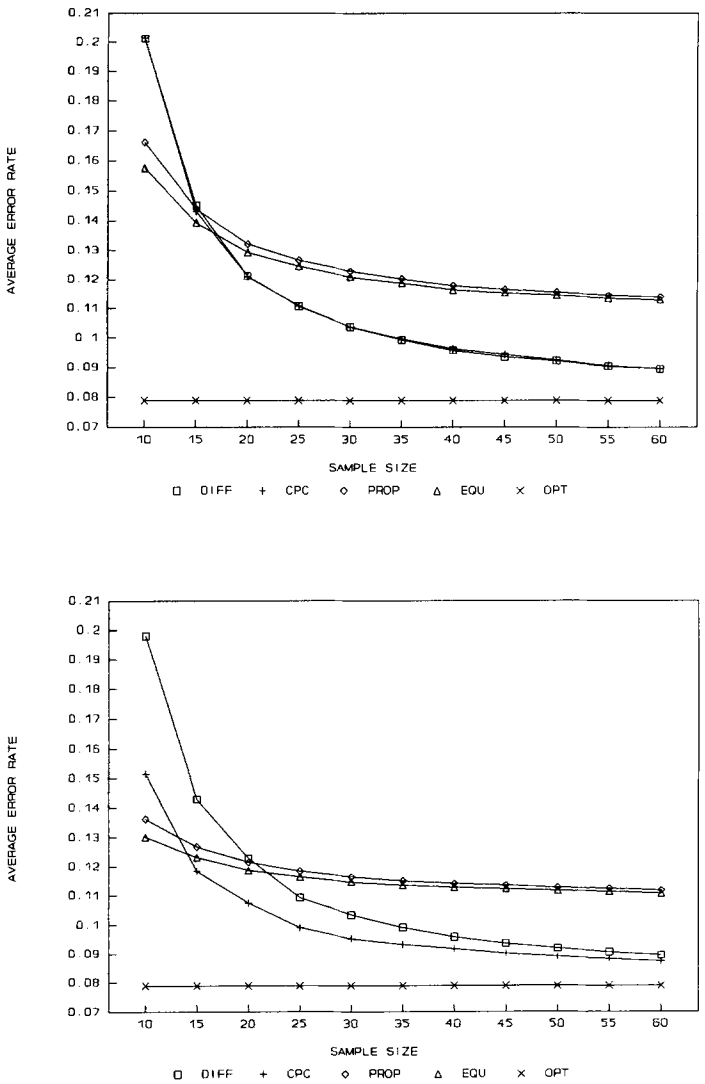

Figure 5. Average Error Rate for Design 5. (a) $k = 2$ groups; (b) $k = 4$ groups.

in view of the fact that the variance ratios range from 1.25 (for the first variable) to 11.0 (for the fifth variable), meaning that the two covariance matrices are far from proportional, and it underlines the usefulness of proportional discrimination due to its flexibility while introducing only a single parameter for each additional group.

Design 5: (A model with "informative noise"; see Figure 5). Recall that the parameters in this design were chosen such as to favor ordinary quadratic discrimination. All three constrained methods are theoretically wrong. For all sample sizes and both $k = 2$ and $k = 4$ groups, PROP and EQU discrimination perform about identically. CPC discrimination performs about as well as DIFF for all sample sizes when $k = 2$, and beats DIFF for all sample sizes when $k = 4$. Of course asymptotically the single correct method (ordinary quadratic) will take over. Some calculations that are not reported in Figure 5 indicate that up to sample sizes of 100 there will be no noticeable advantage of DIFF over CPC.

All five designs reported here should be considered as "representatives" of interesting models, but the numerical values chosen for the parameters are somewhat arbitrary. We chose the parameter values such as to illustrate as many (expected or unexpected) phenomena as possible on limited space.

7. Discussion

It is always difficult to draw general conclusions from limited simulation studies. We try to make some recommendations anyway, based on the simulation results as well as the asymptotic results of Flury and Schmid (1992).

1. Among competing and theoretically correct models, choose the most parsimonious one, i.e., the one which has the smallest number of functionally independent parameters to estimate. Note that this does not necessarily imply computational parsimony, because both the CPC and the PROP methods require iterative computations for parameter estimation.

2. Whenever the assumption of equality of covariance matrices seems questionable, consider proportional discrimination. In the worst case, it may perform slightly worse than linear discrimination (Designs 3 and 5). In more fortunate cases, it may perform much better than linear discrimination (Designs 2 and 4), and beat theoretically correct methods for finite sample sizes even when the covariance matrices are far from proportional (Design 4). This

confirms the suggestion made by the "Panel on Discriminant Analysis, Classification, and Clustering" (1987, p. 61), as well as empirical results obtained by Kirby et al (1991).

3. CPC discrimination may in some situations reduce the expected error rates, but the improvement over ordinary quadratic discrimination is typically not great. Together with the fact that CPC discrimination is not scale-invariant, this indicates that it does not offer much advantage, except perhaps in cases where a CPC model is fitted to the data for its own intrinsic value.

4. Ordinary quadratic discrimination should be avoided whenever possible.

For completeness we should mention that our proposed "intermediate" procedures CPC and PROP are not the only attempts to compromise between linear and quadratic discrimination. An interesting approach named "regularized discriminant analysis" (Friedman 1989, McLachlan 1992) consists of shrinking each sample covariance matrix $\mathbf{S}_i$ towards the pooled covariance matrix $\mathbf{S}$, using a single regularization parameter $\lambda, 0 \leq \lambda \leq 1$. The extreme cases $\lambda = 0$ and $\lambda = 1$ then yield exactly the ordinary quadratic and the linear classification rule, respectively. In addition, a shrinkage parameter $\gamma\,(0 \leq \gamma \leq 1)$ is introduced to control shrinkage of each covariance matrix towards a multiple of the identity matrix, thereby counteracting the bias of the sample eigenvalues. For practical applications, the optimal values of λ and γ are determined by cross-validation. Green and Rayens (1989) and Rayens (1990) propose a similar technique, based on empirical Bayes estimates, to find a proper compromise between linear and quadratic discrimination.

Another approach to reducing the "noise" introduced by parameter estimation is the so-called Euclidean distance classifier of Marco, Young, and Turner (1987). Their procedure is based on Euclidean distance to sample means, and therefore avoids estimation of variances and covariances altogether. This method is optimal when all covariance matrices are equal and proportional to the identity matrix. In our hierarchy of methods the Euclidean distance classifier would be number 5, following linear discrimination. Similarly, a discrimination method recently proposed by Chatterjee and Narayanan (1992) avoids estimation of variances and covariances entirely by using Hausdorff distance. Yet another method that ranks in the hierarchy between levels 1 and 3 (and therefore competes directly with CPC discrimination) is based on estimating a common correlation matrix for all k groups, as in Manly and Rayner (1987). This method avoids the difficulty concerning lack of scale invariance of CPC mentioned in remark 3 above. The common correlation matrix method estimates exactly the same number of parameters

as CPC; we speculate therefore that it would not offer any great advantage over DIFF either.

The obvious question addressed in the above mentioned paper of Friedman (1989), but left open here, is how to decide for a given method in a practical situation. A simple answer is, use likelihood ratio tests or an information criterion (see Flury 1988, Chapter 7). However as Greene and Rayens (1989) point out, testing is usually not appropriate because a parsimonious but theoretically wrong method (LDA) often outperforms the correct method (DIFF) even when a test for equality of covariance matrices rejects the null hypothesis at a very small level. A cross-validatory choice based on a predictive criterion is clearly to be preferred. This is even more evident in view of the fact that non-normality of the distributions may be as important as differences between covariance matrices. All calculations reported in this article are based on multivariate normality and may not be valid if the normality assumptions are severely violated. Nevertheless, just as in regularized discriminant analysis, a cross-validation based choice of one of the four methods of discrimination does not depend on the correctness of the model assumptions.

References

ANDERSON, T. W. (1984), *An Introduction to Multivariate Statistical Analysis*, Wiley, New York.

CHATTERJEE, S., and NARAYANAN, A. (1992), "A New Method of Discrimination and Classification Using a Hausdorff Type Distance," *Australian Journal of Statistics, 34*, 391-406.

DAVIES, R. B. (1973), "Numerical Inversion of a Characteristic Function," *Biometrika, 60*, 415-417.

DAVIES, R. B. (1980), "The Distribution of a Linear Combination of Chi-square Random Variables," Algorithm AS 155, *Applied Statistics, 29*, 323-333.

EFRON, B. (1975), "The Efficiency of Logistic Regression Compared to Normal Discriminant Analysis" *Journal of the American Statistical Association, 70*, 892-898.

FLURY, B. (1986), "Proportionality of k Covariance Matrices" *Statistics and Probability Letters, 4*, 29-33.

FLURY, B. (1988), *Common Principal Components and Related Multivariate Models*, Wiley, New York.

FLURY, B., and CONSTANTINE, G. (1985), "The F–G Diagonalization Algorithm," Algorithm AS 211, *Applied Statistics, 34*, 177-183.

FLURY, B. and SCHMID, M. (1992), "Quadratic Discriminant Functions with Constraints on the Covariance Matrices: Some Asymptotic Results," *Journal of Multivariate Analysis, 40*, 244-261.

FRIEDMAN, J. H. (1989), "Regularized Discriminant Analysis," *Journal of the American Statistical Association, 84*, 165-175.

GREENE, T., and RAYENS, W. S. (1989), "Partially Pooled Covariance Matrix Estimation in Discriminant Analysis," *Communications in Statistics - Theory and Methods, 18*, 3679-3702.

IMHOF, J. P. (1961), "Computing the Distribution of Quadratic Forms in Normal Variables," *Biometrika, 48*, 419-426.

KIRBY, S. P. J., THEOBALD, C. M., PIPER, J., and CAROTHEIS, A. D. (1991), "Some Methods of Combining Class Information in Multivariate Normal Discrimination for the Classification of Human Chromosomes," *Statistics in Medicine, 10*, 141-149.

KNUTH, D. E. (1969), *The Art of Computer Programming, Vol. 2*, Addison-Wesley, Reading, Mass.

LACHENBRUCH, P. A. (1975), *Discriminant Analysis*, Hafner Press, New York.

MANLY, B. F. J., and RAYNER, J. C. W. (1987), "The Comparison of Sample Covariance Matrices Using Likelihood Ratio Tests," *Biometrika, 74*, 841-847.

MARKS, S., and DUNN, O. J. (1974), "Discriminant Functions When the Covariance Matrices are Unequal," *Journal of the American Statistical Association, 69*, 555-559.

MARCO, V. R., YOUNG, D. M., and TURNER, D. W. (1987), "The Euclidean Distance Classifier: An Alternative to the Linear Discriminant Function," *Communications in Statistics-B, Simulation and Computation, 16*, 485-505.

McLACHLAN, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.

O'NEILL, T. J. (1984), "A Theoretical Method of Comparing Classification Rules Under Non-optimal Conditions With Application to the Estimates of Fisher's Linear and Quadratic Discriminant Rules Under Unequal Covariances Matrices," Technical report No. 217, Stanford University, Department of Statistics.

O'NEILL, T. J. (1992a), "The Bias of Fisher's Linear Discriminant Function when the Variances are not Equal," *Statistics and Probability Letters, 14*, 205-210.

O'NEILL, T. J. (1992b), "Error Rates of Non-Bayes Classification Rules and Robustness of Fisher's Linear Discriminant Function," *Biometrika, 79*, 177-184.

Panel on Discriminant Analysis, Classification, and Clustering (1989), "Discriminant Analysis and Clustering," *Statistical Science, 4*, 34-69.

RAYENS, W. S. (1990), "A Rule for Covariance Stabilization in the Construction of the Classical Mixture Surface," *Journal of Chemometrics*, , 159-169.

SCHMID, M. J. (1987), "Anwendungen der Theorie proportionaler Kovzarianzmatrizen und gemeinsamer Hauptkomponenten auf die quadratische Diskriminanzanalyse," Unpublished Ph.D. thesis, University of Berne (Switzerland), Department of Statistics.

SEBER, G. A. F. (1984), *Multivariate Observations*, Wiley, New York.

SMITH, W. B., and HOCKING, R. R. (1972), "Wishart Variate Generator," Algorithm AS 53, *Applied Statistics, 21*, 341-345.

WAHL, P. W., and KRONMAL, R. A. (1977), "Discriminant Functions when Covariances are Unequal and Sample Sizes are Moderate," *Biometrics, 33*, 479-484.

WAKAKI, H. (1990), "Comparison of Linear and Quadratic Discriminant Functions," *Biometrika, 77*, 227-229.