REGULARIZED DISCRIMINANT ANALYSIS

by
Jerome H. Friedman
Stanford Linear Accelerator Center
and
Stanford University

TECHNICAL REPORT NO. 17
OCTOBER 1987

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION
GRANT DMS 84-11300

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
REGULARIZED DISCRIMINANT ANALYSIS

by
Jerome H. Friedman
Stanford Linear Accelerator Center
and
Stanford University

TECHNICAL REPORT NO. 17
OCTOBER 1987

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION
GRANT DMS 84-11300

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Regularized Discriminant Analysis*

JEROME H. FRIEDMAN

Department of Statistics

and

Stanford Linear Accelerator Center

Stanford University, Stanford CA 94305

ABSTRACT

Linear and quadratic discriminant analysis are considered in the small sample high-dimensional setting. Alternatives to the usual maximum likelihood (plug-in) estimates for the covariance matrices are proposed. These alternatives are characterized by two parameters, the values of which are customized to individual situations by jointly minimizing a sample based estimate of future misclassification risk. Computationally fast implementations are presented, and the efficacy of the approach is examined through simulation studies and application to data.

*Work supported by the Department of Energy, contract DE-AC03-76SF00515.
1.0 Introduction

The purpose of classification or discriminant analysis is to assign objects to one of several \( K \) groups or classes based on a set of measurements \( X = (X_1, X_2, \ldots, X_p) \) obtained from each object or observation. An object is assumed to be a member of one (and only one) class and an error is incurred if it is assigned to a different one. The cost or loss associated with such an error is defined to be

\[
L(k, \hat{k}) \quad 1 \leq k, \hat{k} \leq K ,
\]

(1)

where \( k \) is the correct group on class assignment, and \( \hat{k} \) is the assignment that was actually made [\( L(k, k) \) is usually taken to be zero and \( L(k, \hat{k}) \geq 0 \)].

The vector valued measurements associated with all of the members of each class (population) are seldom identical but comprise a distribution of values

\[
F_k(X) = \int_{-\infty}^{X} f_k(X') \, dX' ,
\]

(2)

where \( F_k(X) \) is the fraction of the class \( k \) population with all \( p \) measurement values less than or equal to those of \( X \), and \( f_k(X) \) its corresponding probability density defined by Eq. (2). Sometimes the distribution of measurement values for the individual classes do not overlap so that it is possible to define regions of the measurement space each consisting of members of a single class. In this case the classes are said to be completely separable and it is (at least in principal) possible to do class assignment with perfect accuracy. Often, however, the class distributions do overlap and error free classification is not possible. In either case it is usually the goal to minimize the misclassification risk, which is defined to be the expected misclassification loss [Eq. (1)] over the sample to be classified.

If the class conditional distributions [Eq. (2)] are known, then it is possible to calculate misclassification risk and derive an assignment or classification rule to minimize it. The risk (expected loss) incurred in classifying an object with
measurement vector \( X \) as \( \hat{k} \) is

\[
R(\hat{k}|X) = \frac{\sum_{k=1}^{K} L(k, \hat{k}) f_k(X) \pi_k}{\sum_{k=1}^{K} f_k(X) \pi_k} ,
\]

(3)

where \( \pi_k \) is the unconditional prior probability of observing a class \( k \) member. This can be minimized by choosing \( \hat{k} \) to minimize the numerator in Eq. (3). For the special but commonly occurring case

\[
L(k, \hat{k}) = 1 - \delta(k, \hat{k}) ,
\]

(4)

this reduces to the simple rule: choose \( \hat{k} \) such that

\[
f_k(X) \pi_k = \max_{1 \leq k \leq K} f_k(X) \pi_k .
\]

(5)

The loss matrix [Eq. (4)] assigns a loss of one unit for each mistake irrespective of its type. The misclassification risk is then just the fraction of assignments that are incorrect. The assignment rule [Eq. (5)] simply assigns an object with measurement vector \( X \) to the class with highest posterior probability at \( X \). The rule resulting from choosing \( \hat{k} \) to minimize \( R(\hat{k}|X) \) [Eq. (3) or (5)] is known as the Bayes rule and it achieves minimal misclassification risk among all possible rules.

The class conditional distributions [Eq. (2)] are seldom known. More often we are able to obtain a sample of observations from each class that are correctly classified by some external mechanism. The objective is to use these observations as a training sample to construct a classification rule. Nearly all classification methods use the training sample in some way to estimate the class distributions, and then substitute these estimates into a rule of the form given by Eq. (3) or (5). Since these estimates generally deviate from the population distributions, such a rule will not likely achieve minimal risk, except perhaps asymptotically. Sometimes the unconditional class (prior) probabilities are also unknown. If the pooled (over classes) training data can be regarded as a random sample from the
pooled population distribution, then the prior probabilities can be estimated by the fraction of each class in the pooled sample

$$\pi_k = \frac{W_k}{W},$$  \hspace{1cm} (6)

with

$$W_k = \sum_{c(v)=k} w_v, \hspace{1cm} (7a)$$

and

$$W = \sum_{k=1}^{K} W_k. \hspace{1cm} (7b)$$

Here $v$ labels the observations in the training sample, $c(v)$ is the class of the $v^{th}$ observation, and $w_v$ is a weight or mass assigned to each observation.

The most often applied classification rules are based on the normal distribution

$$f_k(X) = (2\pi)^{-p/2} |\Sigma_k|^{-1/2} e^{-1/2(X-\mu_k)^T \Sigma_k^{-1}(X-\mu_k)} ,$$  \hspace{1cm} (8)

where $\mu_k$ and $\Sigma_k$ are the class $k$ ($1 \leq k \leq K$) population mean vector and covariance matrix. Assuming the simple loss structure [Eq. (4)] and substituting Eq. (8) into Eq. (5) leads to the classification rule

$$d_k(X) = \min_{1 \leq k \leq K} d_k(X),$$  \hspace{1cm} (9)

with

$$d_k(X) = (X - \mu_k)^T \Sigma_k^{-1}(X - \mu_k) + \ln |\Sigma_k| - 2\ln \pi_k.$$  \hspace{1cm} (10)

This quantity is often called the discriminant score for the $k^{th}$ class, whereas $d_k(X) + 2\ln \pi_k$ is referred to as the discriminant function. The first term on the right hand side of Eq. (10) is the well known Mahalonobis distance between $X$ and $\mu_k$. 
Using the classification rule [Eqs. (9) and (10)] is called quadratic discriminant analysis (QDA) since it separates the disjoint regions of the measurement space corresponding to each class assignment by quadratic boundaries. An important special case occurs when all of the class covariance matrices are presumed to be identical

$$\Sigma_k = \Sigma \quad 1 \leq k \leq K \quad .$$  \hspace{1cm} (11)

This is referred to as linear discriminant analysis (LDA) because the quadratic terms associated with Eqs. (9) and (10) cancel, resulting in linear decision boundaries.

Quadratic and linear discriminant analysis can be expected to work well if the class conditional densities are approximately normal and good estimates (for classification purposes) can be obtained for the population parameters defining the distributions (class mean vectors $\mu_k$ and covariance matrices $\Sigma_k$). In the classification context the ellipsoidal symmetry associated with the normal distribution appears to be the important aspect rather than its detailed shape [see Lachenbruch (1975) and James (1985)].

We examine here the problem of obtaining good estimates of the population parameters in settings where the training sample size is small and the dimensionality of the measurement vector space is large, where the goal is accurate classification.

2.0 Regularization and Shrinkage

In most applications of linear and quadratic discriminant analysis the parameters associated with the class densities are estimated by their sample analogues

$$\hat{\mu}_k = \bar{X}_k = \frac{1}{W_k} \sum_{c(v)=k} w_v X_v \quad ,$$  \hspace{1cm} (12)

and

$$\hat{\Sigma}_k = \frac{S_k}{W_k} = \frac{1}{W_k} \sum_{c(v)=k} w_v (X_v - \bar{X}_k)(X_v - \bar{X}_k)^T \quad ,$$  \hspace{1cm} (13)

with $W_k$ given by Eq. (7a). These so-called "plug-in" estimates are straightforward to compute and represent the corresponding maximum likelihood estimates. (Sometimes the covariance matrix estimates are scaled by a factor to
remove bias.) If the class populations are normal, these estimates give rise to
classification rules that approach minimum (Bayes) risk as the class sample sizes
approach infinity.

When the class sample sizes \( N_k, 1 \leq k \leq K, \) are small compared to the
dimension of the measurement space \( p, \) the covariance matrix estimates, especially,
become highly variable. Moreover, when \( N_k < p \) not all of their parameters are
even identifiable. The effect this has on discriminant analysis can be seen by
representing the class covariance matrices by their spectral decomposition

\[
\Sigma_k = \sum_{i=1}^{p} e_{ik} \mathbf{v}_{ik} \mathbf{v}_{ik}^T,
\]

where \( e_{ik} \) is the \( i^{th} \) eigenvalue of \( \Sigma_k \) (ordered in decreasing value) and \( \mathbf{v}_{ik} \) the
corresponding eigenvector. The inverse in this representation is

\[
\Sigma_k^{-1} = \sum_{i=1}^{p} \frac{\mathbf{v}_{ik} \mathbf{v}_{ik}^T}{e_{ik}},
\]

and the discriminant score [Eq. (10)] becomes

\[
d_k(X) = \sum_{i=1}^{p} \frac{[\mathbf{v}_{ik}^T(X - \mu_k)]^2}{e_{ik}}

+ \sum_{i=1}^{p} \ln e_{ik} - 2\ln \pi_k \quad (14)
\]

The discriminant score [Eq. (14)] is seen to be heavily weighted by the smallest
eigenvalues and the directions associated with their eigenvectors. When sample
based plug-in estimates are used, this becomes the eigenvalues and eigenvectors
of \( \hat{\Sigma}_k \) [Eq. (13)].

It is well known that the estimates based on Eq. (13) produce biased estimates
of the eigenvalues; the largest ones are biased high and the smallest ones are
biased towards values that are too low. This bias is most pronounced when
the population eigenvalues tend towards equality, and is correspondingly less severe when their values are highly disparate. In all cases, this phenomenon becomes more pronounced as the sample size decreases. When \( N_k \leq p \) the sample covariance matrix is singular with rank \( \leq N_k \) and the smallest \( p - N_k + 1 \) eigenvalues are estimated to be zero. The corresponding eigenvectors are then arbitrary subject to an orthogonality constraint.

The net effect of this biasing phenomenon on discriminant analysis is to (sometimes dramatically) exaggerate the importance associated with the low variance subspace spanned by the eigenvectors corresponding to the smallest sample eigenvalues. It is often presumed that the low variance subspace tends to be dominated by the noise in the system and the higher variance directions contain the signal. In discriminant analysis this means that the standardized population class mean differences project mostly on the high variance subspace. If this is the case, then the biased eigenvalue estimates have the effect of dramatically amplifying the noise at the expense of the signal. Even when this is not the case, most of the variance incurred in estimating the discriminant scores [Eqs. (10) and (14)] is associated with directions of low sample variance.

One way to attempt to mitigate this problem is to try to obtain more reliable estimates of the eigenvalues by correcting the eigenvalue distortion in the sample covariance matrix. James and Stein (1961), Stein et al. (1972), Stein (1973), Stein (1975), Efron and Morris (1976), Olkin and Sellian (1977), Haff (1980), Lin and Perlman (1984), Takemara (1984) and Dey and Srinivasan (1985) have studied this approach by seeking estimates that minimize particular loss criteria (often some form of squared-error loss) on the eigenvalue estimates. None of these loss criteria that have been studied, however, are related to misclassification risk of a discriminant function. Also, they nearly all require that \( \Sigma_k \) be nonsingular.

Another approach is to employ a regularization method. Regularization techniques have been highly successful in the solution of ill- and poorly-posed inverse problems. [See Titterington (1985) and O'Sullivan (1986) for reviews.] Roughly, a problem is poorly posed if the number of parameters to be estimated is comparable to the number of observations and ill-posed if that number exceeds the sample size. In these cases the parameter estimates can be highly unstable, giving rise to high variance. By employing a method of regularization, one attempts to
improve the estimates by biasing them away from their sample based values towards values that are deemed to be more physically plausible. The high variance associated with the sample based estimate is thereby reduced at the expense of potentially increased bias. This bias variance trade-off is generally regulated by one or more (degree-of-belief) parameters that control the strength of the biasing towards the plausible set of (population) parameter values. For given value(s) of the regularization parameter(s), the increase in bias will depend on how closely the plausible set of parameters actually represent those of the population. Therefore, if a bad guess were made, one would like to employ a small amount of regularization; whereas for a good guess, a high degree of regularization would be appropriate, dramatically decreasing the variance at the expense of low increase in bias. Since one seldom knows the accuracy of the guess, sample based methods are often used to try to estimate values for the regularization parameters as well. These methods are usually based on unbiased risk assessment or cross-validatory choice. [See Titterington (1985) and O'Sullivan (1986) for reviews.]

Quadratic discriminant analysis is clearly ill-posed if $N_k \leq p$ for any class, and poorly posed whenever $N_k$ is not considerably larger than $p$. One method of regularization that is routinely applied in discriminant analysis is to replace the individual class sample covariance matrices by their average

$$\hat{\Sigma}_k = \frac{S}{W},$$  \hspace{1cm} (15)

where

$$S = \sum_{k=1}^{K} S_k,$$  \hspace{1cm} (16)

with $W$ given by Eq. (7b) and $S_k$ by Eq. (13). This applies a considerable degree of regularization by substantially reducing the number of parameters to be estimated. Even if the population class covariance matrices are substantially different, the decrease in variance accomplished by using the pooled covariance estimate can sometimes lead to superior performance, especially in small sample settings. This is a large part of the reason for the success and popularity of linear discriminant analysis.
The choice between linear and quadratic discriminant analysis represents a fairly restrictive set of regularization alternatives. A less limited set of alternatives is represented by

$$
\hat{\Sigma}_k(\lambda) = \frac{S_k(\lambda)}{W_k(\lambda)} ,
$$

(17a)

where

$$
S_k(\lambda) = (1 - \lambda) S_k + \lambda S ,
$$

(17b)

and

$$
W_k(\lambda) = (1 - \lambda) W_k + \lambda W ,
$$

(17c)

with $S_k$ given by Eq. (13), $S$ by Eq. (16), and $W_k$ and $W$ by Eq. (7). The regularization parameter $\lambda$ takes on values $0 \leq \lambda \leq 1$. It controls the degree of shrinkage of the individual class covariance matrix estimates towards the pooled estimate. The value $\lambda = 0$ gives rise to quadratic discriminant analysis (QDA), whereas $\lambda = 1$ yields linear discriminant analysis (LDA). Values between these limits represent degrees of regularization less severe than LDA. Since it is often the case that even small amounts of regularization can largely eliminate quite drastic instability [Titterington (1985)], smaller values of $\lambda$ (than $\lambda = 1$) have the potential of superior performance when the population class covariance matrices substantially differ. It is unlikely that a good value of $\lambda$ will be known in advance. It is imperative, therefore, that a sample based method be provided for its choice. This is discussed in Section 3.0.

The regularization provided by Eqs. (17) is still fairly limited and is not the only natural way to regularize QDA. First of all it might not provide for enough regularization. If the total sample size

$$
N = \sum_{k=1}^{K} N_k
$$

(18)

is less than or comparable to $p$, then even LDA is ill- or poorly-posed. Secondly, biasing the sample class covariance matrices toward commonality may not be
the most effective way to shrink them. For example, if the population class covariance matrices were all (quite different) multiples of the identity matrix, then shrinkage towards LDA would introduce severe bias, whereas shrinking each sample class covariance matrix towards the identity matrix multiplied by its average eigenvalue \( \text{trace } (\hat{\Sigma}_k)/p \) would introduce almost no bias. Ridge regression regularizes ordinary linear least squares regression by shrinking toward a multiple of the identity matrix.

To these ends we further regularize the sample class covariance matrix estimates beyond that provided by Eqs. (17) through

\[
\hat{\Sigma}_k(\lambda, \gamma) = (1 - \gamma) \hat{\Sigma}_k(\lambda) + \frac{\gamma}{p} \text{trace } [\hat{\Sigma}_k(\lambda)] \ I , \tag{19}
\]

with \( \hat{\Sigma}_k(\lambda) \) given by Eqs. (17) and \( I \) being the identity matrix. For a given value of \( \lambda \), the additional regularization parameter \( \gamma, 0 \leq \gamma \leq 1 \), controls shrinkage toward a multiple of the identity matrix. The multiplier is just the average eigenvalue of \( \hat{\Sigma}_k(\lambda) \). This shrinkage has the effect of decreasing the larger eigenvalues and increasing the smaller ones, thereby counteracting the biasing inherent in sample based estimation of eigenvalues.

Equations (17) and (19) represent a two-parameter family of regularized sample class covariance matrix estimators, to be used with the class discriminant scores

\[
d_k(X) = (X - \bar{X}_k)^T \hat{\Sigma}_k^{-1}(\lambda, \gamma)(X - \bar{X}_k) \\
+ \ln \left| \hat{\Sigma}_k(\lambda, \gamma) \right| - 2\ln \pi_k , \tag{20}
\]

to perform discriminant analysis. Values for the two regularization parameters, \( 0 \leq \lambda \leq 1 \) and \( 0 \leq \gamma \leq 1 \), are chosen so as to jointly minimize an unbiased estimate of future misclassification risk (see Section 3.0). We refer to this approach as "regularized discriminant analysis" (RDA).

Regularized discriminant analysis provides for a fairly rich class of regularization alternatives. Figure 1 illustrates the \( \lambda, \gamma \) plane. The four corners defining the extremes of the plane represent fairly well known classification procedures. The lower left corner \( (\lambda = 0, \ \gamma = 0) \) represents quadratic discriminant analysis.
The lower right ($\lambda = 1$, $\gamma = 0$) represents linear discriminant analysis. The upper right corner ($\lambda = 1$, $\gamma = 1$) corresponds to the nearest means classifier well known in pattern recognition; an observation is assigned to the class with the closest (Euclidean distance) mean. The upper left corner of the plane represents a weighted nearest means classifier with the class weights being inversely proportional to the average variance of the measurement variables within the class. Holding $\gamma$ fixed at zero and varying $\lambda$ produces models in between QDA and LDA. Holding $\lambda$ fixed at zero and increasing $\gamma$ attempts to un-bias the sample based eigenvalue estimates. Holding $\lambda$ fixed at one and increasing $\gamma$ gives rise to a ridge regression analogue for LDA.

3.0 Model Selection

A good pair of values for $\lambda$ and $\gamma$ is not likely to be known in advance. We must, therefore, have a (training) sample based method to estimate them. This is a common objective associated with methods of regularization. Two general approaches have been developed toward this end, unbiased risk assessment and cross-validatory choice. Recently techniques based on the bootstrap [Efron (1983)] have also been employed. Most developments of unbiased risk assessment in the predictive setting (regression) are appropriate when one conditions on the design. That is, one assumes that all training sample replications (from the population) give rise to exactly the same set measurement vectors and the class label is the random variable. This setup is sometimes reasonable in regression settings but almost never in classification problems. Cross-validation, on the other hand, is appropriate in exactly the situation where one does not condition on the design. Bootstrapping can be applied in either setting [see Efron (1983)]. The computational advantages associated with the cross-validation approach in this particular application however (see below) make it the most attractive choice here.

The basic idea of cross-validation is to obtain a (nearly) unbiased estimate of the future prediction error associated with a particular observation $X_p$ by removing it from the model building process. That is, the classification rule is developed on the $N - 1$ training observations exclusive of $X_p$ and then it is used to classify $X_p$. Each of the training observations is in turn held out and
then classified in this manner. The resulting misclassification loss averaged over the training sample is then used as an estimate of future misclassification risk. Lachenbruch (1975) calls this the "leaving-one-out" method of estimating the error rate for a classification rule.

Our approach to model selection is to choose values of the covariance matrix mixing parameter \( \lambda \), and the eigenvalue shrinkage parameter \( \gamma \), that jointly minimize this cross-validated estimate of future misclassification risk. This gives rise to a two-parameter numerical minimization problem. Our strategy is to choose a grid of points on the \( \lambda, \gamma \) plane, \( 0 \leq \lambda \leq 1, 0 \leq \gamma \leq 1 \), evaluate the cross-validated estimate of misclassification risk at each prescribed point on the grid, and then choose the point with the smallest estimated risk as our estimate for the optimal regularization parameter values, \( \hat{\lambda} \) and \( \hat{\gamma} \). Typically, the size of the optimization grid \( N_p \) is taken to be from 25 to 50 points.

This strategy, if implemented in a straightforward manner, would require excessive computation. At each grid point, \( N \) [Eq. (18)] sets of discriminant scores [Eq. (20)] would have to be calculated. Thus, the increase in computation for the entire procedure would be \( N_p \times N \) times the computation required for a single discriminant analysis. Fortunately, however, it is possible to develop a strategy based on matrix updating formulae to dramatically reduce this computational burden and bring it to an acceptable level.

In order to apply cross-validation it is necessary to compute the \( K \) discriminant scores [Eq. (20)] with the observation to be classified (say \( X_v \)) left out

\[
\begin{align*}
d_k \setminus v(X_v) &= (X_v - \overline{X}_k \setminus v)^T \hat{\Sigma}_k \setminus v^{-1}(\lambda, \gamma)(X_v - \overline{X}_k \setminus v) \\
&\quad + \ln |\hat{\Sigma}_k \setminus v(\lambda, \gamma)| - 2\ln \pi_k
\end{align*}
\tag{21}
\]

Here the notation \( \setminus v \) refers to the corresponding quantity computed with the \( v^{th} \) observation removed. One could simply recompute the quantities involved from scratch using the \( N - 1 \) observations exclusive of \( X_v \). However, as indicated above, this results in excessive total computation. In the case of linear and quadratic discriminant analysis advantage can be taken of the fact that a covariance matrix with an observation removed differs from the complete covariance matrix by a rank one matrix. One can then express the covariance matrices
through their Cholesky decompositions and take advantage of fast rank-one down
dating formulae to compute $d_{k \setminus v}$ [Eq. (21)] from $d_k$ [Eq. (20)] [see Golub and Van
Loan (1983)].

Unfortunately, removing an observation does not result in a rank-one down-
date of $\hat{\Sigma}_k(\lambda, \gamma)$ [Eqs. (17),(19)]. It can be shown that

$$W_{k \setminus v}(\lambda) \hat{\Sigma}_{k \setminus v}(\lambda, \gamma) = W_k(\lambda) \hat{\Sigma}_k(\lambda, \gamma) - (1 - \gamma) Z_v Z_v^T \frac{\gamma}{p} |Z_v|^2 I \ , \ (22a)$$

with $W_k(\lambda)$ given by [Eq. (17c)], and

$$W_{k \setminus v}(\lambda) = W_k(\lambda) - s_k(v) w_v \ , \ (22b)$$

$$s_k(v) = \begin{cases} 1 & \text{if } c(v) = k \\ \lambda & \text{otherwise} \end{cases} \quad (22c)$$

$$Z_v = \sqrt{b_k(v)} \left( X_v - \bar{X}_{c(v)} \right) \ , \ (22d)$$

and

$$b_k(v) = \frac{s_k(v) W_{c(v)} w_v}{W_{c(v)} - w_v} \ . \ (22e)$$

Thus, removing an observation is equivalent to downdating $\hat{\Sigma}_k(\lambda, \gamma)$ by a rank-
one matrix plus a multiple of the identity matrix. The only matrix representation
for which it is easy to obtain the inverse of a matrix downdated by a multiple of
$I$, from its original inverse, is the spectral decomposition:

$$W_k^{-1}(\lambda) \hat{\Sigma}_k^{-1}(\lambda, \gamma) = \sum_{i=1}^p \frac{v_i v_i^T}{e_i} \ . \ (23a)$$

Then

$$\left[ W_k(\lambda) \hat{\Sigma}_k(\lambda, \gamma) - a I \right]^{-1} = \sum_{i=1}^p \frac{v_i v_i^T}{(e_i - a)} \ , \ (23b)$$

where $e_i$ is the $i^{th}$ eigenvalue of $W_k(\lambda) \hat{\Sigma}_k(\lambda, \gamma)$, $v_i$ its corresponding eigenvector, and $a$ is a real valued scalar. Once this downdate has been performed,
the remaining rank-one downdate can be accomplished through the Sherman–Morrison formula [Golub and Van Loan (1983)]:

\[
(A - rr^T)^{-1} = A^{-1} + \frac{A^{-1}rr^TA^{-1}}{1 - r^TA^{-1}r},
\]  

(24)

where \( A \) is a nonsingular matrix and \( r \) is a vector. In our case \( A^{-1} \) is given by [Eq. (23)] with

\[
a = \frac{\gamma |Z_v|^2}{p},
\]  

(25a)

and

\[
r = \sqrt{1 - \gamma} \ Z_v,
\]  

(25b)

with \( Z_v \) given by Eq. (22).

In addition to the downdated inverse class covariance matrix, we still need to downdate its determinant and the class mean vector, in order to obtain the downdated discriminant score [Eq. (21)]. It is easily verified that

\[
\bar{X}_{k\setminus v} = \begin{cases} X_k & \text{if } c(v) \neq k \\ \frac{W_kX_k - w_vX_v}{W_k - w_v} & \text{otherwise} \end{cases},
\]  

(26)

and

\[
\ln |W_{k\setminus v}(\lambda) \ \bar{S}_{k\setminus v}(\lambda, \gamma)| = \sum_{i=1}^{p} \ln(e_i - a) + \ln \left[ 1 - \sum_{i=1}^{p} \frac{r_i^2}{e_i - a} \right]
\]  

(27)

with \( e_i \) given by Eq. (23) and \( a \) and \( r \) given by Eq. (25).

These quantities [Eqs. (22)–(27)] can be substituted into Eq. (21) to obtain the \( K \) class cross-validated discriminant scores with computation proportional to \( p^2 \) for each observation. The corresponding average misclassification loss over the training sample using these cross-validated scores is then taken to be an estimate of the future misclassification risk for the corresponding values of \( \lambda \) and \( \gamma \).
A substantial amount of additional computation can be saved by taking advantage of the fact that for a fixed value of \( \lambda \) the eigenvectors \( \mathbf{v}_i \) [Eq. (23a)] are independent of \( \gamma \). Changing \( \gamma \) is equivalent to an update by a multiple of the identity matrix. Thus, the \( K \) spectral decompositions and the corresponding rotations \( \mathbf{v}_{ik}^T (\mathbf{X}_o - \bar{\mathbf{X}}_k) \) (\( 1 \leq i \leq p, 1 \leq k \leq K, 1 \leq v \leq N \)) need only be recalculated when the value of \( \lambda \) changes. For each distinct value of \( \lambda \) on the optimization grid, the set of points corresponding to different values of \( \gamma \) can each be cross-validated in time proportional to \( pN \). Therefore, the grid points should be visited in an order that causes \( \lambda \) to change as few times as possible.

4.0 Discussion

The potential for RDA to improve misclassification risk over that of QDA or LDA will depend on the situation (class population distributions and sample size). In situations for which the class sample sizes \( N_k \) are all much larger than the dimension of the measurement space \( p \), no regularization is needed, and the model selection procedure should tend to produce small values of \( \lambda \) and \( \gamma \). However, the estimates of the optimal regularization parameters themselves have an associated bias and variance, so that one would expect the performance of RDA to be slightly worse than QDA. In these large sample settings, however, one might question the use of procedures based on normality, and favor more nonparametrically oriented methods such as nearest neighbors [see Lachenbruch (1975)] or recursive partitioning [Breiman et al. (1984)].

In small sample settings where QDA is either ill- or poorly-posed, it is not likely to be competitive with either LDA or RDA. Situations in which the population class covariance matrices are either very different and/or not too ellipsoidal should favor RDA. (It should be noted that in these settings the sample class covariance matrices are nearly always highly ellipsoidal.)

Another situation that favors RDA is when the (standardized) differences between the class means project mainly on the high variance subspaces. The most difficult situation for RDA is when the population class covariance matrices are all equal and highly ellipsoidal, and the differences between the class means project mostly on the low variance subspace. In this case any regularization away from LDA (\( \lambda = 1, \gamma = 0 \)) will be highly counter productive. Again, owing to the
bias and variance associated with the regularization parameter estimates, RDA should be slightly worse than LDA. When the sample size is small enough so that even LDA is ill- or poorly-posed then, in any situation, the regularization afforded by RDA is the only hope.

It is the goal of the model selection procedure to pick appropriate values for the regularization parameters for each particular situation. For those that are favorable to RDA it should choose a high degree of regularization substantially reducing the variance, while introducing little extra bias, thereby dramatically reducing misclassification risk. On the other hand, when the situation is unfavorable to RDA, the hope is that the model selection procedure will (on average) produce a small degree of regularization so that the performance of RDA will be only slightly worse than that of LDA or QDA. All of this depends of course upon the performance of the model selection procedure. This is investigated in the next section.

5.0 Simulation Studies

In this section we use computer simulation to investigate the performance of RDA compared to LDA and QDA in a variety of settings (class population distributions and ratios of variables to observations). The goal is to study the overall effectiveness of RDA and to identify some situations where one would (and would not) expect substantial improvement with RDA. In all cases the population class conditional distributions were normal [Eq. (8)] and the total sample size was $N = 40$ [Eq. (18)]. A fairly wide spectrum of situations was chosen in terms of the mean and covariance structure of the class populations, some of which would be suspected to be highly favorable, and others highly unfavorable, to RDA. For each situation, simulation experiments were performed for $p = 6, 10, 20, \text{ and } 40$. In all cases there were $K = 3$ groups on classes. The optimization grid of $(\lambda, \gamma)$ values was defined by the outer product of $\lambda = (0, .125, .354, .650, 1.0)$, and $\gamma = (0, .25, .5, .75, 1.0)$. (When the class covariance matrix estimates associated with QDA or LDA happened to be singular, the zero eigenvalues were replaced with a small number just large enough to permit numerically stable inversion. This has the effect of producing a classification rule based on Euclidean distance in the zero variance subspace.)
Each experiment consisted of one hundred replications of the following procedure. First a training sample of size $N = 40$ is randomly drawn from the pooled class population distributions. The prior probability of each of the three classes was taken to be equal so that the expected number of observations in each class was 13.3. However, the actual number in any particular replication was itself a (multinomial) random variable. Each such training data set was used to construct the linear, quadratic and estimated optimal regularized discriminant rules. An additional (test) data set of size $N = 100$ was then randomly generated from the same population and classified with the three rules derived from the training set, thereby obtaining an estimate of the misclassification risk, using the misclassification loss given by Eq. (4).

The tables, summarizing the results for each situation, present the average test misclassification risk (with standard deviations) over the one hundred replications for each of the three classification rules. Also presented are the average (minimizing) cross-validated estimate for the RDA rule, its correlation with the actual test set estimate for the RDA rule, and the mean and standard deviations of the selected regularization parameter ($\hat{\lambda}, \hat{\gamma}$) values over the one hundred replications.

5.1 EQUAL SPHERICAL COVARIANCE MATRICES

This is a situation that might somewhat favor RDA. Each of the three classes was generated from a population with the identity covariance matrix. The population mean of the first class was the origin. The means of the other two classes were taken to be 3.0 in two orthogonal directions. Table 1 summarizes the results.

The quantities in parantheses are the standard deviations of the respective quantities. The standard deviations of the corresponding averages are one tenth these amounts.

As suspected, RDA gives uniformly lower misclassification risk than LDA or QDA. As the dimension of the measurement space increases (relative to sample size) its advantage increases, becoming dramatic for the higher dimensionalities. (It should be noted that the risk estimates for the three methods are not independent when studying uncertainty estimates.) The cross-validated estimate
Table 1

<table>
<thead>
<tr>
<th></th>
<th>$P = 6$</th>
<th>$p = 10$</th>
<th>$p = 20$</th>
<th>$p = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MISCLASSIFICATION RISK:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDA</td>
<td>.11 (.03)</td>
<td>.12 (.04)</td>
<td>.16 (.05)</td>
<td>.19 (.05)</td>
</tr>
<tr>
<td>LDA</td>
<td>.13 (.04)</td>
<td>.16 (.05)</td>
<td>.26 (.05)</td>
<td>.58 (.08)</td>
</tr>
<tr>
<td>QDA</td>
<td>.26 (.08)</td>
<td>.49 (.10)</td>
<td>.57 (.07)</td>
<td>.49 (.06)</td>
</tr>
<tr>
<td><strong>MINIMIZING CROSS-VALIDATED ESTIMATE FOR RDA:</strong></td>
<td>.09 (.05)</td>
<td>.10 (.05)</td>
<td>.12 (.06)</td>
<td>.15 (.06)</td>
</tr>
<tr>
<td><strong>CORRELATION (TEST SET, CROSS-VALIDATION):</strong></td>
<td></td>
<td></td>
<td></td>
<td>.15</td>
</tr>
<tr>
<td></td>
<td>-.11</td>
<td>-.10</td>
<td>.17</td>
<td></td>
</tr>
<tr>
<td><strong>AVERAGE REGULARIZATION PARAMETER VALUES:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda$</td>
<td>.77 (.37)</td>
<td>.79 (.35)</td>
<td>.75 (.37)</td>
<td>.78 (.34)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>.74 (.34)</td>
<td>.72 (.32)</td>
<td>.74 (.28)</td>
<td>.80 (.22)</td>
</tr>
</tbody>
</table>

of RDA risk at its minimum is seen to underestimate the actual risk by about 20% on average. The correlation between them is seen to be surprisingly small. As would be hoped for, RDA is choosing a high degree of regularization for both $\lambda$ and $\gamma$ on average.

5.2 **UNEQUAL SPHERICAL COVARIANCE MATRICES**

This situation should favor RDA even more than the previous example since, unlike the previous one, here LDA is biased. Each of the three classes was generated with covariance matrix $kI$, where $k$ is the class number ($1 \leq k \leq 3$). As before the population mean for the first class is at the origin; the means for classes two and three are shifted in orthogonal directions, class two by a distance of 3.0, and class three by a distance of 4.0. Table 2 summarizes the results.
Table 2

<table>
<thead>
<tr>
<th></th>
<th>$P = 6$</th>
<th>$p = 10$</th>
<th>$p = 20$</th>
<th>$p = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MISCLASSIFICATION RISK:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDA</td>
<td>.17 (.04)</td>
<td>.13 (.05)</td>
<td>.10 (.05)</td>
<td>.05 (.04)</td>
</tr>
<tr>
<td>LDA</td>
<td>.29 (.06)</td>
<td>.32 (.06)</td>
<td>.41 (.07)</td>
<td>.59 (.07)*</td>
</tr>
<tr>
<td>QDA</td>
<td>.33 (.07)</td>
<td>.53 (.09)</td>
<td>.60 (.07)</td>
<td>.53 (.06)</td>
</tr>
<tr>
<td><strong>MINIMIZING CROSS-VALIDATED ESTIMATE FOR RDA:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.14 (.05)</td>
<td>.11 (.04)</td>
<td>.07 (.04)</td>
<td>.04 (.03)</td>
</tr>
<tr>
<td><strong>CORRELATION (TEST SET, CROSS-VALIDATION):</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>-.03</td>
<td>.05</td>
<td>.05</td>
<td>.08</td>
</tr>
<tr>
<td><strong>AVERAGE REGULARIZATION PARAMETER VALUES:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{\lambda}$</td>
<td>.10 (.13)</td>
<td>.06 (.12)</td>
<td>.04 (.08)</td>
<td>.04 (.03)</td>
</tr>
<tr>
<td>$\bar{\gamma}$</td>
<td>.81 (.26)</td>
<td>.88 (.20)</td>
<td>.93 (.16)</td>
<td>.97 (.11)</td>
</tr>
</tbody>
</table>

As conjectured, RDA strongly dominates with smaller risk at all dimensions, the relative improvement again increasing with dimension. The cross-validated estimate for RDA is as before about 20% below its actual risk and essentially uncorrelated with it. The model selection procedure behaved quite reasonably, choosing small values of the covariance matrix mixing parameter $\lambda$, and very large values for the eigenvalue shrinkage parameter $\gamma$.

5.3 **EQUAL HIGHLY ELLIPSOIDAL COVARIANCE MATRICES**

Here we consider two situations that ought to prove difficult for RDA. The covariance matrices of all three class populations are the same and highly ellipsoidal. The first case is constructed so that the location differences between the classes are concentrated in the low variance suspace, whereas in the second they
are concentrated in the high variance subspace. The eigenvalues of the common population covariance matrices are given by

\[ e_i = \left[ \frac{9(i-1)}{p-1} + 1 \right]^2, \quad 1 \leq i \leq p, \quad (28) \]

so that the ratio of the largest to smallest eigenvalues is one hundred.

We first consider the case where the class mean differences project mainly on the low variance subspace. This represents the most difficult problem from the point of view of RDA. The mean of the first class is again located at the origin. The mean vectors for the class two and three populations in terms of the population eigenvectors are

\[ \mu_{2i} = 2.5 \sqrt{\frac{e_i}{e_i}} \cdot \left( \frac{p-i}{p} \right)^{\frac{1}{2}} \]

\[ \mu_{3i} = (-1)^i \mu_{2i}, \quad 1 \leq i \leq p \]

with \( e_i \) given by Eq. (28). The results are given in Table 3.

Linear discriminant analysis performs slightly better in all but the highest dimension where no method does particularly well. This situation, as constructed, is ideal for LDA since any shrinkage away from the point \( (\lambda = 1, \gamma = 0) \) is strongly counterproductive. The regularization parameter values selected by the cross-validation procedure are seen to be concentrated in this corner of the \( \lambda, \gamma \) plane. Note the increase in \( \overline{\gamma} \) as the dimension increases. At the highest dimensions considerable shrinkage is needed to damp the variance even though this introduces substantial bias. Overall the average increased loss in using RDA in this most unfavorable circumstance is slight.

We next modify this problem slightly. The same (unfavorable) covariance structure [Eq. (28)] is used for each class population, but the mean differences are concentrated in the high variance subspace. This provides the shrinkage
Table 3

<table>
<thead>
<tr>
<th></th>
<th>$P = 6$</th>
<th>$p = 10$</th>
<th>$p = 20$</th>
<th>$p = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MISCLASSIFICATION RISK:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDA</td>
<td>.07 (.04)</td>
<td>.07 (.04)</td>
<td>.27 (.07)</td>
<td>.39 (.06)</td>
</tr>
<tr>
<td>LDA</td>
<td>.06 (.03)</td>
<td>.06 (.03)</td>
<td>.24 (.06)</td>
<td>.59 (.07)</td>
</tr>
<tr>
<td>QDA</td>
<td>.17 (.08)</td>
<td>.14 (.12)</td>
<td>.60 (.07)</td>
<td>.60 (.06)</td>
</tr>
<tr>
<td><strong>MINIMIZING CROSS-VALIDATED ESTIMATE FOR RDA:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.05 (.04)</td>
<td>.06 (.04)</td>
<td>.21 (.07)</td>
<td>.34 (.08)</td>
</tr>
<tr>
<td><strong>CORRELATION (TEST SET, CROSS-VALIDATION):</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.19</td>
<td>0.0</td>
<td>0.0</td>
<td>.16</td>
</tr>
<tr>
<td><strong>AVERAGE REGULARIZATION PARAMETER VALUES:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{\lambda}$</td>
<td>.77 (.33)</td>
<td>.83 (.27)</td>
<td>.75 (.30)</td>
<td>.72 (.32)</td>
</tr>
<tr>
<td>$\bar{\eta}$</td>
<td>.02 (.08)</td>
<td>.07 (.16)</td>
<td>.19 (.27)</td>
<td>.45 (.25)</td>
</tr>
</tbody>
</table>

strategy with at least a chance at accomplishing some improvement. For this case the class two and class three means are given by

$$\mu_{2i} = 2.5 \sqrt{\frac{\hat{e}_i}{p \frac{p}{2} - 1}} i^{-1},$$

$$\mu_{3i} = (-1)^i \mu_{2i}, \quad 1 \leq i \leq p$$

while the class one mean is again located at the origin. Table 4 summarizes the results.

Even though the class population covariance matrices are highly ellipsoidal, the rather high degree of shrinkage towards the identity matrix does not increase the bias of the classification rule very much. The population class means differ here mostly in the high variance subspace, so deemphasizing the low variance
subspace has little consequence in terms of biasing the discriminant rule, even though it highly biases the covariance matrix estimates. The corresponding decrease in variance, however, allows RDA to outperform LDA, again especially in the high dimensional settings. Note that here, where the RDA misclassification risk is quite small, the minimizing cross-validated estimate seems to more seriously underestimate the actual risk ($\approx 30\%$).

Table 4

<table>
<thead>
<tr>
<th></th>
<th>$P = 6$</th>
<th>$p = 10$</th>
<th>$p = 20$</th>
<th>$p = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MISCLASSIFICATION RISK:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDA</td>
<td>.06 (.03)</td>
<td>.05 (.02)</td>
<td>.14 (.04)</td>
<td>.18 (.05)</td>
</tr>
<tr>
<td>LDA</td>
<td>.07 (.03)</td>
<td>.07 (.03)</td>
<td>.24 (.06)</td>
<td>.58 (.08)</td>
</tr>
<tr>
<td>QDA</td>
<td>.19 (.08)</td>
<td>.43 (.12)</td>
<td>.57 (.08)</td>
<td>.48 (.07)</td>
</tr>
<tr>
<td><strong>MINIMIZING CROSS-VALIDATED ESTIMATE FOR RDA:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.04 (.03)</td>
<td>.03 (.03)</td>
<td>.11 (.05)</td>
<td>.14 (.06)</td>
</tr>
<tr>
<td><strong>CORRELATION (TEST SET, CROSS-VALIDATION):</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.16</td>
<td>-.20</td>
<td>-.07</td>
<td>.13</td>
</tr>
<tr>
<td><strong>AVERAGE REGULARIZATION PARAMETER VALUES:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda$</td>
<td>.92 (.24)</td>
<td>.86 (.30)</td>
<td>.72 (.38)</td>
<td>.76 (.36)</td>
</tr>
<tr>
<td>$\eta$</td>
<td>.71 (.36)</td>
<td>.66 (.36)</td>
<td>.70 (.29)</td>
<td>.79 (.23)</td>
</tr>
</tbody>
</table>

5.4 UNEQUAL HIGHLY ELLIPSOIDAL COVARIANCE MATRICES

Our last two examples complete the sequence by considering cases where the class population covariance matrices are highly ellipsoidal and very unequal.
The eigenvalues for class one are given by Eq. (28). Those for class two are given by

\[ \lambda_{12} = \left[ \frac{9(p-i)}{p-1} + 1 \right]^2, \quad 1 \leq i \leq p, \]

while those for class three are

\[ \lambda_{13} = \left\{ \frac{9(i - \frac{(p-1)}{2})}{(p-1)} \right\}^2, \quad 1 \leq i \leq p. \]

The population eigenvectors for all three classes are the same. For the first two classes the ratio of the largest to smallest eigenvalues is one hundred, but their high and low variance subspaces are complementary to each other. This ratio for the third class is \((p+1)^2\). It has low variance in the subspace of intermediate variance for the first two classes, and high variance where they have their complementary high/low variances. The first case we consider is where the population means are all identical so that the class distributions differ only in their covariance matrices. Table 5 presents the results.

As would be expected, LDA does very poorly because the population class means are all the same. For the lowest dimension, RDA is slightly worse than QDA, but for the rest RDA is substantially better. Again the model selection procedure is tending to do the right thing. Very little covariance matrix mixing is selected at any dimension, while the eigenvalue shrinkage increases with dimension.

The final simulation example uses the same covariance structure as the previous one. The population class means, however, are different. The class one mean is at the origin. The class two and class three mean vectors are given by

\[ \mu_{2i} = \frac{14}{\sqrt{p}}, \quad \mu_{3i} = (-1)^i \mu_{2i}, \]


Table 5

<table>
<thead>
<tr>
<th></th>
<th>( P = 6 )</th>
<th>( p = 10 )</th>
<th>( p = 20 )</th>
<th>( p = 40 )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MISCLASSIFICATION RISK:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDA</td>
<td>.21 (.06)</td>
<td>.15 (.06)</td>
<td>.12 (.05)</td>
<td>.12 (.06)</td>
</tr>
<tr>
<td>LDA</td>
<td>.61 (.06)</td>
<td>.58 (.06)</td>
<td>.58 (.06)</td>
<td>.63 (.06)</td>
</tr>
<tr>
<td>QDA</td>
<td>.19 (.06)</td>
<td>.35 (.13)</td>
<td>.44 (.10)</td>
<td>.43 (.07)</td>
</tr>
<tr>
<td><strong>MINIMIZING CROSS-VALIDATED ESTIMATE FOR RDA:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.17 (.06)</td>
<td>.13 (.05)</td>
<td>.11 (.05)</td>
<td>.12 (.06)</td>
</tr>
<tr>
<td><strong>CORRELATION (TEST SET, CROSS-VALIDATION):</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.03</td>
<td>-.03</td>
<td>.09</td>
<td>.25</td>
</tr>
<tr>
<td><strong>AVERAGE REGULARIZATION PARAMETER VALUES:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \lambda )</td>
<td>.03 (.05)</td>
<td>.04 (.06)</td>
<td>.06 (.07)</td>
<td>.05 (.07)</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>.17 (.16)</td>
<td>.27 (.18)</td>
<td>.46 (.17)</td>
<td>.60 (.15)</td>
</tr>
</tbody>
</table>

along the respective eigenvectors. The results of this experiment are presented in Table 6.

The presence of the differing class means improves the risk associated with all three methods. Again, RDA substantially dominates the others except at the lowest dimension, where it has comparable risk to QDA.

5.5 **REMARKS ON THE SIMULATION RESULTS**

The model selection procedure based on cross-validatory choice seems to perform surprisingly well. In each of the simulated examples the best joint values for the covariance matrix mixing parameter \( \lambda \), and eigenvalue shrinkage parameter \( \gamma \), are roughly known. The distributions of the sample based estimates are in each case seen to concentrate near these optimal values. This is why RDA


<table>
<thead>
<tr>
<th></th>
<th>$P = 6$</th>
<th>$p = 10$</th>
<th>$p = 20$</th>
<th>$p = 40$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MISCLASSIFICATION RISK:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDA</td>
<td>.07 (.04)</td>
<td>.07 (.03)</td>
<td>.06 (.04)</td>
<td>.07 (.06)</td>
</tr>
<tr>
<td>LDA</td>
<td>.17 (.04)</td>
<td>.20 (.05)</td>
<td>.28 (.06)</td>
<td>.54 (.09)</td>
</tr>
<tr>
<td>QDA</td>
<td>.06 (.05)</td>
<td>.28 (.16)</td>
<td>.35 (.13)</td>
<td>.28 (.08)</td>
</tr>
<tr>
<td><strong>MINIMIZING CROSS-VALIDATED ESTIMATE FOR RDA:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.04 (.03)</td>
<td>.04 (.03)</td>
<td>.05 (.03)</td>
<td>.06 (.04)</td>
</tr>
<tr>
<td><strong>CORRELATION (TEST SET, CROSS-VALIDATION):</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.11</td>
<td>.06</td>
<td>.05</td>
<td>.35</td>
</tr>
<tr>
<td><strong>AVERAGE REGULARIZATION PARAMETER VALUES:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\lambda$</td>
<td>.09 (.12)</td>
<td>.10 (.11)</td>
<td>.10 (.12)</td>
<td>.10 (.12)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>.25 (.25)</td>
<td>.38 (.28)</td>
<td>.54 (.20)</td>
<td>.62 (.18)</td>
</tr>
</tbody>
</table>

seems to lose so little in unfavorable situations and gain so much in favorable ones. It is also surprising how small the observation to variable ratio can be and still permit fairly accurate classification with RDA. It is not surprising that the cross-validated estimate of misclassification risk for RDA somewhat underestimates the actual risk ($\approx 20\%$) on average, since this quantity is minimized with respect to the regularization parameters for each individual training sample. What is surprising is its low correlation with the actual misclassification risk. This means that when an especially favorable or unfavorable training sample is realized (from the population), the minimized cross-validation estimate provides no apparent reflection of this. Cross-validation provides an estimate of the average performance of a procedure but not necessarily its performance with a particular training sample.
The minimal Bayes risk for all of the simulated situations is quite low, but the class means were not widely separated with respect to their covariances. When the means are widely separated, any (reasonable) classification procedure will provide good results and regularization will not be particularly beneficial, although it won’t hurt either. In well-posed situations where the class sample sizes are all very large compared to the number of measurement variables, there is usually little benefit to be derived from regularization, and as the simulations indicate, sometimes there is a small degradation in performance when employing regularization in these settings.

6.0 Data Examples

In this section we study the performance of RDA on two data sets. The first data set is quite large. We use it to study small sample performance by drawing small training samples from it and then validating performance by classifying the large remaining (test) samples. The second data set is a small one, so we employ the 632 bootstrap [Efron (1983)] to evaluate performance. The prior probabilities were taken to be equal, \( \pi_k = 1/K \), for all classes. The optimization grid point values for \( \lambda \) were the same as for the simulation examples. The values for \( \gamma \) were taken to be \( \gamma = (0.0, .037, .105, .192, .30, .414, .544, .686, .838, 1.0) \). The intent here is to use these data sets to study the effect of regularization on misclassification risk, and not to present complete or definitive analyses of these data.

6.1 Kangaroo Skull Measurements

This data set is given in Andrews and Hertzberg (1985). It consists of 18 skull measurements on each of 148 kangaroos of three different species. The objective is to classify the kangaroos to one of the three species based on the skull measurements. Since the data are plentiful we confine our analysis to the 101 animals where all 18 measurements were made. Missing values can be accommodated in RDA by employing the same techniques that are commonly used for LDA and QDA [see Lachenbruch (1975)].

Since this data set is large, it is not likely that regularization will provide much benefit. We can use this data set, however, to examine small sample performance by repeatedly drawing smaller samples from it to be used as training samples,
Table 7

<table>
<thead>
<tr>
<th>( \bar{N} )</th>
<th>20.2</th>
<th>25.25</th>
<th>33.67</th>
<th>50.5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>MISCLASSIFICATION RISK:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RDA</td>
<td>.29 ± .02</td>
<td>.25 ± .01</td>
<td>.24 ± .01</td>
<td>.26 ± .02</td>
</tr>
<tr>
<td>LDA</td>
<td>.46 ± .03</td>
<td>.46 ± .03</td>
<td>.27 ± .04</td>
<td>.23 ± .02</td>
</tr>
<tr>
<td>QDA</td>
<td>.43 ± .03</td>
<td>.42 ± .03</td>
<td>.55 ± .03</td>
<td>.59 ± .06</td>
</tr>
<tr>
<td><strong>MINIMIZING CROSS-VALIDATED ESTIMATE FOR RDA:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.20</td>
<td>.20</td>
<td>.21</td>
<td>.19</td>
</tr>
<tr>
<td><strong>AVERAGE REGULARIZATION PARAMETER VALUES:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \lambda )</td>
<td>.82</td>
<td>.82</td>
<td>.67</td>
<td>.82</td>
</tr>
<tr>
<td>( \gamma )</td>
<td>.20</td>
<td>.25</td>
<td>.06</td>
<td>.36</td>
</tr>
</tbody>
</table>

and then use the remaining left-out observations to validate the classification rules derived from the small training sets. Four exercises of this type were performed, each with different training sample sizes.

In the first exercise every fifth observation of the 101 was taken to be in the training sample; in the second—every fourth; in the third—every third; and in the fourth—every second observation was used for training. The first exercise was repeated five times; the second—four times; the third—three times; and the fourth twice; each time choosing a different subset for training. The results averaged over these trials are presented in Table 7. The uncertainty estimates are obtained by dividing the standard deviations of the corresponding quantities by the square root of the number of trials. Note that none of these quantities are independent. The quantity \( \bar{N} \) is the average training sample size over the corresponding set of trials. Note also, that unlike the simulation studies, here the measurement space dimensionality is being held constant \( (p = 18) \) while the sample size is varying.
Judging from the regularization parameter estimates, this situation appears similar to that presented in Table 3 (Section 5.3): that is, large values of the covariance matrix mixing parameter $\lambda$ and small values of the shrinkage parameter $\gamma$. These are situations highly favorable to LDA, and RDA appears to be trying to produce a classification rule close to LDA. It has better performance for the smaller sample sizes owing to the fact that even LDA is poorly-posed in these cases. It is noteworthy that RDA produces a classification rule with 25 observations that has misclassification risk almost as good as LDA or RDA based on twice that much data.

6.2 WINE TASTING DATA

This data consists of 38 different wine samples made from the Pinot Noir (Burgundy) grape [Kwan and Kowalski (1980)]. The wines were subjected to taste tests by 16 judges and graded with numerical scores on 14 sensory characteristics. These characteristics were: clarity, color, aroma intensity, aroma character, undesirable odor, acidity, sugar, body, flavor intensity, flavor character, oakiness, astringency, undesirable taste, and overall quality. These wines originate from three different geographical regions: 9 from California, 17 from the Pacific Northwest and 12 from France. The purpose is to classify the geographical origins of the wine samples from the 14 sensory characteristics.

Two analyses were performed. In the first RDA, LDA and QDA were applied to the entire data set. In the second the data were divided into two samples each of size 19. Each half sample was then used as a training set and the three classification rules so obtained were validated on the other sample. In the first analysis there is no validation sample, so we must use a sample reuse technique to estimate the future misclassification risk of the classification rules. We use the 632 bootstrap [Efron (1983)] which has shown superior performance over other sample reuse techniques for this purpose in several simulation studies [Efron (1983), Gong (1982) and Crawford (1986)]. One hundred bootstrap replications were employed.

Applying RDA to the entire sample ($N = 38$) gave a minimizing cross-validated misclassification risk of 0.14 at $\lambda = 0.35$ and $\gamma = 0.04$. The 632 bootstrap estimates for RDA, LDA and QDA were respectively 0.18, 0.26,
and 0.36. The distribution of $\lambda$ over the 100 bootstrap replications had a mean of $\bar{\lambda} = 0.49$ and a standard deviation of $\sigma(\lambda) = 0.37$. The corresponding quantities for the distribution of $\gamma$ were $\bar{\gamma} = 0.40$ and $\sigma(\gamma) = 0.31$.

The results for RDA averaged over the two half sample runs ($N = 19$) gave $\bar{\lambda} = 0.56$, $\bar{\gamma} = 0.48$, with an averaged minimizing cross-validated risk of 0.19. The average misclassification risks of RDA, LDA and QDA, obtained from the half sample complementary to the corresponding training sample, were respectively 0.21, 0.50 and 0.59.

Judging from the chosen values of the regularization parameters, this does not appear to be a situation favorable to LDA. This is also indicated by the substantially superior performance of RDA for the larger ($N = 38$) sample where LDA is fairly well-posed for $p = 14$. When the sample size is reduced to $N = 19$ the performance of RDA seems to be degraded surprisingly little while LDA appears to completely collapse.

Concluding Remarks

Both the simulation studies and the data examples indicate that the method of regularization as applied here has the potential for increasing the power of discriminant analysis in settings for which sample sizes are small and the number of measurement variables is large. There appears to be at most a small loss in applying RDA in unfavorable situations and often substantial to dramatic gains in favorable circumstances. Of course, one generally does not know the type of situation in advance when confronted with a particular data set.

The regularization method presented here is rotationally invariant. That is, if the measurement variables of the training data and future test data are subjected to the same orthonormal rotation, the RDA classification rule would not change. The same is of course true for LDA and QDA. Unlike LDA and QDA, however, RDA is not generally scale invariant. That is, changing the relative scales of the measurement variables, or their linear combinations, can change the classification rule. This lack of scale invariance results from the introduction of the eigenvalue shrinkage parameter $\gamma$. If $\gamma = 0$ then RDA is scale invariant. This lack of scale invariance is a common property of many regularization methods that shrink eigenvalues, such as ridge and principal components regression.
Lack of scale invariance raises the question of which particular scaling is most appropriate. One natural choice is to use the original measurement scales. That was the choice used in our examples. An equally natural choice is to standardize or "auto-scale" all variables to have the same variance. The best scaling for the measurement variables, or their linear combinations, depends on the particular problem at hand and is usually unknown in advance. One can experiment with several scalings using the minimized cross-validated risk estimate as a guide.

Changing the scales of the measurement variables or their linear combinations is equivalent to changing the regularization matrix for \( \hat{\Sigma}_k(\lambda, \gamma) \) in Eq. (19). Here \( \hat{\Sigma}_k(\lambda) \) [Eq. (17)] was regularized by shrinking it toward a multiple of the identity matrix \( I \). There is clearly nothing special about this particular choice and one could consider more general regularizations of the form

\[
\hat{\Sigma}_k(\lambda, \gamma) = (1 - \gamma) \hat{\Sigma}_k(\lambda) + \gamma t M,
\]

with \( M \) a prespecified positive definite symmetric matrix and

\[
t = \frac{\text{trace}(\hat{\Sigma}_k(\lambda))}{\text{trace}(M)}.
\]

One can implement this generalized approach, using the techniques outlined in Sections 2 and 3, by first applying a transformation to the data (rotation and scaling) that takes \( M \) to the identity matrix. Let

\[
M = LL^T
\]

be the Cholesky factorization of \( M \), where \( L \) is a lower triangular matrix. Then applying the transformation

\[
Y_v = L^{-1} X_v, \quad 1 \leq v \leq N,
\]

or equivalently, solving the linear system

\[
LY_v = X_v, \quad 1 \leq v \leq N,
\]

for \( Y_v \) performs a rotation and scaling such that the matrix \( M \) is represented by the identity matrix in the transformed coordinate system. Then applying RDA to
the transformed data [Eq. (32)] is equivalent to specifying $M$ as the regularizing matrix [Eqs. (29) and (30)] in the original coordinate system.

Standardizing or auto-scaling the data is equivalent to using the diagonal matrix

$$M = \text{diag} (\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_p^2) ,$$

for regularization, where $\hat{\sigma}_i$ is the sample standard deviation of the $i^{th}$ measurement variable. Choice of a particular matrix $M$ is analogous to choosing a metric ($M^{-1}$) for a nearest means classification procedure. In the absence of any prior information, there is no clear best choice and, as indicated above, one might experiment with several choices.

There can be situations, however, where particular regularizations are suggested. When the data measurement vectors $X_e$ arise from a signal or image, there is a natural distance measure between variables or, more precisely, their indices. Each signal digitization point, or each image pixel, corresponds to a measurement variable. If one believes that in the absence of error, close measurement variables ought to have similar values, then a natural regularization matrix to try would be

$$M = H^T H , \quad (33a)$$

with $H$ being the matrix representation of some smoothing kernel

$$H_{ij} = h \left( \frac{d_{ij}}{s} \right) . \quad (33b)$$

where $h$ is (usually) a positive monotonically decreasing function such that

$$\sum_{j=1}^{p} H_{ij} = 1 , \quad (33c)$$

$d_{ij}$ is a distance between the indices $i$ and $j$, and $s$ is the bandwidth parameter for the smoothing kernel. Using $M$ [Eq. (33)] for regularization tends to deemphasize directions in the measurement variable space dominated by differences of those variables that correspond to close pixels or digitization points. This approach attempts to use to advantage the spatial nature of the problem in suggesting a particular regularization matrix $M$. 

31
A common method of regularization used with LDA and QDA is measurement variable subset selection. One attempts to reduce variance while not introducing excessive bias by the judicious selection of a small subset of the original set of variables. Stepwise and "all subsets" strategies are often employed. Subset selection is scale invariant, but clearly not rotationally invariant. If the mean vector and covariance matrix differences between the class populations happen to align principally along a very small number of the original measurement variables, then subset selection strategies can be effective. Variable subset selection can be used in addition to, or in conjunction with, the regularization methods presented here. It should be kept in mind, however, that although variable subset selection seems very natural and readily understandable, it can be fairly ineffective in these settings where variance dominates the prediction error. A heuristic explanation for this is as follows.

The bias of a prediction rule depends largely on the true underlying (population) means and covariance matrices, about which there is often little prior knowledge. The variance, on the other hand, depends mostly on the particular estimation method being used, about which there is considerable knowledge. Covariance matrix shrinkage techniques basically use this information to attempt to achieve maximal reduction in variance (for a given level of regularization) by preferentially damping the influence of those directions (eigenvectors) associated with the smallest eigenvalues. These are the directions (linear combinations of the variables) that contribute most strongly to the variance, and are of course obtainable from the sample covariance matrix. Therefore, in the absence of any prior knowledge of how one is affecting the bias, it makes sense to regularize in a way that achieves the largest reduction in variance for a given level of regularization.

Variable subset selection, on the other hand, assumes fairly specific prior knowledge concerning the population class means and covariance matrices. Namely, that the (standardized) class means and covariance matrices differ mostly in a small subset of the measurement variables and are similar in the larger complement subset. If this is true and if one can reliable identify the small subset, then by damping the influence of the complement subset of variables, one introduces very little bias while achieving some reduction in variance.
The relative efficacy of the two approaches in particular situations depends on the degree to which the assumption inherent in the subset selection method is valid. The size of the influential subset must be surprisingly small, however, for subset selection techniques to be competitive with other regularization methods, or even no regularization at all [see Copas (1983)].

As the examples indicate (and as is well known) QDA is only viable in situations where the ratio of sample size to variable count is large. For these situations the largest contributor to prediction error is not usually variance but lack of ellipsoidal symmetry of the population class densities. In such cases nonparametric classification techniques are more appropriate [see Lachenbruch (1975) and Breiman et al (1983)]. For the situations that we have been considering here (small samples and high variable count) LDA has been the method of choice in the past. The additional regularization alternatives provided by RDA can substantially improve misclassification risk when the population class covariance matrices are not close to being equal and/or the sample size is too small for even LDA to be viable.

A FORTRAN program implementing the RDA procedure is available from the author.
References


